technologies

MDPI

Article

Open-Weight Multimodal LLMs Versus Manual Data Entry for Legacy ERP Digitization: A Comparative Evaluation of Accuracy, Cost, and Verifiability

Chacharin Lertyosbordin [1,*] and Boonyakorn Trangadisaikul [2]

[1] School of Information Technology, King Mongkut's Institute of Technology Ladkrabang, Bangkok 10520, Thailand
[2] Ruamrudee International School, Bangkok 10510, Thailand; boonyakornt27@rism.ac.th
* Correspondence: chacharin.le@kmitl.ac.th

Abstract

Decades-old enterprise-resource-planning (ERP) systems lock operational data inside unstructured, human-readable reports, forcing slow, costly, error-prone manual re-keying. Because multimodal large language model (MLLM) capability is uneven, deploying MLLMs for extraction means trusting outputs without a labeled reference. We test this with a within-document controlled experiment on 400 controlled-substance stock-ledger documents (2951 records, 11 fields, predominantly Thai) from a Thai pharmaceutical factory, comparing trained human double-entry against four open-weight MLLMs (2 × 2 design: vendor × architecture) via OpenRouter. Human double-entry left 14 discrepancies against the adjudicated gold standard, none common to both operators. The strongest model, Qwen3-VL-32B-Instruct (Dense), reached 93.95% cell accuracy; among these four models, field accuracy varied more across vendors, whereas structural completeness differed consistently between dense models (0 missing records) and Mixture-of-Experts models (up to 51 of 2951 dropped). Deterministic accounting invariants flagged 0.61% of its records, leaving the unflagged majority 94.1% accurate across all 11 fields; adding calendar rules flagged 4.61% and raised residual date accuracy from 92.1% to 95.9%. We report both operating points and recommend the extended level where date fidelity is regulatory-critical. The pipeline is 13.5–29.4× faster in wall-clock terms and 97.5–99.5% cheaper. Gold-free, rule-based verification thus locates where MLLM reliability holds, giving human–AI collaboration quantified, disclosed residual risk rather than an implied guarantee. Even at the more conservative operating point, unflagged records average 94.5% accuracy across all 11 fields but only 43.7% on the free-text Remarks field, which the triage cannot check; the results support risk reduction and the localization of review effort, not unrestricted regulatory reliability across all fields.

Keywords: multimodal large language models; intelligent document processing; ERP digitization; open-weight models; data verifiability; human-in-the-loop; Thai document understanding; Mixture-of-Experts

Check for updates

Academic Editor: Zhipeng Wu

Received: 13 July 2026
Revised: 18 August 2026
Accepted: 20 August 2026
Published: 24 August 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Legacy technology infrastructure is a persistent obstacle to applying new AI capabilities inside established organizations: automation tools keep improving, but the systems to which they need to connect often do not. This study examines this connection problem in a specific, high-stakes form: how open-weight multimodal large language models (MLLMs),

models whose trained weights are openly released, can be systematically evaluated for the task of unlocking data trapped inside legacy enterprise-resource-planning (ERP) reports in a highly regulated setting.

Organizations worldwide continue to depend on legacy information systems for core operations. These systems are expensive to maintain and carry security risk, yet they persist because critical data and processes are bound to them. The U.S. Government Accountability Office reports that federal agencies spend over one hundred billion dollars annually on information technology, much of it sustaining systems that are decades old [1]. One consequence is that vast quantities of data become trapped in forms that are difficult to reuse, turning into "dark data" that an organization holds but never exploits; surveys estimate that roughly 55% of enterprise data is dark and goes unused or unreachable [2].

The problem is acute in legacy ERP systems, many of which export reports as cleanly laid-out PDFs for human reading but, when exported to spreadsheets, yield unstructured output that cannot be converted into standard formats such as JSON for integration with modern systems. Organizations therefore fall back on manual re-keying, a process that is slow, costly, and systematically error-prone. Empirical work confirms that human data entry introduces errors capable of materially distorting downstream results and conclusions [3], and industry estimates place manual-entry error at roughly 1–4% of fields, even for trained staff [4]. Preparing and structuring data correctly is thus a precondition for any data-driven decision system [5].

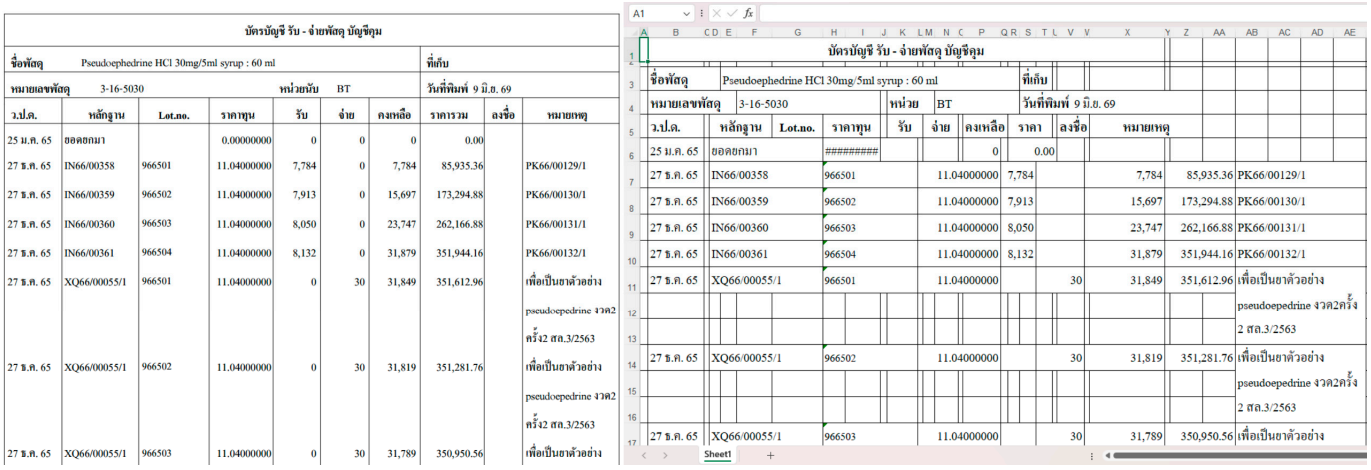
Advances in MLLMs open a new route to converting document images directly into structured data, outperforming traditional optical character recognition on complex layouts, tables, and handwriting, albeit at a higher per-page cost [6]. Realizing this opportunity in a regulated setting raises two concerns beyond raw accuracy: many documents of interest are written in a low-resource, orthographically complex language such as Thai [7–9], and the model output must be sufficiently trustworthy and verifiable for high-precision work, since even capable generative models are known to produce plausible but incorrect structured output when unchecked [10]. Both concerns are addressed later in this paper: Thai document understanding in Section 2.1, and verifiability, the central concern of this work, from Section 2.3 onward.

AI capability, however, is not uniformly high across tasks. Recent field-experimental work on knowledge workers found that large language models exceed human performance on some tasks while underperforming on others, with no single boundary separating what the technology can and cannot do reliably: a "jagged" frontier of capability rather than a smooth one [11]. That original finding comes from open-ended, creative knowledge work; we do not claim it transfers directly to structured information extraction. We extend the concept rather than the empirical claim, as a way of naming a pattern that our own results independently reproduce (Sections 4.2 and 5.2): the same model that nearly saturates numeric fields can fail on free-text fields, and which vendor or architecture performs best depends on which axis of performance is asked about. This unevenness compounds a well-documented tension in AI-assisted work (automation can as easily degrade a process as improve it, depending on how the work is divided between humans and the system [12]) and means a deployment cannot responsibly assume uniform reliability across a corpus. What is needed instead is a way to know, for each record, whether the extracted output falls inside or outside the reliable region of the frontier, without requiring a human-labeled reference for every new document. That capability (gold-free, rule-based verification of MLLM output) is the central technical contribution of this paper.

These challenges converge, with the highest stakes, in the pharmaceutical manufacturing of controlled substances. Thailand maintains a comprehensive legal framework for drugs spanning many statutes and several categories, including the control of medicinal

products [13]. For psychotropic and narcotic substances, possession, manufacture, and distribution fall under dedicated legislation now consolidated into the Narcotics Code, with periodic statutory reporting required of operators [14]. A controlled-substance stock ledger is therefore not merely an internal operating record but a legal obligation, demanding high accuracy and traceability.

The case study in this work is a state pharmaceutical factory in Thailand that has kept stock ledgers for controlled substances in a legacy ERP system for more than a decade. The system exports legible PDF reports, but its spreadsheet export is unstructured, with cells merged and misaligned, and cannot be converted into structured data (Figure 1). More than a decade of data therefore remains locked in documents, and the current process requires staff to re-key it by hand.



บัตรบัญชี รับ - จ่ายพัสดุ บัญชีคุม

ชื่อพัสดุ	Pseudoephedrine HCl 30mg/5ml syrup : 60 ml						ที่เก็บ	
หมายเลขพัสดุ	3-16-5030			หน่วยนับ	BT		วันที่พิมพ์ 9 มิ.ย. 69	
ว.ป.ค.	หลักฐาน	Lot.no.	ราคาทุน	รับ	จ่าย	คงเหลือ	ราคารวม	ลงชื่อ	หมายเหตุ
25 ม.ค. 65	ยอดยกมา		0.00000000	0	0	0	0.00		
27 ธ.ค. 65	IN66/00358	966501	11.04000000	7,784	0	7,784	85,935.36		PK66/00129/1
27 ธ.ค. 65	IN66/00359	966502	11.04000000	7,913	0	15,697	173,294.88		PK66/00130/1
27 ธ.ค. 65	IN66/00360	966503	11.04000000	8,050	0	23,747	262,166.88		PK66/00131/1
27 ธ.ค. 65	IN66/00361	966504	11.04000000	8,132	0	31,879	351,944.16		PK66/00132/1
27 ธ.ค. 65	XQ66/00055/1	966501	11.04000000	0	30	31,849	351,612.96		เพื่อเป็นยาตัวอย่าง pseudoepedrine จวค2 ครั้ง42 สค.3/2563
27 ธ.ค. 65	XQ66/00055/1	966502	11.04000000	0	30	31,819	351,281.76		เพื่อเป็นยาตัวอย่าง pseudoepedrine จวค2 ครั้ง42 สค.3/2563
27 ธ.ค. 65	XQ66/00055/1	966503	11.04000000	0	30	31,789	350,950.56		เพื่อเป็นยาตัวอย่าง

Figure 1. Legible PDF export versus unstructured spreadsheet export of one ledger card from the legacy ERP system. Numeral formatting shown reflects the legacy system's native, unedited output; inconsistencies are part of the digitization problem this figure illustrates and have been deliberately left unaltered.

Four considerations motivated this study. First, before building the extraction pipeline evaluated here, the team piloted a purpose-built Thai optical character recognition (OCR) model, Typhoon OCR [15], on a sample of ten stock-ledger documents from the same corpus. The output was legible text with reasonable layout, but not integration-ready records: inspecting the underlying structured (HTML table) output mode directly, the number of columns produced per row was not constant from one transaction row to the next, and document-header metadata (item code, storage location, print date) was fused into the same table as the transaction rows rather than kept separate (Figure 2). A representation that cannot guarantee a fixed field count per record cannot be loaded into a downstream schema without additional, document-specific repair. That is precisely the manual burden this study set out to remove. This structural finding, not a comparison of extraction accuracy, is what motivated moving to a multimodal large language model pipeline built around an explicit, fixed 11-field JSON schema (Section 3.4) rather than a general-purpose OCR-to-table approach. Second, the organization operates an active digital-transformation program under which this pilot was chartered. Third, the re-keying burden described above is a recurring, unsolved operational problem that a systematic solution could measurably reduce. Fourth, the organization required empirical evidence of feasibility and value before committing to larger-scale investment in the technology.



Figure 2. The same ledger card processed by Typhoon OCR: document header metadata (item number, storage location, print date) is fused into the same table as the transaction rows, and the number of output columns is not constant row-to-row, illustrating that transcription output does not reliably preserve a fixed field count.

We designed a controlled experiment comparing open-weight MLLM extraction with human data entry on 400 real stock-ledger documents (2951 records) exported from the legacy ERP described above. We built a prototype that renders PDF pages to images and extracts structured JSON, and evaluated four models in a 2 × 2 factorial design (two vendors × two architectures: dense, which activates all model parameters for every token, and Mixture-of-Experts (MoE), which routes each token through a small subset of specialized "expert" subnetworks). The evaluation spans three axes (accuracy, cost, and verifiability) and proposes a gold-free quality-control mechanism based on internal accounting-consistency rules. To keep the comparison controlled and transparently priced, all models are accessed through the OpenRouter API, used here as a standardized measurement harness rather than as a claim about any particular production deployment; token pricing and latency are therefore reported as measured through this harness, while the accuracy and verifiability results (properties of the model weights and of the deterministic checks applied to their output) are nonetheless reported as measured under the serving configurations observed during the evaluation period, not as properties established to hold under other deployments (Section 5.4).

This work makes four contributions. First, we design and build a prototype that renders legacy ERP PDF pages to images and uses open-weight MLLMs to extract structured, integration-ready JSON under a fixed 11-field schema. Second, we evaluate and compare four open-weight MLLMs in a 2 × 2 factorial design (vendor × dense/MoE architecture) against an independently double-entered human gold standard, at field and record level and stratified by data type, on real Thai documents in a regulated task, including a repeated-runs analysis of run-to-run reliability (Sections 3.1 and 4.3). Third, we analyze the cost–time–accuracy trade-off between MLLMs and human entry on the same corpus under conservative, worst-case review-time assumptions. Fourth, and most centrally, we propose and evaluate a gold-free quality-control mechanism (deterministic accounting and calendar invariants, framed as integrity constraints in the established data-quality-assessment literature [16] rather than as an ad hoc engineering trick) that screens extracted records for human review without requiring a labeled reference for the document being checked.

The overarching question of this research is whether open-weight MLLMs can convert semi-structured legacy ERP documents into integration-ready data at a level of accuracy, cost, and verifiability sufficient to augment human data entry in a regulated process. Three specific research questions guide the investigation:

(1) RQ1 (Accuracy): How accurately do MLLMs extract field-level data relative to a human gold standard, and how does accuracy vary by data type, vendor, and architecture?
(2) RQ2 (Operational cost): On the same corpus, how do MLLMs and human operators differ in the cost–time–accuracy trade-off?
(3) RQ3 (Verifiability): Can internal accounting-consistency rules serve as an automated, gold-free quality signal that screens records for human review and raises system trustworthiness?

2. Related Work

This review is organized into three connected themes: the evolution from OCR to MLLM-based extraction (Section 2.1); model architecture and deployment, in particular, dense versus Mixture-of-Experts (Section 2.2); and gold-free data-quality assessment and human-in-the-loop verification (Section 2.3). Each section closes with the gap that motivates our research questions.

2.1. From OCR to Multimodal Extraction

Deep-learning document understanding evolved well before large language models, with early surveys consolidating OCR and document-understanding techniques [17]. Intelligent document processing (IDP) subsequently combined computer vision, natural-language processing, and machine learning to automate extraction and structuring through staged pipelines [18]. The commercial momentum behind IDP is substantial: market analyses valued the sector at roughly USD 3.0 billion in 2025, projecting a compound annual growth rate near 33% over the following decade [19], a figure corroborated by independent estimates of comparable magnitude [20]. MLLMs extend this capability by extracting structured output such as JSON directly from images without predesigned templates [21], and outperform traditional OCR on complex layouts, tables, and handwriting [6].

Language models are nonetheless not a turnkey answer for every document task. Multimodal models have clear limits on structured visual content: benchmark evaluations show that most models score poorly on layout perception and the disambiguation of complex elements [22], and remain limited at interpreting abstract visual content such as charts, maps, and layouts [23]. For structured extraction specifically, surveys frame the conversion of text into structured knowledge with generative models [24], while dedicated benchmarks find that the validity of JSON output depends more on prompting strategy than on model size [25]. Hybrid OCR–LLM frameworks further show that copy-heavy documents with much repeated structural content can be extracted efficiently via table-based extraction [26], a property that matches the stock ledgers studied here. Low-resource, orthographically complex languages such as Thai add a further layer of difficulty, compounded by stacked diacritics, context-dependent vowel placement, and the absence of explicit word boundaries [8]; task-specific OCR on templated cards (close in spirit to the fixed-schema ledgers studied here) has been demonstrated for Thai documents specifically [27]. A cross-lingual study of an edge small language model further finds that content selection and prioritization are governed mainly by model architecture rather than by language context, with Thai attenuating but not eliminating the architectural gap [28]. This links the architectural theme in Section 2.2 to the linguistic dimension considered here.

A recurring theme across this literature is that visual recognition and integration-readiness are not the same property: a model can correctly read the content of a page while still failing to emit a representation with a fixed, predictable schema that a downstream system can load without repair [17,18,26]. Our own project's preliminary evaluation is consistent with this observation: piloting a purpose-built Thai OCR model on ten documents from the corpus studied here produced legible, well-laid-out text, but the

number of output columns per transaction row was not constant, and document-header metadata was fused into the same table as transaction data (Section 1, Figure 2).

This is a concrete instance of the recognition/integration-readiness gap rather than a shortcoming specific to that one tool. Gap: most benchmarks still test English or laboratory data, and very few address structured extraction (as opposed to transcription) on real legacy ERP documents with a fixed record schema.

2.2. Architecture and Deployment: Dense, MoE, and Open-Weight Models

The Mixture-of-Experts (MoE) architecture scales model capacity while controlling compute by sparsely activating only part of the parameters, yielding very large total parameter counts at near-constant compute [29]. Mixtral, for example, holds eight experts per layer and routes two per token, accessing 47 billion total parameters while activating only about 13 billion [30]. This gap between total and active parameters is the defining feature of the MoE models tested here, and design principles are catalogued in dedicated surveys [31]. We evaluate only open-weight models in this comparison: doing so keeps model weights, and therefore the accuracy and verifiability results computed from them, inspectable and independent of any single vendor's API availability or pricing changes, and keeps the comparison auditable from the released weights, prompt and scoring code (Section 3.4).

MoE's efficiency carries hidden inference-time costs (expert load imbalance and routing overhead) that make performance highly configuration-dependent across both language and vision models [32]. Most relevant to this work is the "seeing but not thinking" phenomenon in multimodal MoE models, which perceive visual content correctly yet fail to reason over it, because visual input induces routing to drift away from the mid-layer experts that text input would engage [33]. This offers a plausible mechanistic explanation, developed further in Section 5.2, for the structural-omission pattern we observe empirically: Mixture-of-Experts models in this study drop transaction records at a non-zero, architecture-linked rate across every independent run, while both dense models never do (Sections 4.3 and 4.4). This empirical pattern is what the routing-distraction mechanism, proposed in other multimodal MoE settings, was meant to explain.

Gap: there is very little empirical work directly comparing MoE and dense models on the structural completeness of real document extraction while controlling for both vendor and data type, and none that tests whether a structural-omission pattern replicates across independent runs.

2.3. Gold-Free Data Quality Assessment and Human-in-the-Loop Verification

Deploying language models in high-stakes domains such as finance and accounting confronts the risk of fluent but fabricated output, which is especially severe in tasks demanding high numeric precision and long context [10]. Rather than treat this purely as a language-model problem, we position the verification challenge within the older and more general literature on data quality assessment: correctness of a data record can be checked against integrity constraints, consistency rules (does the record contradict itself or a companion record?) and completeness rules (is anything missing that should be present?), independent of any particular data-collection technology [16]. Deterministic accounting identities of the kind we use in Section 3.5 are, in this framing, an instance of a well-established technique family, not an ad hoc engineering trick specific to language models; the broader practice of encoding such constraints and repairing or flagging violations is treated at length in the data-cleaning literature [34].

A parallel language-model literature approaches the same problem from the model side. One line analyzes a model's internal states (hidden states and attention maps) to

detect fabrication from a single response [35]. Another relies on self-consistency, sampling multiple answers or perspectives and comparing them [36]. Neuro-symbolic frameworks encode instructions as constraints and check them as a constraint-satisfaction problem [37], and other work integrates domain knowledge with language models to raise accuracy in expert tasks [38]. The visualization community has proposed design spaces for validating model-generated tabular data [39]. Closest to ours are two strands: selective human-in-the-loop review, which uses model uncertainty to route only suspect cases to humans (raising accuracy while cutting workload, but relying on repeated sampling to estimate uncertainty [36]), consistent with governance architectures that enforce policy constraints and route failing cases to review [40]; and the use of a document's internal structure to check numeric correctness, as in fiscal-document extraction that builds multi-level internal validation from hierarchical table totals [41].

Selective review of this kind only pays off, however, if the division of labor between model and human is well-designed: automation that routes the wrong cases to the wrong reviewers can degrade a process as easily as improve it [12], and technology more broadly displaces some tasks while reinstating and creating others rather than eliminating work outright [42]; the cognitive burden of deciding when to trust a model's output is itself a documented obstacle to productive delegation [43]. Collaborative-intelligence framings treat this division of labor as the central design problem rather than an afterthought [44]. Gap: existing verification approaches require access to model internals, repeated sampling, or generic constraints, and (apart from narrow fiscal-document totals checks [41]) none quantifies a domain accounting invariant as a gold-free screening signal, with precision, recall, and negative predictive value against an independently constructed human gold standard, in the context of regulated accounting documents; nor, to our knowledge, has any prior study tested whether such a signal's detection performance is stable across independent repeated runs of the underlying model.

Closing this three-part gap (across Sections 2.1–2.3) is what Sections 3 and 4 of this paper are designed to do, mapped directly onto the three research questions stated in Section 1: field accuracy by data type and architecture (RQ1), the cost–time–accuracy trade-off against human labor (RQ2), and a gold-free invariant-verification mechanism whose reliability is itself tested under repetition (RQ3).

3. Methodology

This section describes the methodology for evaluating the conversion of legacy ERP documents into structured data: a systematic comparison between open-weight multimodal large language models (MLLMs) and manual data entry. Subsequent subsections detail the experimental design, data preparation, extraction pipeline, and measurement instruments used to address the three research questions.

3.1. Experimental Design

This study is a comparative experiment on a real case from a Thai pharmaceutical factory, contrasting two routes for converting stock-ledger documents into structured data: manual data entry by operators and automatic extraction by open-weight MLLMs. The design is within-document repeated-measures—the same 400 documents are processed by every condition—enabling paired comparison and bootstrap uncertainty estimation. Conditions comprise two operators performing independent double-entry (H1, H2) and four AI models (A1–A4) in a 2 × 2 factorial design (vendor × architecture), listed in Table 1. Separating the two operators lets us report cost and time for both single-entry and double-entry, the latter being the verified practice actually used.

Table 1. Experimental conditions, processed over the identical 400-document set.

Condition	Vendor	Architecture	Params (Total/Active)	n Docs
A1 Gemma-MoE	Google	Mixture-of-Experts	26B total/4B active	400
A2 Gemma-Dense	Google	Dense	31B	400
A3 Qwen-MoE	Alibaba	Mixture-of-Experts	30B total/3B active	400
A4 Qwen-Dense	Alibaba	Dense	32B	400
H1/H2 (human)	—	Human double-entry, independent, blind	—	400

In addition to the main run above, a repeated-runs reliability check was carried out: 50 documents, drawn by simple random sampling from the full 400-document corpus (seed = 42, via the same experiment script used for the main run), were each processed 3 additional times per model, under the identical pipeline, prompt, and API access as the main experiment. The mean record count per document in this 50-document subsample (7.14 record/doc) is close to the full-corpus mean (7.38 record/doc), so the subsample is a reasonable, if not formally stratified, representative slice. This design supports Section 4.3 and directly addresses whether the single-pass main-run numbers in Sections 4.2 and 4.4 are representative of model behavior under repetition, or an artifact of one draw.

Table 2 lists the variables of the experiment. For transparency we also state the relevant confounds and limitations: total parameter sizes in the 2 × 2 pairing are not perfectly matched (26B vs. 31B, and 30B vs. 32B), model access is via an intermediary service (OpenRouter) that may affect the reproducibility of exact provider routing, and inference is a single pass per document in the main run, with repetition characterized separately above and in Section 4.3 rather than assumed away.

Table 2. Variables in the experiment.

Category	Variable	Levels/Definition
Controlled	Vendor	Google vs. Alibaba (Qwen)
Controlled	Architecture	Mixture-of-Experts vs. Dense
Controlled	Document set	Same 400 documents for every condition
Controlled	Prompt	Identical extraction prompt across all 4 models
Measured	Cell accuracy	3 equivalence tiers × 3 alignment strategies
Measured	Record-level exact match	All 11 fields correct simultaneously
Measured	Structural completeness	Extracted record count vs. gold record count
Measured	Gold-free triage performance	Precision/recall/NPV of invariant flags vs. actual errors
Measured	Run-to-run variability	SD/range of overall accuracy across repeated runs
Measured	Cost & time	USD and seconds per 100 records, full workflow

3.2. Corpus

The data are receipt–issue stock-ledger cards for controlled substances exported from the factory's legacy ERP: 400 documents containing 2951 transaction records (32,461 cells across the 11-field schema). Originals are PDFs, rendered to PNG images that serve both as model input and as the identical stimulus for operator entry. Most content is Thai—column headers, Buddhist-era dates and remarks—mixed with Latin characters and Arabic numerals for generic drug names and reference codes. A masked sample ledger card with all eleven field positions annotated is shown in Figure 3. Every record follows the 11-field schema in Table 3, grouped into three data types.

The 400 documents analyzed here were drawn via simple random sampling from the factory's digital archive. Factory staff executed this randomized selection and exported the set of single-page ledger cards, to which the research team applied its scoring pipeline.

The 400 documents span transaction dates from 31 May 2022 to 29 May 2026, computed directly from the transaction dates recorded across all 2951 gold-standard records. Each PDF page was rendered to a PNG image at 96 DPI, producing images of 1653 × 2337 pixels; PDF pages had already been split into one page per document by the factory prior to this study, so every document analyzed here is a single-page ledger card.

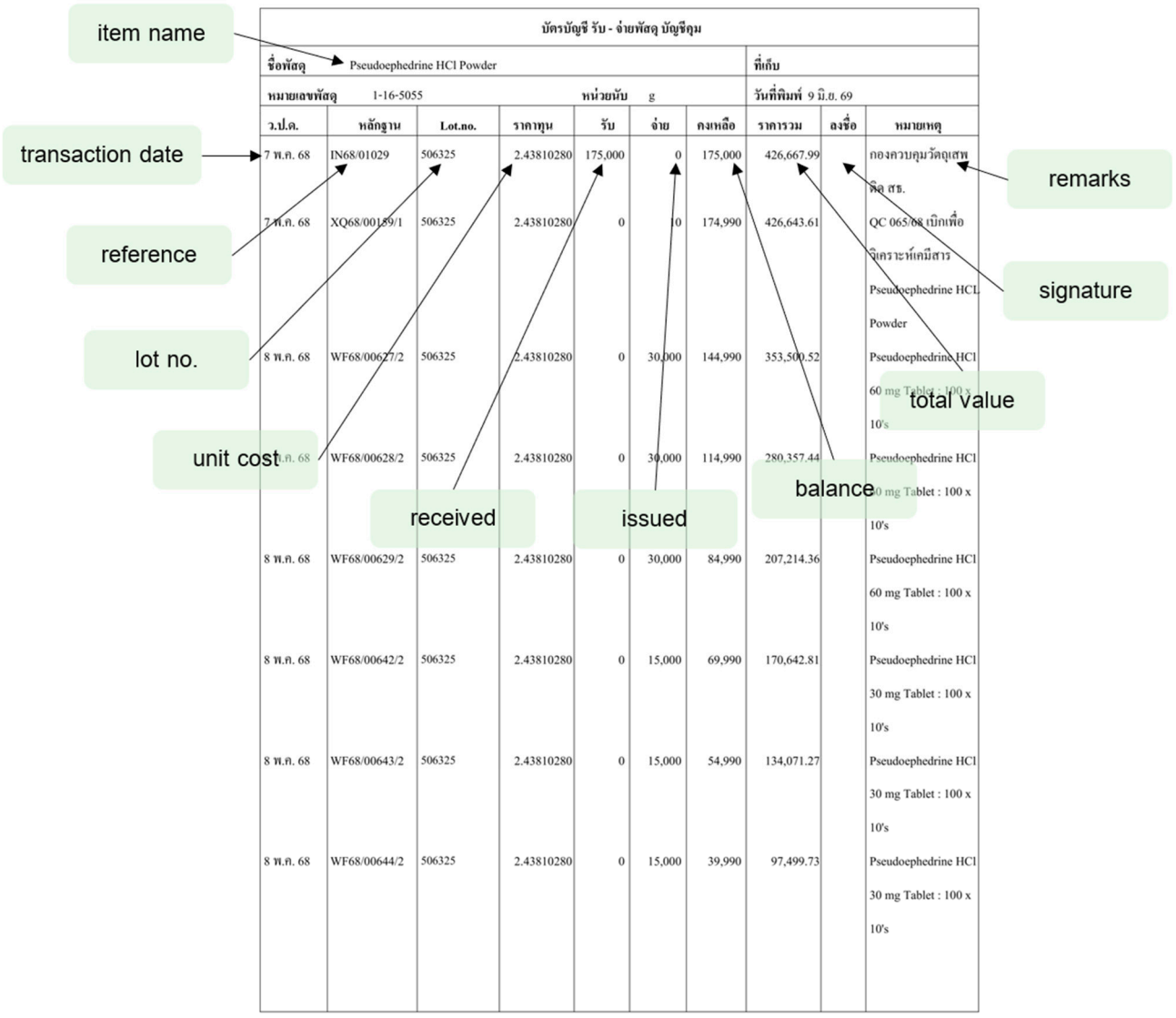


Figure 3. Masked sample ledger card with the eleven annotated field positions.

Table 3. The eleven-field record schema.

Data Type	Fields	Count
Numeric	Unit cost, Received, Issued, Balance, Total value	5
Text	Item name, Reference, Lot no., Signature (constant "-" placeholder), Remarks	5
Date	Transaction date (Buddhist-era format)	1

Under explicit institutional permission granted for this evaluation, the experiment processed the unredacted data; sensitive identifiers such as partner names and internal reference numbers are redacted in all published media and in the openly released repository (see the Data Availability Statement).

3.3. Gold-Standard Construction and Human Baseline

The gold standard was built by two trained operators who independently keyed all 400 documents (double-entry), following the practice known to minimize human entry error. Throughout this paper, "double-entry" denotes this duplicate data-capture procedure; it is unrelated to double-entry bookkeeping. In practice, each operator created a fresh spreadsheet per document, entered all 11 fields for every transaction row, and self-checked before saving. The two operators' outputs were then compared cell-by-cell; every disagreement was adjudicated against the original document image by the research team in two further review passes to produce the final reference set. This adjudication process, not majority agreement between the two operators, is the source of the published gold standard.

Independent entry by the two operators, checked against this adjudicated gold standard, produced 9 erroneous cells for H1 and 5 for H2 out of 32,461 (99.972% and 99.985% cell accuracy respectively)—all transcription slips (a misplaced decimal point, a single wrong digit) rather than comprehension errors, detailed in Section 4.1. These near-perfect individual accuracies do not imply that comparing the two operators against each other would find near-zero discrepancies: cell by cell, H1 and H2 disagree at 14 cells, matching the union of both operators' errors against gold, which is likewise 14 cells. We highlight this equality because it defines the practical value of double-entry as a quality-control procedure: comparing the two operators' entries against each other flags a cell whenever they disagree, without ever needing to consult the correct answer, and each of the 14 actual errors in this corpus surfaced as such a disagreement—none of the 14 was a case where both operators independently entered the same wrong value (zero concordant errors). Equivalently, this cross-check alone would have caught 100% of the errors present in this corpus—an empirical detection rate we report as a finding of this dataset, not a property guaranteed by the double-entry design. The explicit caveat is that such cross-checking cannot, in principle, catch an error both operators make identically; that is a limitation we did not encounter in this corpus (Section 5.4).

The resulting cell-level error rate (0.03–0.04%) is far below the roughly 1–4% often cited for general manual-entry tasks [4]. We attribute this gap to task characteristics rather than an implausibly capable workforce: the source documents are cleanly printed reports rather than handwritten forms, the fields being copied are predominantly numeric, operators self-check before saving, and this is routine work performed by trained, experienced staff—closer in kind to structured medical-record abstraction by trained abstractors than to general free-form data entry. Medical-record-abstraction studies report a comparable range of low, sub-1% error rates for trained abstractors performing structured transcription under similar conditions [45,46], which we take as a more apt comparison group than generic manual-entry statistics for interpreting the low error rate observed here.

3.4. Extraction Pipeline and Prompting

To run the experiment systematically, we developed a web-based prototype that acts as an extraction pipeline with three stages. First, the user uploads a legacy ERP PDF and the system renders each page to an image. Second, the system sends the page image together with a system prompt to the selected multimodal model through the OpenRouter API. Third, the model returns a JSON array matching the prescribed 11-field schema, ready for integration. Image-rendering time was logged separately from model-inference time so that the cost analysis (Section 3.5) reflects inference cost alone.

Accessing all models through OpenRouter is a deliberate methodological choice that makes inference systematic, easily accessible, and transparently priced per token, despite minor hardware-level variations across providers. The system prompt is zero-shot and schema-constrained: it directs the model to read the document from the rendered image

(rather than relying on an embedded text layer, which is frequently corrupted for Thai fonts), and it specifies transaction-row grouping rules, number normalization, and a strict JSON format. No decoding parameters were set: every request used the provider's default decoding configuration as served through OpenRouter, and output is therefore not assumed to be deterministic—repeated-runs data (Sections 3.1 and 4.3) characterize the residual run-to-run variability observed under that configuration as a measured property of the pipeline. The full system prompt is openly published in the GitHub repository referenced in the Data Availability Statement.

For each AI condition, all 400 documents were processed through the pipeline with one inference per document in the main run, logging image-rendering time, inference time, token counts, and cost per document. For the human condition, the two operators keyed the same image set with timing captured from file metadata (Section 3.3). All JSON outputs were stored for comparison against the gold standard in the measurement stage described next.

Inference for the main run and for the repeated-runs reliability check (Section 3.1) was carried out between 1 July and 5 August 2026. OpenRouter discloses the following provider-side weight precision for the models served during this window: A1 (Gemma-MoE, routed model gemma4-26B-A4) and A2 (Gemma-Dense, routed model gemma4-31b) were both served at BF16 precision; A3 (Qwen-MoE, routed model Qwen3-VL 30B-A3B) and A4 (Qwen-Dense, routed model Qwen3-VL 32B) were both served at FP8 precision. Beyond this disclosed precision, which upstream provider serves any individual request, and any further provider-side routing logic, are outside our control.

3.5. Measurement

Correctness is scored at three nested levels: the cell (one field in one record), the record (one transaction row of 11 fields), and the document (one ledger card of many records). Because a single scoring convention can itself be a source of dispute, every result in Section 4 is reported under three equivalence tiers and, for cell-level scores, three alignment strategies simultaneously, rather than a single chosen convention.

The three equivalence tiers are as follows: strict, literal token identity as written; normalized, type-aware normalization of whitespace, case, and thousands separators; and semantic, full type-aware equivalence—numeric fields compared by parsed value, date fields compared by parsed (day, month, Buddhist-era year) regardless of punctuation or spacing, and text fields compared after whitespace removal and case folding, with empty/placeholder values unified. The three alignment strategies are the following: positional, matching the i-th extracted record to the i-th gold record regardless of any structural mismatch; key-anchored, matching records via a composite key (reference number, lot number, and transaction date) using longest-common-subsequence alignment, so that a single dropped or inserted row does not misalign every subsequent record in the document; and key-anchored with missing-as-wrong, identical to key-anchored but additionally counting every gold record that has no extracted counterpart as fully incorrect across all 11 fields, so structural omissions are penalized rather than silently excluded from the denominator. All headline results in Section 4 use the semantic tier under key-anchored, missing-as-wrong alignment, the strictest of the nine combinations that still credits records displaced by a structural error rather than penalizing every field after the displacement; the full 3×3 sensitivity analysis is reported in Section 4.2.

Every metric is reported as a point value with a 95% confidence interval from document-level cluster bootstrap (resampling documents, not individual records, with replacement, B = 10,000), consistent with records nested within documents—records within the same document can share header fields and are therefore not statistically indepen-

dent, a dependency directly relevant to the header-error propagation pattern analyzed in Sections 4.2 and 5.1. We emphasize estimation and uncertainty over traditional significance testing; all analysis was performed in Python version 3.14 with the scoring script released alongside this paper.

Verifiability is measured with two, then five, deterministic accounting and calendar rules that a correct ledger record must satisfy, applied post hoc to each model's own extracted output with no human-keyed reference required at check time. Framed in the terms of Section 2.3, these are integrity constraints in the data-quality-assessment sense [16]: a flow rule (consistency constraint)—balance equals the previous balance plus received minus issued, tolerance 0.5—and a valuation rule (consistency constraint)—total value equals balance times unit cost, tolerance 0.5—form the numeric-only triage tested first; a set of date rules (well-formed Thai month token, plausible Buddhist-era year, non-decreasing chronological order within a document) are added afterward as an extended triage. Each rule is evaluated with the record flag rate, precision, recall, and negative predictive value (NPV—the share of unflagged, "passing" records that are actually correct, i.e., clean-pass accuracy) against the gold standard. Critically, the system prompt (Section 3.4) does not instruct the model to perform accounting arithmetic or any consistency check; the invariants are therefore an unbiased, deterministic post hoc validator of extraction fidelity, not a measure of model reasoning ability.

Cost is computed as cost per 100 records: AI cost is the sum of token charges actually billed by OpenRouter, and human cost is time spent multiplied by wage, using Thailand's average minimum wage of about USD 11.33/day (374 THB/day, 8 h day [47]) as a conservative lower-bound baseline. Reviewing a flagged record is costed, deliberately, at the same time and price as re-keying that record from scratch—the most conservative assumption available, since in practice confirming a flagged value is typically faster than a fresh full-record entry (Section 4.5). Because pharmaceutical stock-ledger accounting requires specialized, trained staff, true operational wages are substantially higher than the statutory minimum; the cost reduction reported here is therefore a lower bound, and the real economic gap would be larger than reported. Time is measured as processing time per document and per record and as throughput, separating image-rendering from inference for the AI side and using recorded times for the human side; we report both labor-time (the sum of both operators' time for double-entry) and wall-clock time (the shorter elapsed time when both operators work in parallel), since these differ by roughly a factor of two and conflating them would overstate one side of the comparison.

The scoring script that produces the accuracy, alignment and verifiability results reported in Section 4—the three equivalence tiers, the alignment strategies, the deterministic invariants, and the document-level cluster bootstrap—is released as a single script alongside this paper, together with the raw first-pass files from both human operators, the adjudicated gold standard, the main-run outputs of all four models, and the system prompt. Running that script on the released data reproduces the accuracy, structural-completeness, and gold-free triage results reported in Sections 4.2 and 4.4. The human-baseline figures in Section 4.1, and the model error taxonomy and representative error cases in Section 4.4, are obtained from the same released files by the procedures described in Sections 3.3 and 3.5, using the equivalence tiers released with the script. Two sets of figures are not reproducible from the released data: the run-to-run reliability results in Section 4.3, which require the repeated-run outputs, and the cost and latency figures in Section 4.5, which are measured from the provider's billing and usage logs recorded during the evaluation window (Data Availability Statement).

Figure 4 summarizes the complete workflow described across this section, from document ingestion through the human-in-the-loop review stage: page rendering, schema-

constrained extraction, gold-free invariant triage, selective human review of flagged records, and loading into the target database.

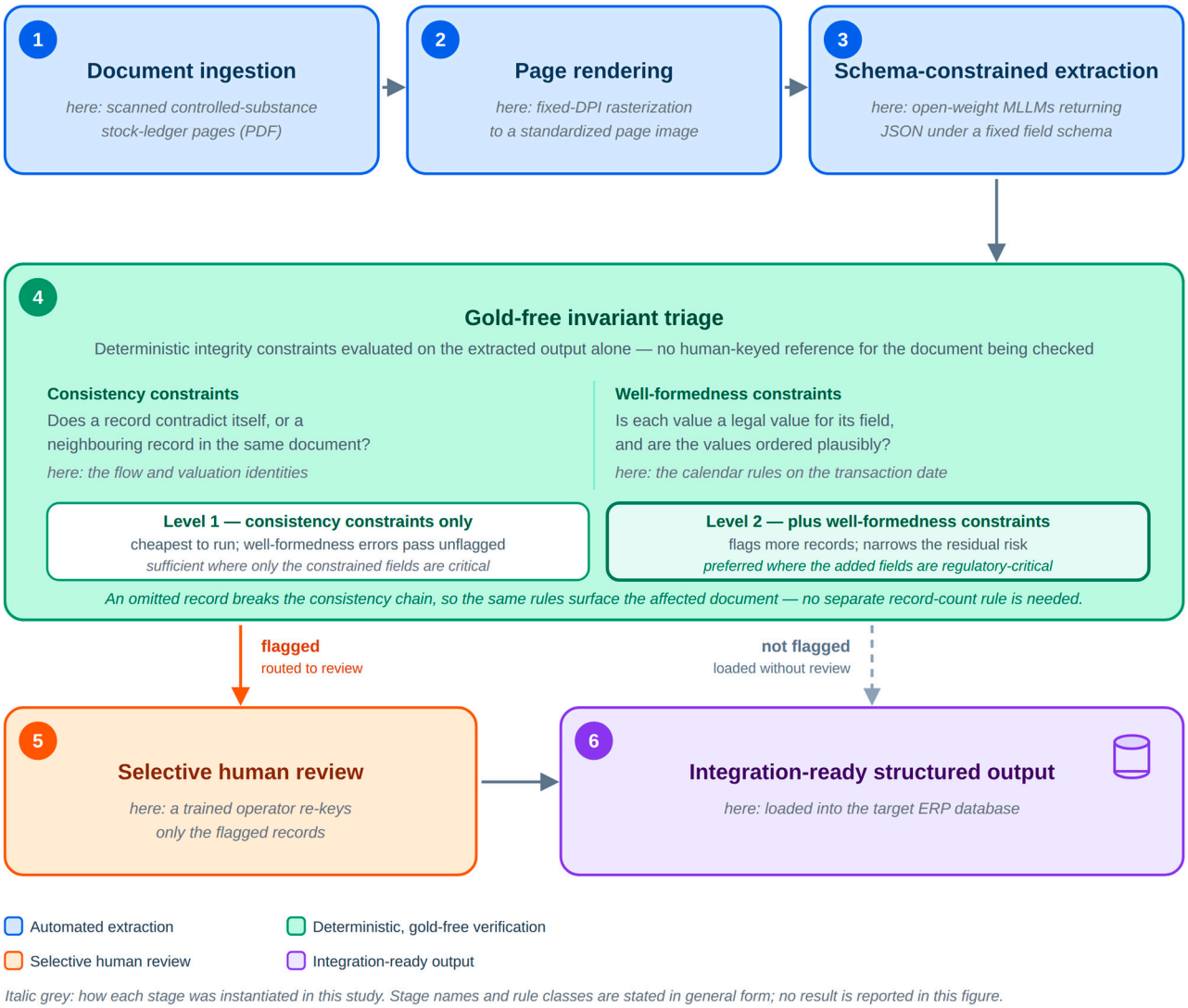


Figure 4. The extraction and verification workflow evaluated in this study, stated in general form; italic text records how each stage was instantiated here. Every triage rule is evaluated on the model's own output alone. Transfer conditions are discussed in Section 5.4.

3.6. Typhoon OCR Pilot: A Structural Anchor

As introduced in Section 1, the team piloted Typhoon OCR (model version typhoon-ocr-preview, the version available at the time of the pilot and since superseded) on ten documents from this corpus before building the MLLM pipeline evaluated in this paper, in both the tool's default (Markdown) and structure (HTML table) output modes. This pilot is reported here only as a structural, factual anchor for the motivation given in Section 1 and the literature gap identified in Section 2.1: inspecting the structure-mode HTML output directly confirms that the number of table cells per transaction row is not constant across rows of the same document, and that document-header metadata is fused into the same table as transaction rows via cell spanning. We do not report extraction accuracy for this pilot in the body of this paper, do not compare it against the four MLLMs evaluated in Section 4, and do not extend it to the full 400-document corpus; the sample size (N = 10) is a deliberate, permanent qualitative case, not a claim of statistical comparison. The parser, scoring script, and full numeric results for this pilot are released in

the accompanying repository for transparency but are not part of the manuscript's reported findings (Section 3.5).

4. Results

This section reports results in the order the three research questions require: the quality of the human gold standard against which every model is judged (Section 4.1); extraction accuracy itself, field by field and under the 2 × 2 vendor-by-architecture design (Section 4.2); the reliability of these numbers under repetition (Section 4.3); the performance of the gold-free invariant triage and the residual risk it leaves behind (Section 4.4); and the resulting throughput and full-workflow cost (Section 4.5). Unless stated otherwise, all point estimates are reported with 95% confidence intervals from the document-level cluster bootstrap described in Section 3.5 (B = 10,000).

4.1. Human Baseline and Gold-Standard Quality

Independent double-entry by the two trained operators, checked against the adjudicated gold standard, produced 9 erroneous cells for H1 and 5 for H2 out of 32,461 cells (99.972% and 99.985% cell accuracy, respectively). Table 4 gives the full taxonomy: every error on both sides is a transcription slip—a dropped decimal point, an extra digit, a single wrong digit—with no comprehension errors of the kind that would indicate an operator misunderstood the document. Comparing the two operators directly against each other, rather than each against gold, finds 14 cells where H1 and H2 disagree; the same 14 cells make up the union of both operators' errors against gold, and 0 cells are wrong in the same way for both operators. On this corpus, therefore, cross-checking the two operators against each other alone—without ever consulting the gold standard—would have caught all 14 errors: a 100% empirical double-entry detection rate. We report this as an observed property of this dataset, not a general guarantee: agreement between two people cannot, by construction, catch an error both of them make in the same way, and we simply did not observe such an error here (Section 5.4).

Table 4. Human error taxonomy: cell-level errors by field for each of the two independent operators (H1, H2) relative to the adjudicated gold standard.

Field	H1 Errors	H2 Errors	Nature of Error
Total value	4	4	off-by-decimal/extra digit (746,611.11 → 74,661,111; 919,909.57 → 0.54)
Reference	3	0	trailing digits of code wrong (WF69/00395/1 → 00398/1), suffix dropped
Lot no.	1	0	single-digit slip (269,097 → 269,094)
Unit cost	0	1	decimal point lost (9.91 → 991)
Issued	1	0	single-digit slip (8034 → 8031)
Total	9	5	all transcription slips; no comprehension errors

The resulting cell-level error rate (0.03–0.04%) is well below the roughly 1–4% often cited for general manual-entry tasks [4], a gap we attribute to task characteristics—cleanly printed source documents, predominantly numeric fields, operator self-checking, and routine work by trained staff—rather than to an implausibly error-free workforce (Section 3.3).

4.2. Extraction Accuracy

Table 5 reports semantic-tier, key-anchored (missing-as-wrong) per-field accuracy for all four models. Two fields stand out. Remarks, the only long free-text field, is the hardest field for every model (20.8–42.8%), reflecting the difficulty in transcribing extended, unconstrained Thai prose rather than a fixed-format value. Item name is near-ceiling for three of the four models (98.3–100.0%) but only 51.2% for Gemma-Dense (A2), driven by a

systematic header-propagation failure mode in which a unit token from a nearby header field overwrites the correct unit in the item-name string (e.g., "Analgesic cream: 30 g" extracted as "Analgesic cream: 30 mg"); this single failure mode, not a general weakness on text fields, accounts for most of A2's field-accuracy gap.

Table 5. Per-field accuracy (%, [95% CI]), semantic equivalence tier, key-anchored alignment with missing records scored as wrong.

Field	Type	A1 Gemma-MoE	A2 Gemma-Dense	A3 Qwen-MoE	A4 Qwen-Dense
Item name	text	99.2 [98.4, 99.7]	51.2 [46.2, 56.6]	98.3 [97.8, 98.8]	100.0 [100.0, 100.0]
Transaction date	date	89.9 [86.6, 92.8]	93.5 [91.0, 95.7]	89.6 [86.9, 92.0]	91.7 [89.1, 94.1]
Reference	text	98.7 [97.8, 99.5]	98.0 [96.8, 99.1]	97.5 [96.6, 98.3]	99.9 [99.7, 100.0]
Lot no.	text	97.3 [95.8, 98.6]	99.1 [98.2, 99.8]	97.8 [97.0, 98.5]	100.0 [100.0, 100.0]
Unit cost	num	89.7 [86.8, 92.2]	84.3 [80.7, 87.7]	98.0 [97.4, 98.5]	99.2 [98.4, 99.8]
Received	num	99.3 [98.6, 99.8]	99.9 [99.8, 100.0]	97.7 [96.9, 98.5]	100.0 [100.0, 100.0]
Issued	num	99.3 [98.9, 99.7]	99.9 [99.8, 100.0]	97.0 [95.9, 98.0]	100.0 [100.0, 100.0]
Balance	num	95.0 [92.9, 96.9]	98.5 [97.6, 99.3]	96.1 [94.7, 97.3]	100.0 [100.0, 100.0]
Total value	num	75.3 [72.2, 78.4]	94.1 [92.4, 95.6]	96.3 [95.1, 97.5]	99.8 [99.5, 100.0]
Signature	text	99.7 [99.4, 100.0]	100.0 [100.0, 100.0]	98.3 [97.8, 98.8]	100.0 [100.0, 100.0]
Remarks	text	20.8 [18.3, 23.3]	39.2 [36.1, 42.3]	36.6 [33.5, 39.9]	42.8 [39.8, 45.8]

Table 6 reports overall cell accuracy under all nine tiers × alignment combinations, to guard against the appearance of having selected the single most favorable scoring convention. Across all nine combinations, the spread between the strictest and the most permissive convention is 0.1–0.8 percentage points, and the ranking of the four models never changes. The headline results below therefore use the semantic tier under key-anchored, missing-as-wrong alignment—the strictest of the nine combinations that still credits records displaced by a structural error.

Table 6. Overall cell accuracy (%, [95% CI]) under three equivalence tiers × three alignment strategies.

Tier	Alignment	A1 Gemma-MoE	A2 Gemma-Dense	A3 Qwen-MoE	A4 Qwen-Dense
strict	positional	87.58 [86.80, 88.31]	86.71 [86.01, 87.39]	90.17 [89.20, 91.04]	93.19 [92.79, 93.57]
strict	key	87.68 [86.92, 88.39]	86.71 [86.01, 87.39]	92.08 [91.56, 92.58]	93.19 [92.79, 93.57]
strict	key + penalized	87.56 [86.78, 88.30]	86.71 [86.01, 87.39]	90.49 [89.65, 91.26]	93.19 [92.79, 93.57]
normalized	positional	87.58 [86.80, 88.31]	86.71 [86.01, 87.40]	90.17 [89.20, 91.04]	93.19 [92.79, 93.57]
normalized	key	87.68 [86.92, 88.39]	86.71 [86.01, 87.40]	92.08 [91.56, 92.58]	93.19 [92.79, 93.57]
normalized	key + penalized	87.56 [86.78, 88.30]	86.71 [86.01, 87.40]	90.49 [89.65, 91.26]	93.19 [92.79, 93.57]
semantic	positional	87.66 [86.89, 88.40]	87.07 [86.37, 87.75]	90.88 [89.91, 91.76]	93.95 [93.53, 94.35]
semantic	key	87.77 [87.01, 88.48]	87.07 [86.37, 87.75]	92.80 [92.26, 93.31]	93.95 [93.53, 94.35]
semantic	key + penalized	87.65 [86.87, 88.39]	87.07 [86.37, 87.75]	91.20 [90.34, 91.99]	93.95 [93.53, 94.35]

Under this headline convention, Qwen-Dense (A4) reaches 93.95% [93.53, 94.35] overall cell accuracy, the highest of the four models, followed by Qwen-MoE (A3) at 91.20% [90.34, 91.99], Gemma-MoE (A1) at 87.65% [86.87, 88.39], and Gemma-Dense (A2) at 87.07% [86.37, 87.75]. Table 7 reports the corresponding record- and document-level structural results: record-level exact match, which requires all 11 fields of a transaction row to be simultaneously correct, ranges from 14.0–14.5% for the two Gemma models to 33.7–40.1% for the two Qwen models—far lower than any per-field number—because a single wrong field anywhere in an 11-field row is enough to fail the whole record. Structurally, both dense models (A2, A4) reproduce the full record count in all 400 documents with zero missing or spurious records; Gemma-MoE (A1) drops records in 4 of 400 documents (4 missing, 1 spurious), and Qwen-MoE (A3) drops records in 43 of 400 documents (51 missing records total).

Table 7. Record-level exact match and document-level structural completeness by model.

Model	Record Exact-Match %	Docs Structurally OK/400	Missing Records	Spurious Records
A1 Gemma-MoE	14.0 [11.9, 16.2]	397	4	1
A2 Gemma-Dense	14.5 [11.6, 17.5]	400	0	0
A3 Qwen-MoE	33.7 [30.5, 36.9]	357	51	0
A4 Qwen-Dense	40.1 [36.9, 43.3]	400	0	0

Figure 5 and, read jointly, Tables 5–7 support a 2 × 2 conclusion rather than a single claim that dense models are simply "more accurate" than Mixture-of-Experts models. On field-level accuracy, architecture does not act uniformly across vendors: within Google's models, Gemma-MoE slightly outperforms Gemma-Dense (87.65% vs. 87.07%, +0.6 pp); within Alibaba's models, Qwen-Dense outperforms Qwen-MoE by a wider margin (93.95% vs. 91.20%, +2.8 pp). This is an interaction between vendor and architecture, not a main effect of architecture on accuracy. On structural completeness, by contrast, the two dense models never drop a record in any of the 400 documents, while both Mixture-of-Experts models do, at very different rates (4 documents affected for Gemma-MoE, 43 for Qwen-MoE)—the same direction for both vendors. We therefore report field-level accuracy as varying more strongly across vendors, and structural completeness as differing consistently between the dense and Mixture-of-Experts models; the mechanism proposed for this pattern is developed in Section 5.2.

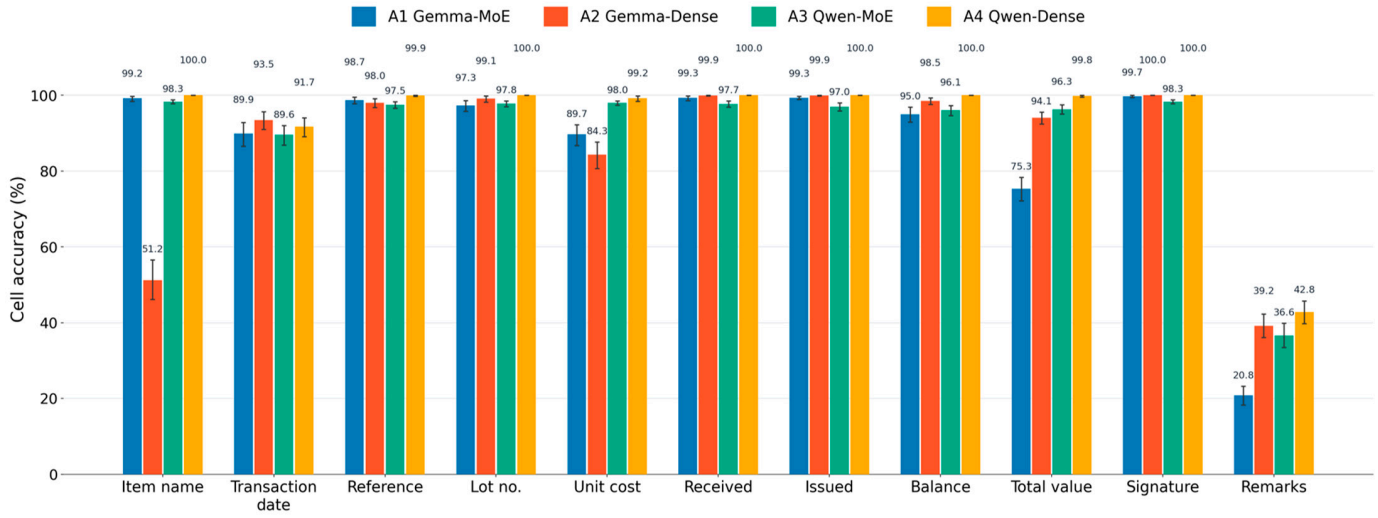


Figure 5. Per-field accuracy with 95% confidence intervals, by model and data type.

4.3. Run-to-Run Variability

The repeated-runs design (Section 3.1) processed 50 documents three additional times per model under the identical pipeline, prompt, and API access as the main run, to test whether the single-pass main-run numbers above are representative of model behavior under repetition. Table 8 reports two distinct standard deviations for each model, deliberately kept separate rather than pooled: the SD within the three repeat runs alone, and the SD when the main-study result on the same 50 documents is added as a fourth independent sample. Every model, including the main run, raises the SD—for Gemma-Dense (A2), from 1.01 pp within the three repeat runs to 2.00 pp once the main run is included, the largest shift of the four models (Figure 6a). Because the pipeline, prompt, and API access were confirmed unchanged across all four samples, we interpret this as evidence of model-level stochastic variance that is larger than the document-level cluster bootstrap confidence interval alone would suggest, rather than as evidence of setup drift.

Table 8. Run-to-run variability in overall accuracy: standard deviation and range within the 3 repeat runs alone, and including the main-study result as a fourth independent sample.

Model	SD Within 3 Repeat Runs (pp)	Range Within 3 Runs (pp)	SD Include Main-Study as 4th Sample (pp)	Range Include Main-Study (pp)
A1 Gemma-MoE	0.22	0.53	1.77	4.35
A2 Gemma-Dense	1.01	2.44	2.00	5.47
A3 Qwen-MoE	0.35	0.81	1.05	2.67
A4 Qwen-Dense	0.08	0.20	0.09	0.23

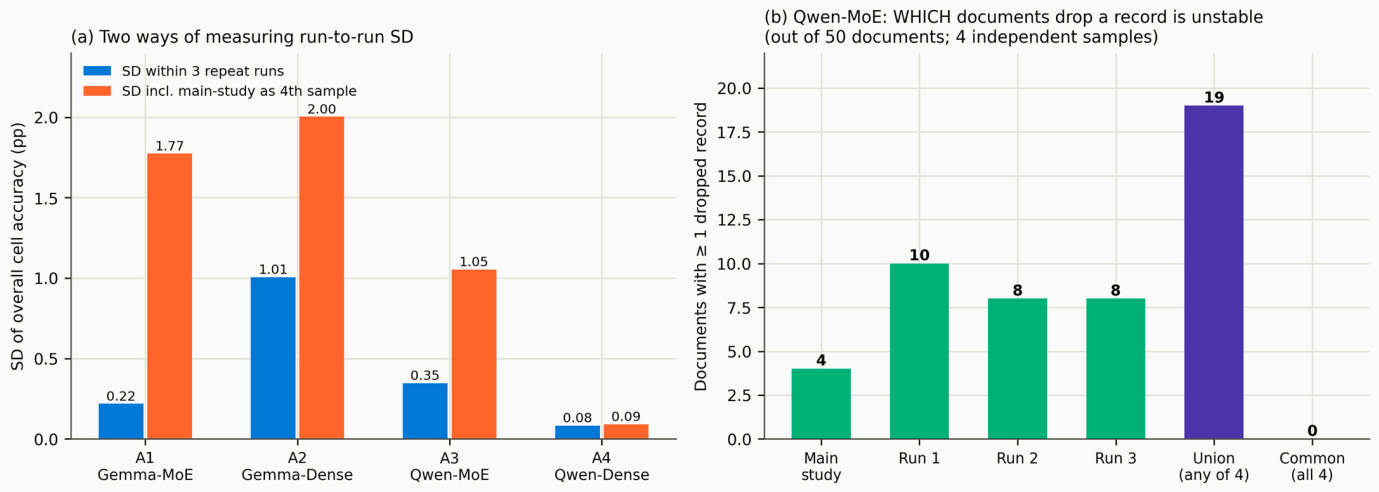


Figure 6. Run-to-run reliability of the verification layer. (**a**) Standard deviation of overall accuracy under the two computations in Table 8, by model. (**b**) Overlap, across the four independent samples, of documents with at least one missing record for Qwen-MoE (union = 19, intersection = 0).

Structural completeness is likewise not a one-off artifact. Table 9 shows that Qwen-MoE (A3) drops at least one record in every one of the four independent samples—the main run and all three repeat runs—while both dense models (A2, A4) never drop a record in any of the four. The specific documents affected are largely, though not entirely, distinct from run to run: the union of documents with at least one missing record across all four independent samples is 19 out of 50, while the intersection—documents affected in every single sample—is 0 (Figure 6b). This pattern is more consistent with a largely stochastic drop location than with a fixed subset of structurally difficult documents. Critically, the gold-free invariant triage (Section 4.4) still catches every affected document in every sample: 4/4 in the main run, and 10/10, 8/8, and 8/8 in the three repeat runs, respectively—a 100% document-level detection rate reproduced independently four times, not a property of one lucky draw.

Table 9. Documents with at least one missing record, by model and by run (main study plus 3 repeat runs).

Model	Missing Records: Main Study (on 50-Doc Subset)	run1	run2	run3
A1 Gemma-MoE	0	0	0	0
A2 Gemma-Dense	0	0	0	0
A3 Qwen-MoE	4	10	8	8
A4 Qwen-Dense	0	0	0	0

4.4. Gold-Free Verification and Residual Risk

The deterministic invariants introduced in Section 3.5 are, on the gold standard itself, exact: the flow rule holds for 2551 of 2551 applicable records (100.00%), the valuation rule

holds for 2941 of 2941 applicable records (100.00%), and the date rules hold for 2951 of 2951 records (100.00%). The rules are therefore valid characterizations of a correct ledger record on this corpus, not noisy heuristics, before they are ever applied to model output.

Applied to each model's own extracted output, the numeric-only triage (flow and valuation rules) achieves its best result on the model it is meant to protect against error: for Qwen-Dense (A4), only 0.61% of records are flagged for review, at 100.00% precision and 64.29% recall, leaving a negative predictive value (NPV) of 99.66%—records that pass the triage unflagged are correct on the numeric fields 99.66% of the time. Table 10 reports the full numeric-triage results for all four models.

Table 10. Gold-free invariant triage using numeric-only rules (flow, valuation): flag rate, precision, recall, and negative predictive value against the gold standard.

Model	Flagged %	Precision %	Recall %	NPV (Numeric) %	tp/fp/tn/fn
A1 Gemma-MoE	21.17	99.04	73.57	90.44	618/6/2101/222
A2 Gemma-Dense	17.69	98.47	91.79	98.11	514/8/2383/46
A3 Qwen-MoE	3.21	62.37	63.04	98.79	58/35/2773/34
A4 Qwen-Dense	0.61	100.00	64.29	99.66	18/0/2923/10

Extending the triage with calendar rules (Table 11) raises Qwen-Dense's residual date accuracy from 92.1% to 95.9%, at the cost of raising its flag rate from 0.61% to 4.61%. We report this trade honestly rather than selling it: the extended rules help, but do not close the residual-risk gap, and the choice of which triage level to deploy is a genuine cost–accuracy decision rather than a free improvement.

Table 11. Extended gold-free triage: numeric invariants plus calendar (date) rules. No reference to the gold standard is required at check time.

Model	Flagged %	Precision %	Recall %	NPV (Numeric + Date) %	Residual Date Accuracy %
A1 Gemma-MoE	23.69	99.28	66.51	84.48	93.5
A2 Gemma-Dense	17.82	98.67	73.41	92.25	94.1
A3 Qwen-MoE	8.10	85.96	63.32	95.61	96.3
A4 Qwen-Dense	4.61	100.00	52.11	95.56	95.9

The residual accuracy that matters operationally is not the NPV of the numeric rules alone but the accuracy across all 11 fields of the records that pass triage unflagged—the records that would enter the database untouched under a deployment that reviews only flagged output. For Qwen-Dense (A4), this residual accuracy is 94.1% across all 11 fields while flagging 0.61% of records for human review (Table 12), and 94.5% while flagging 4.61% under the extended triage (Table 13). Both figures are the disclosed residual risk of the workflow at their respective operating points. This is not an unqualified accuracy claim: it is the accuracy of the unflagged majority, explicitly conditioned on its flag rate, and the two weakest fields—Transaction date (92.1%) and Remarks (43.0%)—remain the dominant contributors to the residual 5.9% error because neither field is covered by the numeric-only triage rules.

The invariants' most consequential capability is not numeric-error detection but structural-omission detection—catching the dropped records identified in Sections 4.2 and 4.3 without ever comparing against a gold record count. For Qwen-MoE (A3), the model with the largest number of dropped records, the invariant triage flags at least one violation in 43 of the 43 documents that actually have a dropped record (100%), against only 9 of the 357 documents that do not (3%)—a document-level sensitivity of 100% and specificity of 97.5% (Table 14). Gemma-MoE (A1) shows the same directional pattern (4 of

4 affected documents flagged) but with much less discriminative power because 58% of its unaffected documents are already flagged for other reasons, given its higher baseline numeric-error rate; we do not generalize the clean Qwen-MoE separation to every model. Neither dense model drops any records, so structural-omission detection does not apply to A2 or A4. This detection capability replicates across all four independent samples in Section 4.3, not just the single main run.

Table 12. Residual per-field accuracy (%) among records that pass the numeric-only gold-free triage unflagged.

Field	A1 Gemma-MoE	A2 Gemma-Dense	A3 Qwen-MoE	A4 Qwen-Dense
Item name	99.9	46.3	100.0	100.0
Transaction date	90.6	93.9	91.4	92.1
Reference	99.0	97.8	99.4	99.9
Lot no.	98.0	99.6	99.8	100.0
Unit cost	93.9	98.2	99.7	99.7
Received	100.0	100.0	100.0	100.0
Issued	99.8	100.0	99.5	100.0
Balance	98.9	100.0	99.2	100.0
Total value	90.9	98.6	99.1	100.0
Signature	99.9	100.0	100.0	100.0
Remarks	21.8	40.4	37.8	43.0
All 11 fields	90.2	88.6	93.3	94.1

Table 13. Residual per-field accuracy (%) among records that pass the extended gold-free triage unflagged (numeric invariants plus calendar rules); the counterpart of Table 12.

Field	A1 Gemma-MoE	A2 Gemma-Dense	A3 Qwen-MoE	A4 Qwen-Dense
Item name	100.0	46.2	100.0	100.0
Transaction date	93.5	94.1	96.3	95.9
Reference	99.0	97.9	99.8	99.9
Lot no.	98.1	99.6	100.0	100.0
Unit cost	93.7	98.2	99.7	99.6
Received	100.0	100.0	100.0	100.0
Issued	99.8	100.0	99.8	100.0
Balance	98.9	100.0	99.7	100.0
Total value	90.9	98.6	99.5	100.0
Signature	99.9	100.0	100.0	100.0
Remarks	21.9	40.5	38.0	43.7
All 11 fields	90.5	88.6	93.9	94.5
All except Remarks	97.4	93.5	99.5	99.5
Records passing triage (of 2951)	2249	2425	2665	2815

Table 14. Structural-omission detection: invariant flag rate on documents with vs. without at least one dropped record.

Model	Docs with Dropped Rows		Docs Without Drops		Row After Drop Flagged
	n	≥1 Invariant Flag	n	≥1 Invariant Flag	
A1 Gemma-MoE	4	4 (100%)	396	231 (58%)	2/4 (50%)
A2 Gemma-Dense	0	—	400	—	—
A3 Qwen-MoE	43	43 (100%)	357	9 (3%)	31/51 (61%)
A4 Qwen-Dense	0	—	400	—	—

Figure 7 plots this trade-off—unflagged-record accuracy against the percentage of records flagged for review—across all four models and both triage levels (Tables 10 and 11). Table 15 characterizes the specific errors underlying these numbers (semantic tier, key-anchored). Remarks misreadings dominate the error count for every model (1582–1750 instances), consistent with Remarks being the hardest field in Table 5. Item-name misreadings are concentrated almost entirely in Gemma-Dense (1441 of the 1462 total instances across all models), matching the header-propagation pattern described in Section 4.2. Among date errors, invalid-month-token errors—the model producing a three-letter Thai month abbreviation that does not correspond to any of the twelve valid tokens—are, by construction, catchable with 100% certainty by a deterministic well-formedness check, unlike month-confusion (Table 16) errors that substitute one valid month for another; this asymmetry is the empirical justification for including a date well-formedness rule in the extended triage above.

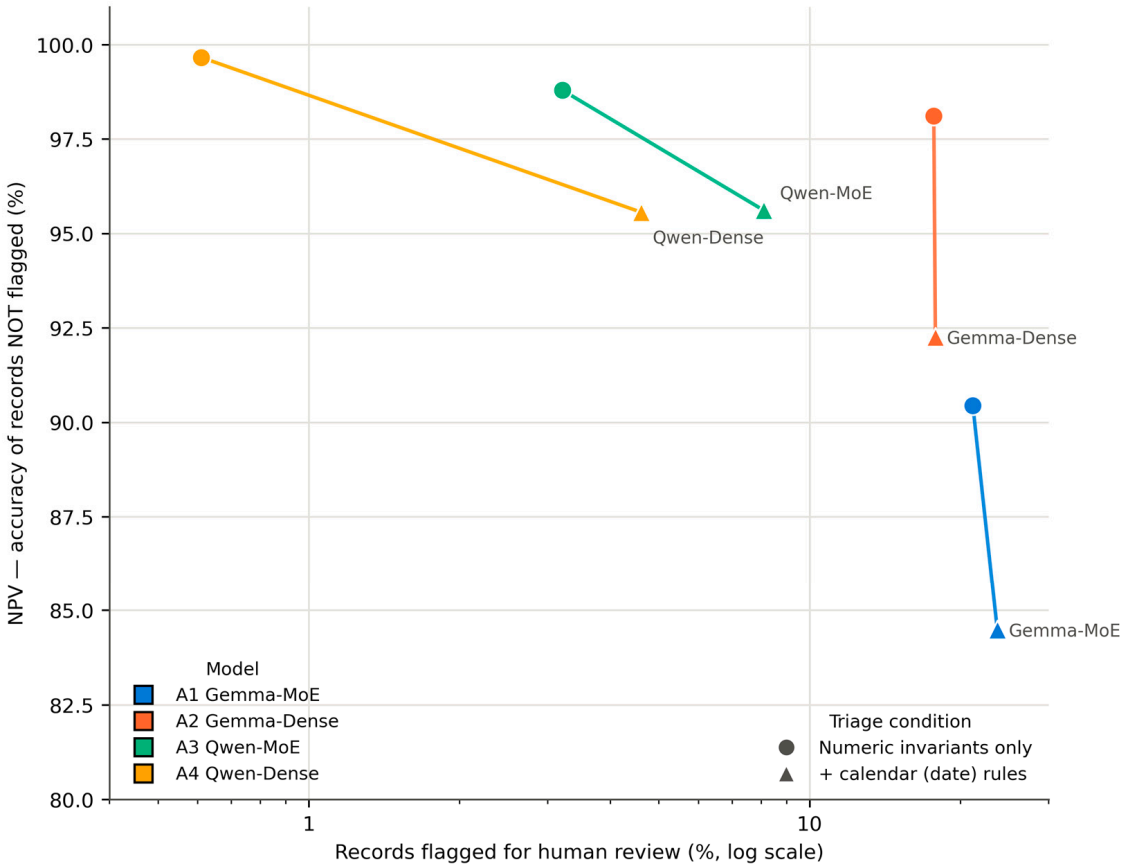


Figure 7. Residual risk versus review burden: unflagged-record accuracy against the percentage of records flagged for human review, across models and triage levels.

Table 15. Model error taxonomy by category and field (semantic tier, key-anchored alignment).

Error Category	A1 Gemma-MoE	A2 Gemma-Dense	A3 Qwen-MoE	A4 Qwen-Dense
Remarks—misread	1582	1750	1696	1619
Remarks—omitted	676	3	7	29
Remarks—truncated/extended	75	42	116	40
Item name—misread	21	1441	0	0
Transaction date—month	216	191	102	126
Transaction date—unparseable	79	2	152	118
Total value—value	724	174	57	5
Unit cost—value	292	452	0	13
Balance—value	143	43	65	0

Table 16. Representative extraction errors by type, with gold-standard and model-output examples.

Case	Model	Result	Note
Misread (text)	A2 Gemma-Dense	In the Remarks field, the gold text read "QC 069/68 เบิกเพื่อวิเคราะห์ Pseudoephedrine HCl Powder" (*boek phuea wikhro*, "withdrawn for analysis"). The model output "QC 069/68 เปิดเพื่อวิเคราะห์ Pseudoephedrine HCl Powder" (*poet phuea wikhro*, "opened for analysis") instead.	The Thai verb เบิก (*boek*, "to withdraw") was misread as เปิด (*poet*, "to open"). The two words differ only in their initial consonant (บ vs. ป) and final consonant (ก vs. ด), each pair being closely similar in shape, so the substitution is orthographically plausible; it nevertheless changes the meaning of the record rather than only its formatting.
Omitted (text)	A1 Gemma-MoE	In the Remarks field, the gold text read "Pseudoephedrine HCl 60 mg Tablet: 100 × 10's." The model output a placeholder dash (-) with no content.	A long free-text field was dropped entirely rather than partially transcribed, matching this model's broader pattern of 676 Remarks omissions (Table 15).
Header-propagation (item name)	A2 Gemma-Dense	In the Item name field, the gold text read "Analgesic cream: 30 g." The model output "Analgesic cream: 30 mg" instead.	A unit token from a nearby header field overwrote the correct unit, changing grams to milligrams, matching this model's broader pattern of 1441 item-name errors (Table 15).
Month-confusion (date)	A1 Gemma-MoE	In the Transaction date field, the gold date was 9 เม.ย. (9 April) on the Thai calendar. The model output 9 มิ.ย. (9 June) instead.	The Thai abbreviation for April (เม.ย.) was misread as the abbreviation for June (มิ.ย.). Both are valid month tokens, so this error would not be caught by a well-formedness check alone.

Thai text is reproduced verbatim from the source documents because the errors illustrated here are orthographic and cannot be demonstrated in translation. Romanization follows the Royal Thai General System of Transcription; English glosses appear in quotation marks.

Table 16 gives one concrete case for each of the four major error categories in Table 15, describing the gold value and the model's output in place of the image-based qualitative figure used in earlier drafts of this work—a table communicates the same evidence without the redaction and resolution concerns that accompany publishing crops of real controlled-substance ledger images.

4.5. Throughput and Full-Workflow Cost

Table 17 reports the cost and time of the complete proposed workflow—extraction plus gold-free triage plus human review of every flagged record—against human double-entry on the same corpus (USD 8.73 per 100 records; 221.8 labor-seconds per record; 114.9 wall-clock-seconds per record when both operators work in parallel), under the deliberately conservative assumption that reviewing one flagged record costs as much time and money as re-keying that record from scratch. Under Qwen-Dense's numeric-only triage (0.61% flagged), the workflow costs USD 0.041 per 100 records—a 99.53% cost reduction—and completes 56.7× faster on a labor-time basis and 29.4× faster on a wall-clock basis. Under the extended triage (4.61% flagged), the reduction is 97.53% and the speed-up is 26.1× (labor)/13.5× (wall-clock); the two triage levels therefore bracket a wall-clock speed-up of 13.5–29.4×. Even in the deliberately unrealistic worst case of reviewing every single record, the workflow remains 49.84% cheaper than human double-entry and no slower (1.9× labor-time, 1.0× wall-clock), so the qualitative business conclusion does not reverse, even under the most pessimistic assumption tested.

We report both labor-time and wall-clock speed-ups throughout because they differ by roughly a factor of two—the wall-clock figure reflects the two human operators working in parallel, and conflating the two would overstate the comparison on whichever side is measured only in labor-time.

Table 17. Full-workflow cost and throughput against human double-entry: complete best-to-worst-case range for A4 (the recommended model), plus one baseline comparison point per remaining model.

Model	Review Level	Flag %	USD/100 rec	Cost Reduction vs. Double-Entry	s/rec	Speed-Up (Labor)	Speed-Up (Wall-Clock)
A4 Qwen-Dense	numeric	0.61	0.041	99.53	3.91	56.7×	29.4×
A4 Qwen-Dense	extended	4.61	0.215	97.53	8.51	26.1×	13.5×
A4 Qwen-Dense	every record	100	4.379	49.84	118.1	1.9×	1.0×
A3 Qwen-MoE	numeric	3.21	0.159	98.18	7.03	31.6×	16.3×
A3 Qwen-MoE	extended	8.10	0.372	95.73	12.65	17.5×	9.1×
A2 Gemma-Dense	numeric	17.69	0.782	91.04	28.41	7.8×	4.0×
A1 Gemma-MoE	numeric	21.17	0.932	89.33	28.83	7.7×	4.0×

5. Discussion

This section interprets the Section 4 results against the research questions posed in Section 1: what the combined accuracy, cost, and verifiability results imply for a practical deployment decision (Section 5.1); how the vendor and architecture effects identified in Section 4.2 point to a candidate mechanism rather than a simple accuracy ranking (Section 5.2); the specific scope of what gold-free invariant verification can and cannot guarantee (Section 5.3); and the limitations and conditions under which these single-site findings should be expected to transfer (Section 5.4).

5.1. What the Numbers Mean for Practice

Taken together, Section 4 supports a specific, bounded answer to the overarching question posed in Section 1: open-weight MLLMs can convert this class of legacy ERP document into integration-ready data at a level of accuracy, cost, and verifiability sufficient to augment, not replace, human data entry, provided the deployment includes the gold-free triage layer rather than trusting raw model output directly. This bounded answer is deliberately narrow: it holds for the fields the triage layer can check—numeric-relational and date fields—and explicitly excludes the free-text Remarks field, where even the extended triage leaves only 43.7% field accuracy on unflagged records (Table 13) and where no gold-free check currently exists (Section 5.3). We do not claim the pipeline is sufficiently reliable for unrestricted regulatory use across all 11 fields; we claim it reduces and localizes the human review burden for the fields it can verify, while free-text fields require continued full human review regardless of triage outcome. This scope also reflects the limitations in Section 5.4: the comparison covers four specific open-weight models under a fixed prompt, and prompt-sensitivity and an OCR-plus-template baseline were not tested. The strongest configuration in this study, Qwen-Dense (A4), reaches 93.95% raw cell accuracy, drops zero records across all 400 documents, and leaves a workflow in which 0.61% of records require human review under the numeric-only triage, or 4.61% under the extended triage, while the unflagged 99.39% remain 94.1% accurate across all 11 fields—a small, disclosed residual error rate rather than an implicit claim of perfection.

This workflow is simultaneously the cheapest and the fastest option tested, at 13.5–29.4× the wall-clock speed and 97.5–99.5% the cost of human double-entry, and the qualitative conclusion survives even the deliberately pessimistic assumption of reviewing every single record (Section 4.5). The practical reading is not "MLLMs are accurate enough to trust," which the field-level numbers alone would not support, but "MLLMs plus a cheap, deterministic screening layer are accurate enough to deploy," which is a materially different and more defensible claim.

This reframing matters because the field-level numbers alone are uneven: near-ceiling on structured numeric fields, weaker on free text (Remarks, 20.8–42.8%), and never perfect on dates. A deployment that judged trustworthiness from an aggregate accuracy figure

would be blind to exactly this unevenness. The triage layer is what converts an uneven but strong model into a workflow with a stated, bounded risk: the central practical contribution of this paper, developed conceptually next.

Choosing between the two triage levels is an explicit risk decision rather than a free improvement. The numeric-only level sends 0.61% of records to review and accepts a residual transaction-date error rate of 7.9%; the extended level sends 4.61% and reduces that to 4.1%, at a residual accuracy of 94.5% across all 11 fields and 99.5% across the ten fields other than the unconstrained Remarks column. Because Remarks-field errors are invisible to the triage by construction (Section 5.3), the 43.0% Remarks accuracy is not reduced by either triage level and should be treated as a standing requirement for full human review of that field, independent of the flagged/unflagged split reported here. In absolute terms, 1919 of the 32,461 extracted cells reach the database uncorrected under the numeric-only level, against 1713 under the extended level. Where dates carry regulatory weight, as in controlled-substance recordkeeping, the extended level is the appropriate default; where only the numeric fields are consequential, the numeric-only level is sufficient and roughly eight times cheaper to review.

5.2. Vendor, Architecture, and Structural Completeness

Section 4.2 reported an interaction, not a uniform ranking: the direction of the dense/Mixture-of-Experts difference in field-level accuracy reverses between vendors (Gemma-MoE edges out Gemma-Dense by 0.6 pp; Qwen-Dense outperforms Qwen-MoE by 2.8 pp), whereas the same contrast in structural completeness is consistent across vendors: both dense models never drop a record, both Mixture-of-Experts models do. The two properties therefore vary along different axes in this sample: field-level accuracy tracked the vendor, whereas the dense/Mixture-of-Experts distinction was the only factor separating the models that preserved a document's full record count from those that did not. Because architecture is confounded with active parameter capacity in this design (Section 5.4), we report this as an observed pattern among the four evaluated models rather than as an isolated architectural effect. A single claim that dense models are simply more accurate than Mixture-of-Experts models would not be supported by the vendor-reversed field-accuracy result.

A plausible mechanism for the structural-completeness half of this pattern comes from recent work on multimodal Mixture-of-Experts models, which finds that visual input can induce token routing to drift away from the mid-layer experts that the same content would engage as text, a "seeing but not thinking" failure in which the model perceives visual content correctly yet reasons over it poorly, termed routing distraction [33]. If page-image tokens near a document's table boundaries are routed away from the experts responsible for maintaining row-by-row structural continuity, a plausible symptom is exactly what we observe: an entire transaction row silently missing from the output, rather than a row present but locally garbled. We offer this as a candidate explanation consistent with the literature, not a mechanism we directly instrument in this study; the routing internals of the evaluated models are not observable through the OpenRouter API used here (Section 3.4). Two observations support treating this as architecture-linked rather than incidental: the drop rate is non-zero for both Mixture-of-Experts models and zero for both dense models across the main run and all three repeated runs (Section 4.3), and it holds despite the two Mixture-of-Experts models differing substantially in overall field accuracy, a pattern more consistent with a shared architectural property than with a shared vendor-specific weakness (Section 2.2).

The practical consequence is favorable rather than purely negative. Because the invariant triage flags 100% of the documents affected by a dropped record for Qwen-MoE,

and reproduces that detection rate independently in all four samples in Section 4.3, a record that a Mixture-of-Experts model never extracts is not invisible to every downstream check: it is exactly the kind of failure the flow-rule invariant catches at the document level, without ever consulting a gold record count. This reverses the intuition that an unextracted record would be invisible to every downstream check, and that silent omissions are therefore more dangerous than value errors; the data instead show silent omissions to be the failure mode this triage layer catches most reliably.

5.3. The Reach and the Limits of Gold-Free Verification

The triage layer performs two structurally different jobs, and it is important not to conflate them. On numeric-relational consistency (the flow and valuation rules), it is close to definitive for the fields it covers: for Qwen-Dense, records that pass unflagged are correct on those specific rules essentially always, and the residual 11-field accuracy of 94.1% reflects mainly the fields the numeric rules do not check (Transaction date, Remarks) rather than failures of the rules themselves. On structural omission, it is close to definitive at the document level, catching every Qwen-MoE document with a dropped record across all four independent samples tested (Sections 4.3 and 4.4). Extending it to date well-formedness closes part, not all, of the remaining gap: residual date accuracy rises from 92.1% to 95.9% at the cost of a more than sevenfold increase in flag rate (0.61% to 4.61%), a trade we report as a genuine cost–accuracy decision rather than a free improvement (Section 4.4). No rule in this study checks Remarks, the field responsible for the largest share of the residual error, because free text has no deterministic internal constraint analogous to an accounting identity or a valid month token.

This split (strong where an internal constraint exists, weak where it does not) is the concrete, task-level shape of the capability unevenness described abstractly in Section 1 as a jagged frontier of model reliability, extended there from its original setting in open-ended knowledge work to structured document extraction. The contribution of this paper is not a claim that MLLMs are uniformly reliable extractors, which the field-level numbers directly contradict, nor a claim that no automated signal can help, which the triage results directly contradict. It is a demonstration that the region of the frontier where a given model's output can be trusted without a human-labeled reference for that specific document is itself identifiable, deterministically and cheaply, from properties of the extracted output alone (flow and valuation consistency, date well-formedness, and structural omission surfaced by the flow rule) whenever the underlying document type provides such internal constraints. Where it does not, as with unconstrained free text, the frontier's jagged edge remains invisible to this class of gold-free check, and broader human review is the honest fallback (Section 5.4).

5.4. Limitations and Transfer Conditions

Several limitations bound the external validity of these findings. The study is a single site, a single document type, and Thai-language only, so generalization beyond this setting is untested. In the 2×2 design, architecture is confounded with active parameter capacity: although total parameter counts are roughly matched within each vendor pair (26B vs. 31B for Google, 30B vs. 32B for Alibaba), the Mixture-of-Experts models activate only about 3–4B parameters per token against roughly 31–32B for the dense models, an 8–10$\times$ active-capacity gap, so the dense/Mixture-of-Experts differences reported here cannot be attributed to architecture alone and may partly reflect this gap; the comparison also reflects only the specific open-weight models available at the time of the study and should not be read as a general law that dense architectures always preserve structure better than Mixture-of-Experts architectures. All models were accessed through the OpenRouter API as a standardized measurement har-

ness (Section 3.4): provider-side routing and serving configuration were not controlled, and local replication at a known precision was not performed. The study establishes performance only under the serving configurations observed during the evaluation period; replication under alternative serving configurations is left to future work. The cost and latency figures in Section 4.5 are likewise measured through this particular API and could differ elsewhere.

The gold-free invariants cover numeric-relational consistency, date well-formedness, and structural omission surfaced by the flow rule; they do not cover free-text fields, so Remarks-field errors are by construction never caught by triage and remain a residual risk that only broader human review (or an as-yet-undeveloped free-text check) can close. The gold standard itself comes from one institution's two operators; while its construction and the double-entry detection-rate analysis are described in detail (Section 4.1), we cannot rule out an error pattern that both operators would make identically, which by definition neither operator agreement nor this study would detect (Section 5.3). Two further analyses were outside the scope of this study. The system prompt was held constant across all models by design, so we did not test whether the comparative conclusions hold under alternative prompts; a prompt-sensitivity analysis is left to future work. We also did not run a quantitative OCR-plus-template baseline: the Typhoon OCR pilot (N = 10, Section 3.6) is a qualitative structural anchor only, and this paper makes no claim of superiority over OCR-plus-template approaches.

Specifying the conditions under which these findings should and should not be expected to transfer is, for a single-site case study, the more defensible response than an unqualified generalization claim [48]. We expect the approach to transfer to settings that share two properties with this one: documents with repeated, structured fields (copy-heavy in the sense that makes table-based extraction efficient [26]) and the presence of at least one internal domain invariant, such as an accounting identity or a well-formedness rule, that can be checked deterministically on the extracted output alone. Where both properties hold, the extract/gold-free-triage/selective-human-review workflow evaluated here is likely to reproduce a similar accuracy–cost–verifiability trade-off. Where the second property is largely absent (documents whose critical fields are free text or otherwise lack a checkable internal constraint), the gold-free guarantee weakens correspondingly, and broader, less targeted human review is the appropriate and honest fallback rather than an unsupported extension of this study's triage numbers. The general form of the workflow is given in Figure 4; the two properties stated here are the conditions under which that general form is expected to hold.

6. Conclusions

This study evaluated whether open-weight multimodal large language models can convert semi-structured legacy ERP documents into integration-ready data at a level of accuracy, cost, and verifiability sufficient to augment human data entry in a regulated setting. On a controlled, within-document comparison of 400 real controlled-substance stock-ledger documents against an independently double-entered human gold standard, the strongest model (Qwen-Dense) reached 93.95% overall cell accuracy with zero dropped records across the corpus; among the four evaluated models, field-level accuracy varied more strongly across vendors, whereas structural completeness differed consistently between the dense and Mixture-of-Experts models—an interaction rather than a uniform ranking (RQ1). A deterministic, gold-free invariant triage (requiring no human-labeled reference for the document being checked) flagged 0.61% of that model's records for review while the unflagged majority remained 94.1% accurate across all 11 fields; extending the triage with calendar rules flagged 4.61% and raised residual transaction-date accuracy from 92.1% to 95.9%. Both operating points are reported, and the extended level is the

appropriate default wherever date fidelity is regulatory-critical. The triage also caught the structural omissions produced by Mixture-of-Experts models, detecting every affected document in all four independent samples tested (RQ3). Under conservative, worst-case review-time assumptions, the resulting augmentation workflow was 13.5–29.4× faster on a wall-clock basis and 97.5–99.5% cheaper than human double-entry, and remained favorable even under the deliberately pessimistic assumption that every record required review (RQ2). This finding applies unevenly across fields—near-ceiling on structured numeric fields, materially weaker on the free-text Remarks field (20.8–43.7% depending on model and triage level)—and should not be read as uniform regulatory-grade reliability across the full 11-field schema.

The proposed workflow composes three explicit steps: an MLLM extracts every record from the rendered document images; deterministic accounting, calendar, and structural invariants screen every extracted record without requiring a human-prepared reference for that document; and trained reviewers examine only the flagged minority. Because this screening layer is what turns an unevenly reliable model into a workflow with a stated, bounded risk, it should be treated as a required component of deployment rather than an optional add-on. This conclusion applies to the ten structured fields the verification layer covers; the free-text Remarks field remains a residual risk requiring full, unconditional human review, and we do not claim regulatory-grade reliability for that field from this study. The approach is expected to transfer to other document types and settings that share the same two properties exploited here (repeated, structured fields and at least one internally checkable domain invariant) and to weaken where documents are dominated by free text lacking such a constraint (Section 5.4).

Future work includes testing transfer to other document types, languages, and institutions; extending gold-free verification to free-text fields; and characterizing cost and latency under alternative serving configurations beyond the OpenRouter measurement harness used here.

Author Contributions: Conceptualization, C.L. and B.T.; methodology, C.L.; software, B.T.; validation, C.L. and B.T.; formal analysis, B.T.; investigation, B.T.; resources, B.T.; data curation, C.L. and B.T.; writing—original draft preparation, C.L.; writing—review and editing, C.L. and B.T.; visualization, C.L.; supervision, C.L.; project administration, C.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Ethical review and approval were waived for this study because it evaluated routine occupational data-entry tasks performed by staff under the explicit authorization of the state pharmaceutical factory; no personal, medical, or sensitive data of the operators were collected, and all operator-related data were anonymized (recorded only as H1 and H2).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study. The two data-entry operators were informed of the purpose of the experiment and consented to the anonymized use of their task-timing and data-entry outputs for research and publication.

Data Availability Statement: The adjudicated gold standard, the first-pass files from both human operators, the main-run outputs of all four evaluated models, the system prompt, and the scoring script that reproduces the accuracy, alignment and verifiability tables are openly available in the GitHub repository, cited as reference [49].

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. U.S. Government Accountability Office. Information Technology: Agencies Need to Plan for Modernizing Critical Decades-Old Legacy Systems; Rep. GAO-25-107795; Washington, DC, USA, July 2025. Available online: https://www.gao.gov/products/gao-25-107795 (accessed on 16 May 2026).
2. Raza, M. Dark Data: An Introduction. *Splunk Blog*, 26 March 2025. Available online: https://www.splunk.com/en_us/blog/learn/dark-data.html (accessed on 16 May 2026).
3. Barchard, K.A.; Pace, L.A. Preventing human error: The impact of data entry methods on data accuracy and statistical results. *Comput. Hum. Behav.* **2011**, *27*, 1834–1839. [CrossRef]
4. DigiParser. Manual Data Entry Error Rate: How Many Typos Are Hiding in Your Systems? *DigiParser*, 2026. Available online: https://www.digiparser.com/statistics/manual-data-entry-error-rate (accessed on 16 May 2026).
5. El Morr, C.; Jammal, M.; Ali-Hassan, H.; El-Hallak, W. Data preprocessing. In *Machine Learning for Practical Decision Making: A Multidisciplinary Perspective with Applications from Healthcare, Engineering and Business Analytics*; International Series in Operations Research & Management Science; Springer: Cham, Switzerland, 2022; Volume 334, pp. 117–163. [CrossRef]
6. Bazerbachi, T. LLM OCR vs Traditional OCR: When AI Wins (and When It Doesn't)—2026 Benchmark. *Parsli Blog*, 27 March 2026. Available online: https://parsli.co/blog/llm-ocr-vs-traditional-ocr (accessed on 16 May 2026).
7. Dey, S.; Alaei, A.; Roy, P.P. Handwritten text recognition for low resource languages. *arXiv* **2025**, arXiv:2512.01348.
8. Sukhahuta, R.; Smith, D. Information extraction for Thai documents. In *Proceedings of the 5th International Workshop on Information Retrieval with Asian Languages (IRAL)*; Association for Computing Machinery: New York, NY, USA, 2000; pp. 103–110.
9. Nonesung, S.; Jaknamon, T.; Chaiophat, S.; Nitarach, N.; Wittayasakpan, C.; Sirichotedumrong, W.; Na-Thalang, A.; Pipatanakul, K. ThaiOCRBench: A task-diverse benchmark for vision-language understanding in Thai. *arXiv* **2025**, arXiv:2511.04479.
10. Ji, L.; Seyler, D.; Kaur, G.; Hegde, M.; Dasgupta, K.; Xiang, B. PHANTOM: A benchmark for hallucination detection in financial long-context QA. In *Proceedings of the 39th Conference Neural Information Processing Systems (NeurIPS)*; Datasets and Benchmarks Track; Neural Information Processing Systems Foundation, Inc.: San Diego, CA, USA, 2025.
11. Dell'Acqua, F.; McFowland, E., III; Mollick, E.; Lifshitz, H.; Kellogg, K.C.; Rajendran, S.; Krayer, L.; Candelon, F.; Lakhani, K.R. Navigating the jagged technological frontier: Field experimental evidence of the effects of artificial intelligence on knowledge worker productivity and quality. *Organ. Sci.* **2026**, *37*, 403–423. [CrossRef]
12. Raisch, S.; Krakowski, S. Artificial intelligence and management: The automation-augmentation paradox. *Acad. Manag. Rev.* **2021**, *46*, 192–210. [CrossRef]
13. Pitaknitinun, K.; Lawanprasert, Y.; Vimolkittipong, S. Thailand's legal framework concerning drugs. *J. Health Sci. Thail.* **2020**, *29*, S15–S30.
14. Kingdom of Thailand. Narcotics Act, B.E. 2522 (1979) Bangkok, Thailand. 22 April 1979. Available online: https://www.thailawforum.com/laws/Narcotics%20Act%202522.pdf (accessed on 16 May 2026).
15. Nonesung, S.; Nitarach, N.; Jaknamon, T.; Taveekitworachai, P.; Pipatanakul, K. Typhoon OCR: Open vision-language model for Thai document extraction. *arXiv* **2026**, arXiv:2601.14722.
16. Batini, C.; Cappiello, C.; Francalanci, C.; Maurino, A. Methodologies for data quality assessment and improvement. *ACM Comput. Surv.* **2009**, *41*, 1–52. [CrossRef]
17. Subramani, N.; Matton, A.; Greaves, M.; Lam, A. A survey of deep learning approaches for OCR and document understanding. In Proceedings of the NeurIPS 2020 Workshop on ML Retrospectives, Surveys & Meta-Analyses (ML-RSA), Vancouver, BC, Canada, 11 December 2020.
18. Hyland Software, Inc. What Is Intelligent Document Processing? A Beginner's Guide to IDP. *Hyland*. Available online: https://www.hyland.com/en/resources/articles/intelligent-document-processing-guide (accessed on 24 June 2026).
19. Grand View Research. Intelligent Document Processing Market (2026–2033) Size, Share & Trends Analysis Report; Rep. GVR-4-68040-023-3; San Francisco, CA, USA, June 2026. Available online: https://www.grandviewresearch.com/industry-analysis/intelligent-document-processing-market-report (accessed on 16 May 2026).
20. Research and Markets. Intelligent Document Processing Market Report 2026. Dublin, Ireland, February 2026. Available online: https://www.researchandmarkets.com/reports/5806873/intelligent-document-processing-market-report (accessed on 16 May 2026).
21. elDoc. Intelligent Document Processing with LLM: No Templates, No Configuration—Just Intelligence. *elDoc Blog*, 29 October 2025. Available online: https://eldoc.online/blog/intelligent-document-processing-with-llm/ (accessed on 16 May 2026).
22. Fu, L.; Kuang, Z.; Song, J.; Huang, M.; Yang, B.; Li, Y.; Zhu, L.; Luo, Q.; Wang, X.; Lu, H.; et al. OCRBench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning. In *Proceedings of the 39th Conference Neural Information Processing Systems (NeurIPS)*; Datasets and Benchmarks Track; Neural Information Processing Systems Foundation, Inc.: San Diego, CA, USA, 2025.

23. Zhang, W.; Cheng, Z.; He, Y.; Wang, M.; Shen, Y.; Tan, Z.; Hou, G.; He, M.; Ma, Y.; Lu, W.; et al. Multimodal self-instruct: Synthetic abstract image and visual reasoning instruction using language model. In *Proceedings of the 2024 Conference Empirical Methods in Natural Language Processing (EMNLP), Miami, FL, USA, 16 November 2024*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; Volume 2024, pp. 19228–19252. [CrossRef]
24. Xu, D.; Chen, W.; Peng, W.; Zhang, C.; Xu, T.; Zhao, X.; Wu, X.; Zheng, Y.; Wang, Y.; Chen, E. Large language models for generative information extraction: A survey. *Front. Comput. Sci.* **2024**, *18*, 186357. [CrossRef]
25. Tenckhoff, S.; Koddenbrock, M.; Rodner, E. LLMStructBench: Benchmarking large language model structured data extraction. *arXiv* **2026**, arXiv:2602.14743.
26. Wang, Z.; Shen, X. Hybrid OCR-LLM framework for enterprise-scale document information extraction under copy-heavy task. *arXiv* **2025**, arXiv:2510.10138.
27. Thongtaweechaikij, P.; Tangpong, P.; Inthiam, J.; Tangsuksant, W. Text extraction by optical character recognition-based on the template card. In *Proceedings of the 2024 1st International Conference Robotics, Engineering, Science, and Technology (RESTCON)*; IEEE: New York, NY, USA, 2024; pp. 188–192. [CrossRef]
28. Lertyosbordin, C.; Iamruttanawong, K. Content selection behavior of an edge small language model under word-budget compression: A cross-lingual study of English and Thai person descriptions. *Appl. Sci.* **2026**, *16*, 5754. [CrossRef]
29. Fedus, W.; Zoph, B.; Shazeer, N. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *J. Mach. Learn. Res.* **2022**, *23*, 1–39.
30. Jiang, A.Q.; Sablayrolles, A.; Roux, A.; Mensch, A.; Savary, B.; Bamford, C.; Chaplot, D.S.; de las Casas, D.; Hanna, E.B.; Bressand, F.; et al. Mixtral of experts. *arXiv* **2024**, arXiv:2401.04088.
31. Cai, W.; Jiang, J.; Wang, F.; Tang, J.; Kim, S.; Huang, J. A survey on mixture of experts in large language models. *IEEE Trans. Knowl. Data Eng.* **2025**, *37*, 3896–3915. [CrossRef]
32. Chitty-Venkata, K.T.; Howland, S.; Azar, G.; Soboleva, D.; Vassilieva, N.; Raskar, S.; Emani, M.; Vishwanath, V. MoE-Inference-Bench: Performance evaluation of mixture of expert large language and vision models. *arXiv* **2025**, arXiv:2508.17467.
33. Xu, H.; Hong, H.; Li, H.; Zhou, R.; Zhang, Y.; Huang, L.; Xue, H.; Shen, Y.; Lu, W.; Zhuang, Y. Seeing but not thinking: Routing distraction in multimodal mixture-of-experts. *arXiv* **2026**, arXiv:2604.08541.
34. Chu, X.; Ilyas, I.F.; Krishnan, S.; Wang, J. Data cleaning: Overview and emerging challenges. In *Proceedings of the 2016 International Conference on Management of Data (SIGMOD)*; Association for Computing Machinery: New York, NY, USA, 2016; pp. 2201–2206. [CrossRef]
35. Sriramanan, G.; Bharti, S.; Sadasivan, V.S.; Saha, S.; Kattakinda, P.; Feizi, S. LLM-Check: Investigating detection of hallucinations in large language models. In *Proceedings of the 38th Conference Neural Information Processing Systems (NeurIPS)*; Neural Information Processing Systems Foundation, Inc.: San Diego, CA, USA, 2024; pp. 34188–34216.
36. Korthals, L.; Akrong, E.; Geller, G.; Rosenbusch, H.; Grasman, R.; Visser, I. Towards reliable LLM grading through self-consistency and selective human review: Higher accuracy, less work. *Mach. Learn. Knowl. Extr.* **2026**, *8*, 74. [CrossRef]
37. Su, Y.; Xu, K.; Gao, Y.; Yang, F.; Li, C.; Yang, M.; Xu, T. Neuro-symbolic verification on instruction following of LLMs. *arXiv* **2026**, arXiv:2601.17789.
38. Wu, Y.; Zhou, S.; Liu, Y.; Lu, W.; Liu, X.; Zhang, Y.; Sun, C.; Wu, F.; Kuang, K. Precedent-enhanced legal judgment prediction with LLM and domain-model collaboration. In *Proceedings of the 2023 Conference Empirical Methods in Natural Language Processing (EMNLP)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 12060–12075.
39. Sachdeva, M.; Narayanan, C.; Wiedenkeller, M.; Sedlakova, J.; Bernard, J. A design space for the critical validation of LLM-generated tabular data. In *Proceedings of the EuroVis Workshop on Visual Analytics (EuroVA)*; The Eurographics Association: Eindhoven, The Netherlands, 2025. [CrossRef]
40. Hinov, N.; Ivanova, M. LLM-augmented algorithmic management: A governance-oriented architecture for explainable organizational decision systems. *AI* **2026**, *7*, 102. [CrossRef]
41. Aggarwal, V.; Kulkarni, J.; Mascarenhas, A.; Narang, A.; Raman, S.; Shah, A.; Thomas, S. Information extraction from fiscal documents using LLMs. In *Proceedings of the ACM International Conference AI in Finance (ICAIF)*; Association for Computing Machinery: New York, NY, USA, 2025.
42. Acemoglu, D.; Restrepo, P. Automation and new tasks: How technology displaces and reinstates labor. *J. Econ. Perspect.* **2019**, *33*, 3–30. [CrossRef]
43. Fügener, A.; Grahl, J.; Gupta, A.; Ketter, W. Cognitive challenges in human-artificial intelligence collaboration: Investigating the path toward productive delegation. *Inf. Syst. Res.* **2022**, *33*, 678–696. [CrossRef]
44. Wilson, H.J.; Daugherty, P.R. Collaborative intelligence: Humans and AI are joining forces. *Harv. Bus. Rev.* **2018**, *96*, 114–123.
45. Garza, M.Y.; Williams, T.B.; Ounpraseuth, S.; Hu, Z.; Lee, J.; Snowden, J.; Walden, A.C.; Simon, A.E.; Devlin, L.A.; Young, L.W.; et al. Comparing Medical Record Abstraction (MRA) error rates in an observational study to pooled rates identified in the data quality literature. *BMC Med. Res. Methodol.* **2024**, *24*, 304. [CrossRef] [PubMed]

46. Garza, M.Y.; Williams, T.; Ounpraseuth, S.; Hu, Z.; Lee, J.; Snowden, J.; Walden, A.C.; Simon, A.E.; Devlin, L.A.; Young, L.W.; et al. Error rates of data processing methods in clinical research: A systematic review and meta-analysis of manuscripts identified through PubMed. *Int. J. Med. Inform.* **2025**, *195*, 105749. [CrossRef] [PubMed]
47. Ministry of Labour, Kingdom of Thailand. Starting July 1: Ministry of Labour Revises Minimum Wage to Align with Economy and Improve Workers' Quality of Life. Available online: https://www.mol.go.th/en/news/starting-july-1-ministry-of-labour-revises-minimum-wage-to-align-with-economy-and-improve-workers-quality-of-life (accessed on 15 July 2026).
48. Flyvbjerg, B. Five misunderstandings about case study research. *Qual. Inq.* **2006**, *12*, 219–245. [CrossRef]
49. Lertyosbordin, C. "old_erp_pdf_json," GitHub Repository. 2026. Available online: https://github.com/chacharin/old_erp_pdf_json (accessed on 13 August 2026).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.