Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Bridging the QC–QA Cognitive Handoff: An Agentic Verification Harness for Error Detection, Specification Cross-Checking and Quality Documentation Assembly in Pharmaceutical Quality Assurance

**Anil Bhusal**
**Diwas Kunwar**
**Pushkar Niroula**
**Shashwot Rimal**
shashwot.rimal@gmail.com

**Article**

**Keywords:** pharmaceutical quality assurance, regulatory dossier, data integrity, tool-augmented language models, document verification, benchmark, GAMP 5, evaluation protocol

**Posted Date:** November 4th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-11344126/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** No competing interests reported.

# Bridging the QC–QA Cognitive Handoff: An Agentic Verification Harness for Error Detection, Specification Cross-Checking and Quality Documentation Assembly in Pharmaceutical Quality Assurance

Anil Bhusal[1], Diwas Kunwar[2], Pushkar Niroula[3], Shashwot Rimal[4]*

*Corresponding author: Shashwot Rimal, shashwot.rimal@gmail.com

## Abstract

**Background.** Quality control (QC) results reach regulatory dossiers through manual transcription and review. Such a benchmark does not yet exist. No collection of pharmaceutical quality documents carries discrepancies of known class and location, so automated verification cannot yet be evaluated against measured performance across realistic transfer errors and document types.

**Objectives.** To specify (i) an architecture in which large language models only read and propose values, deterministic code performs every calculation and acceptance test, and a qualified reviewer makes every decision; (ii) a protocol for building a reference corpus and evaluating such systems; and (iii) sensitivity estimates from three dossiers. The plan covers design, evaluation, and measurement.

**Methods.** Records from three ASEAN Common Technical Dossier Part II submissions (13, 14 and 10 records) were de-identified, and 112, 100 and 120 discrepancies across nine classes were introduced into dossiers A, B and C. A finding counted as a detection when it identified an injected discrepancy at the correct location with the correct class. The framework ran blind to the register. Sensitivity to injected discrepancies is reported per dossier with Wilson 95% confidence intervals and pooled by random-effects meta-analysis; clustering within records is not accounted for.

**Results.** The harness detected 95 of 112 (84.8%; 95% CI 77.0–90.3), 80 of 100 (80.0%; 71.1–86.7) and 105 of 120 (87.5%; 80.4–92.3) injected discrepancies, a pooled sensitivity of 84.1% (72.4–91.5), with detection in every class in every dossier. False-positive rate, specificity,

precision, run-to-run repeatability and performance relative to human reviewers or baseline systems were not measured, and the injection registers have not been independently validated.

**Conclusions.** The architecture detected a similar share of injected discrepancies in three dossiers, one run each. Although single runs across three dossiers cannot show reduced reviewer workload or superiority over existing practice, the observed pattern supports further testing rather than any claim of practical benefit. The protocol defines the required next experiments.

**Keywords:** pharmaceutical quality assurance; regulatory dossier; data integrity; tool-augmented language models; document verification; benchmark; GAMP 5; evaluation protocol

## 1. Introduction

Every result in the quality section of a marketing application passes through a manual step. Quality control (QC) generates analytical, stability and process-validation data inside qualified systems, and quality assurance (QA) must then carry those results into the finished-product specification, the validation summary, the batch analyses, the certificate of analysis and the stability tabulation of the dossier(1,2). Manual comparison governs each of these transfers. At each transfer a reviewer compares a value with an acceptance criterion held in another document, recalculates it where the criterion requires, and confirms that batch identifiers, dates and method revisions agree wherever they appear.

Each individual check is simple. Volume alone creates the real difficulty here. Quality control and quality assurance form the core of a pharmaceutical quality system(3), and the manual handoff between them adds a human-factor problem: human error is widely reported as a major contributor to quality deviations and product recalls, which carry economic and reputational cost(4,5). A single dossier contains hundreds of values that must each agree with several others, and data-integrity guidance and enforcement records show that transcription errors, missing results, inconsistent identifiers and unreconciled records remain common inspection findings(6,7). A defect that survives review can return as a deficiency letter or an approval delay, while a result recorded as conforming against the wrong limit becomes a product-quality risk(8). The quality module is a recurring source of regulatory objections. Quality issues alone were behind about 5% of first-cycle non-approvals of new molecular entities; in one FDA review of first-cycle generic-application complete response letters, more than 70% of major deficiencies were quality related(9). Analyses of US and EU submissions likewise document

recurring chemistry, manufacturing and controls (CMC) deficiencies that delay approval and patient access(10,11).

Artificial intelligence has been applied extensively to data generation, through process analytical technology, continued process verification and process monitoring(12,13), and to regulatory text(14), but rarely to checking reported data against its own specification and its parallel records(15,16). Its output can be fluent and wrong(17), it can vary between runs on identical input, and it provides no inherent link between a statement and its source. Unconstrained LLM use offers no direct fix. These properties conflict with the attributability and traceability expected under ALCOA+ and with the computerised-system expectations of 21 CFR Part 11(18).

A known remedy in the machine-learning literature is to let the model delegate computation to external code and tools(19–21). While that pattern is established, this paper claims no novelty for it. What is missing for pharmaceutical quality is twofold: a way to apply it so that the division of roles maps onto GxP requirements, and, more fundamentally, a reference standard against which any such system can be measured. No published corpus of pharmaceutical quality documents carries expert-validated discrepancies at known locations, so claims of automated verification performance in this domain cannot currently be tested.

This paper makes three contributions. First, it describes a verification harness in which LLMs only read documents and propose values, deterministic code performs every calculation and acceptance test, and a qualified reviewer makes every decision. Second, it specifies a protocol for building a reference corpus and for evaluating such systems against it, including negative controls, comparators, ablations and repeated runs. Third, it reports sensitivity to injected discrepancies in three dossiers, each processed once. Although the third result shows the architecture can find planted discrepancies in more than one dossier, it does not demonstrate practical utility. Because the test material consists of submitted dossiers rather than raw laboratory records, the work addresses pre-submission dossier verification, the last point at which QC-to-QA transfer errors can be caught before submission.

## 2. Background and related work

### 2.1 Where transfer discrepancies arise

The conventional workflow moves quality information through eight stages, from collection of QC records to regulatory submission (Figure 1). At one stage values are transcribed into QA documentation; at three further stages compilation and assembly carry them forward, while two stages rely on manual comparison. Once transcribed, no stage re-derives a value from its source. An early error therefore persists unless a reviewer goes back to the original record. Confirming agreement across every document that mentions each printed value taxes working memory far more than calculation, and that comparison defines the practical limit at each checking stage.

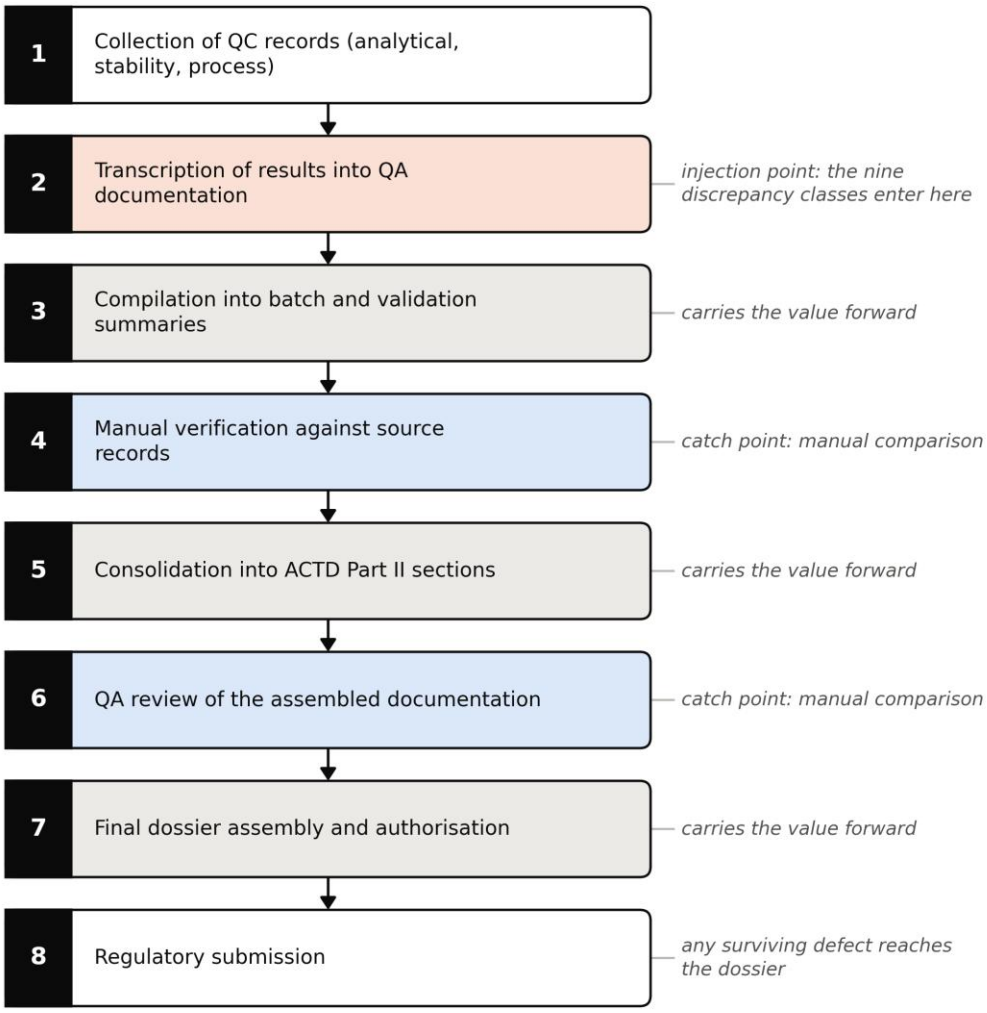



**Figure 1.** The conventional QC-to-QA documentation workflow as eight sequential stages. Discrepancies enter at transcription (stage 2), are carried forward without re-derivation (stages 3, 5 and 7) and can be caught only by manual comparison (stages 4 and 6) before any surviving defect reaches the submission (stage 8).

## 2.2 Regulatory expectations for AI in GxP settings

Any system used in this workflow must produce attributable, accurate, complete and consistent records(6), and software used in quality decisions must be validated and audit-trailed(18). Validation becomes harder with non-deterministic outputs, yet remains possible. GAMP 5 (second edition) treats AI and machine learning in a dedicated appendix, the ISPE GAMP Guide: Artificial Intelligence (2025) describes a life-cycle approach for validating AI-enabled systems(22), while both the FDA draft guidance on AI for regulatory decision-making and the EMA reflection paper outline risk-based credibility assessment(23). The design question is not whether AI can be validated at all, but which roles a generative model can occupy without taking a quality decision.

## 2.3 AI in pharmaceutical quality

Published work concentrates on the two ends of the quality lifecycle (Table 1). Upstream, reviews and case studies examine process analytical technology, continued process verification, analytical monitoring and manufacturing applications under GxP(12,13,16). Regulatory translation(14), regulatory views on AI in GMP environments(24) and broad surveys of AI across the pharmaceutical lifecycle(15) make up the downstream literature. In the indexed set examined, no report in either group describes a system that checks reported quality data against its specification and parallel records, nor performance measured against a ground truth holding both altered and unaltered records.

## 2.4 Tool-augmented language models and automated verification

Program-aided language models and program-of-thoughts prompting pass arithmetic to an interpreter(20); Toolformer and ReAct allow a model to call external tools during reasoning(21,25); and neurosymbolic approaches join learned perception with symbolic reasoning more broadly(26). The architecture described here applies this pattern to one domain. It differs from general-purpose tool use in three ways, because the model cannot select its next step, only four fixed tools are exposed, and each proposed value must be located on the cited page before any tool accepts it.

Verification of claims against evidence has well-established benchmarks. FEVER and SciFact check sentences against text(27,28); TabFact and FEVEROUS check statements against tables(29,30); FinQA and TAT-QA call for numerical reasoning over tables and text(31,32); and EHRCon tests consistency between notes and structured tables inside one health

record(33). The reading layer behind such systems comes from layout-aware document models that jointly encode text, spatial layout, and visual features(34,35), together with table-structure recognition that recovers rows, columns, and cell relationships from document images(36). Built on this layer, downstream systems perform numerical and hybrid tabular-textual reasoning over financial reports(31,32), as well as consistency checking between clinical notes and structured tables and question answering over document images(33,37)., from retrieval-augmented generation grounding generated text in retrieved passages(38,39). Each of these benchmarks checks one claim against one source. Reconciling the same quantity across several documents under a fixed order of source precedence, together with acceptance rules that carry regulatory force, is what dossier verification instead demands.

**Table 1.** Prior work relevant to automated verification of pharmaceutical quality documentation, and what each leaves open.

| Work | What it addresses | What remains open for dossier verification |
|---|---|---|
| Huanbutta et al. 2024(15) | Survey of AI from discovery to post-market surveillance, including QC | No verification of reported data against specifications or parallel records |
| Nakka et al. 2026(16) | AI applications and GxP implementation in manufacturing | Manufacturing focus; no document-level verification or ground-truth evaluation |
| Gaikwad and Pawar 2026(13) | AI-assisted analysis and quality monitoring | Analytical and process monitoring rather than dossier consistency |
| Nagy et al. 2022(12) | Neural networks in process analytical technology | Acts before data become documents |
| Chen et al. 2025(14) | LLM for regulatory-affairs translation | Measures translation fidelity, not detection of data discrepancies |
| Niazi 2025(24) | Regulatory perspectives on AI/ML in GMP | Conceptual; no evaluated system |
| PAL, PoT, Toolformer, ReAct(19–21,25) | Delegating computation and actions from LLMs to tools | General-purpose; no fixed control flow or source-location requirement |
| TabFact, FEVEROUS, FinQA, TAT-QA(29–32) | Verification and numerical reasoning over tables | One claim against one source; no cross-document reconciliation or regulatory limits |
| FEVER(27) | Verification of textual and scientific claims | Sentence-level entailment, not numerical agreement |
| EHRCon(33) | Consistency between notes and tables in one record | Single record; no order of source precedence |

## 2.5 Literature search and the gap

The literature was searched in PubMed, Scopus, Web of Science, IEEE Xplore, the ACM Digital Library and arXiv (cs.CL, cs.CV, cs.IR) from January 2015 to March 2026, combining terms for pharmaceutical quality documentation, AI methods and verification tasks as a structured search without systematic-review methods or record counts. No PRISMA 2020 flow diagram(40) is reported. In the indexed literature searched, no dataset of pharmaceutical quality documents annotated with discrepancies of known class and location was found.

Two limitations run across the work reviewed, and they shape both the design and the testing that follow. Methodologically, no published approach extracts information from several quality documents, normalises it into a common representation, compares it with specifications and related records, and links each finding to its source before human review. More consequentially, the evaluative gap means claims that AI improves documentation efficiency rest on no performance measured against a predefined ground truth with both discrepant and unmodified records, so they cannot be falsified, which this paper tackles by setting out how such a ground truth should be built and used. The first is handled with one architecture for evaluation against it.

## 3. Aim, objectives and scope

The aim is to deliver an architecture and an evaluation protocol for automated verification of pharmaceutical quality documentation before dossier submission, with detection estimates from three dossiers. The objectives for this work are threefold:

- **O1.** To specify a verification architecture that separates probabilistic reading, deterministic checking and human decision, and binds every finding to its source.
- **O2.** To specify a protocol for constructing a reference corpus and evaluating verification systems against it, covering negative controls, comparators, ablations, repeatability and independent validation of the ground truth.
- **O3.** To report the sensitivity of the architecture to injected discrepancies in three dossiers, one run each, with its limitations stated.

Specificity, precision, false-alarm burden, extraction accuracy, repeatability, and comparison with human reviewers and baseline systems are outcomes of the protocol (Section 4.8). They were not obtained here and are not claimed. Running three different dossiers once each tests whether detection holds across dossiers, but it does not measure run-to-run repeatability, which can only be assessed through repeated runs of the same cases under identical conditions.

**Context of use.** The system only reads and checks values. It reads documents, proposes values, recomputes derivable statistics, compares results with bound acceptance criteria, reconciles repeated values across documents, and assembles checked values into the required dossier structure with a link to each source. By design it writes no narrative text, reaches no batch-release or disposition decision, closes no finding, applies no correction without reviewer

acceptance, and does not judge whether an acceptance criterion is scientifically appropriate, which remains a matter for pharmaceutical development(41).

# 4. Methods

## 4.1 Corpus

The corpus comprises three authentic ASEAN Common Technical Dossier (ACTD) Part II (Quality) submissions(1), termed dossiers A, B and C. Dossier A holds 13 records from a small-molecule, immediate-release, film-coated tablet submission. Dossier B holds 14 records from a film-coated tablet submission, and dossier C 10 records from a capsule submission. All three come from one manufacturer. Every record is a born-digital PDF, one file per record, with included sections listed in Supplementary Table S2. To protect confidentiality, commercial names, personnel, sites, active substances and reference numbers were anonymized throughout with a fixed substitution key. Technical data, analytical findings and surrounding text were left untouched, and the identical procedure covered all three dossiers.

Because the corpus consists of submitted dossiers, the top two levels of the source-precedence order (Section 4.4.3), raw instrument output and run logs, appear only where a dossier carries raw-data annexes. All three dossiers do include such annexes. Their coverage is incomplete (Section 4.3). The study therefore tests pre-submission dossier verification, not verification at the laboratory handoff. Each dossier also retains its own pre-existing defects from the original submission, still uncensused (Section 4.8).

## 4.2 Discrepancy taxonomy and injection

Nine discrepancy classes were defined from data-integrity guidance, published analyses of FDA warning letters, and the ICH quality guidelines governing each affected field (Table 2). The study team introduced discrepancies into copies of the records, in the ACTD section where each class occurs in practice and always where contradicting evidence sits in another document or section. The unmodified originals were kept. Every discrepancy went into an injection register before the harness ran, with class, host document, page, field, original and altered values, and the contradicting record. The harness never saw the register.

The study introduced 112 discrepancies into dossier A, 100 into dossier B and 120 into dossier C, spread across the nine classes in roughly the same proportions per dossier (Table 3). Members of the team that built the harness injected all three dossiers, and one harness version

processed each dossier as a single package in a separate run. That gives about 8.6 discrepancies per record in dossier A (112 in 13 records), 7.1 in dossier B (100 in 14) and 12.0 in dossier C (120 in 10). Such density far exceeds any real submission, and independent experts have not rated the realism of individual cases. The host record for each discrepancy is logged in the injection register (Supplementary Table S1).

**Table 2.** The nine discrepancy classes, where each was introduced, the record that contradicts it, and the sources on which the class is based.

| Class | Discrepancy | Sections injected | Contradicting record | Basis |
|---|---|---|---|---|
| D1 | Reported value differs from the raw record (for example, transposed digits) | P.5.3, P.5.4, P.2.3.3 | Raw-data annex; derived statistics | (7) |
| D2 | Result outside its limit presented as conforming | P.5.4, P.8.3, P.2.6, P.5.5 | Limits in P.5.1 and the monograph | (7,8) |
| D3 | Result reported in a unit other than the one specified | P.5.4, P.8.3, P.4.2 | Unit declared in P.5.1 | (42) |
| D4 | Batch identifier disagrees with the identifier used elsewhere | P.8.3, P.5.4 | P.5.4, P.3.5 and P.6 | (43) |
| D5 | Required result missing from a table | P.5.4, P.2.3.3, P.8.3 | Test panel in P.5.1; stability pull schedule | (6,44) |
| D6 | Dates conflict or place events in an impossible order | P.8.3, P.5.4 | Manufacturing record in P.3.5; adjacent pull dates | (6,45) |
| D7 | The same result or method revision stated differently in two documents | P.5.4 vs P.8.3; P.5.3 vs P.5.2 | Every section citing the same fact | (2,46) |
| D8 | Acceptance limit altered, or set below what the validated method can support | P.5.1, P.4.1 | Monograph; method capability in P.5.3 | (42,47) |
| D9 | Document cited by another section is absent | Manifest, with P.5.3, P.6 or P.9.2 withheld | Sections citing the missing document | (1,2) |

## 4.3 Reachability

An injected discrepancy was classed as reachable if the package holds a record that contradicts it or from which the correct value can be derived, and as blind otherwise. The original protocol tied this grouping to checks the harness performs, which made the secondary denominator partly depend on the system under test. Restated here in document-only terms, the five discrepancies first called blind stay blind: each is a D1 alteration inside its limit, never restated elsewhere, entering no printed statistic, with the contradicting raw record absent from the package. Whether the other 107 discrepancies in dossier A count as reachable under the document-based definition has not been independently re-adjudicated. The reachable denominator is thus secondary and provisional, for dossier A only. Reachability in dossiers B and C remains unclassified.

## 4.4 Architecture of the verification harness

### *4.4.1 Overview: stages, roles and control flow*

Five fixed stages handle each dossier: Extract, Source of truth, Identify, Fix and Compile, followed by review of every finding by a qualified QA pharmacist (Figure 2). A controlling program, the orchestrator, runs the stages in order and moves defined inputs and outputs between them. One responsibility per component. Three roles stay apart throughout:

- **Language models propose.** Confined to reporting what a page contains, they never declare it correct. A dense page may take them several passes, while correctness remains someone else's job.
- **Deterministic code checks.** Whereas models only describe, code verifies each proposed value against its source, recomputes every derivable statistic, tests each acceptance criterion and then assembles the output.
- **A qualified QA pharmacist decides.** Only the pharmacist can close a finding, accept a correction, or sign the output, which keeps every quality decision with a qualified person.

Components using language models are called sub-agents. A sub-agent decides how to read a page — how many passes a dense chromatographic annex needs, for instance. Run order, inputs and output handling are fixed in code and frozen for the run. Every sub-agent output is checked independently before use.

Built as a nine-state machine — idle, ingest, ground, consolidate, dispatch, assert, commit, quarantine and halt — the orchestrator starts every run in idle and finishes in commit or halt. Each state-event pair maps to exactly one next state. No transition reads a model-proposed value. All run paths can therefore be enumerated and tested beforehand.

Removed entirely from the process are shell access, code evaluation (eval and exec), file-system writes and web access. Sub-agents come from a fixed registry and cannot launch other agents. Four tools exist for a run (Supplementary Table S4):

- **bind_anchor** links a value to the page region from which it was read.
- **recompute_statistic** passes raw values to a calculation module (Section 4.4.4).
- **cross_match_entity** tests a consistency rule across documents.
- **assert_specification** tests a result against its acceptance limit.

Any request for another action counts as a system fault and halts the run. Each stage is detailed in Sections 4.4.2 to 4.4.6.

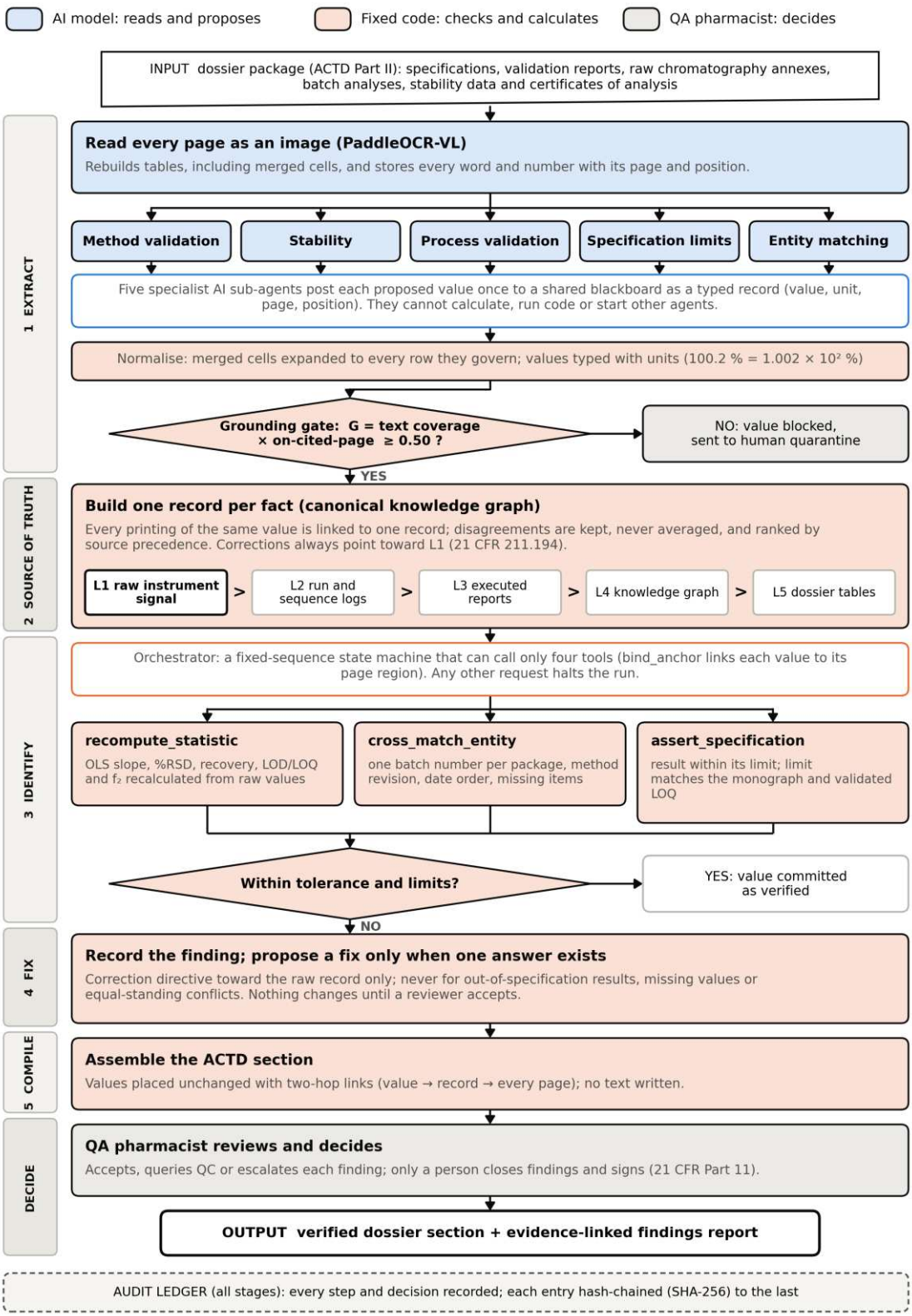



**Figure 2.** How the verification system works. Blue boxes are performed by AI models, which only read pages and propose values; orange boxes are performed by fixed code, which checks every value and performs all calculations; grey boxes are performed by a qualified QA pharmacist, who makes every decision. The left rail

shows the five stages and the human decision. G is the grounding score (Equation 1); L1 to L5 are the levels of source precedence (Section 4.4.3). Every step is written to an append-only, hash-chained audit ledger.

### *4.4.2 Stage 1 (Extract): reading, extraction and the grounding gate*

**Reading.** PaddleOCR-VL, a layout-preserving document model, reads each page as an image at native resolution (up to 1536 × 1536 pixels). Tables come back reconstructed directly, spanned rows and columns included, with every word and number stored alongside page and normalised coordinates. Layout preservation was chosen because line-by-line text recognition fails validation documents in three ways:

- **Merged labels.** Spanning several rows, a category label sticks to the first row only, so limits below lose their governing attribute unless spanning cells are resolved.
- **Unruled tables.** Without ruling lines, chromatography tables collapse into interleaved text and column structure is lost.
- **Lost positions.** When the stored position is discarded, no finding can be traced back to its source and provenance is lost.

**Extraction.** Five specialist sub-agents extract values, one document domain apiece:

- **Method validation.** Calibration, recovery, replicate-precision and forced-degradation tables under ICH Q2(R2)(47).
- **Stability.** Multi-station matrices, arranged by ICH Q1A(R2) climatic zone and packaging configuration(44,48).
- **Process validation.** Critical process parameters, stage yields and blend uniformity(49).
- **Specification binding.** Each monograph and internal-specification limit, tied to the attribute it governs.
- **Entity matching.** Links values referring to the same item; extracts no new values itself (Section 4.4.3).

The sub-agents draw on a pool of two language models with fixed versions, Gemini 3.8 Flash and Mistral Large 3 (Section 4.5). Difficulty decides the routing: the lower-cost model takes clean ruled tables, while the stronger model takes harder passages, such as footnotes tying one merged label to several limits. The models never check each other and never calculate. Proposals go to a shared blackboard — a workspace where records accrue but cannot be overwritten — as typed records (value, unit, page and position). One reading per page feeds all sub-agents.

**Normalisation.** Two deterministic operations then standardise proposals. Spanned cells resolve first: a value covering several rows is copied to every record it governs with its layout data, so later comparisons never see an empty cell. Unit normalisation follows, converting variants like 100.2 %, 100.2 % w/w and $1.002 \times 10^2$ % to one decimal value plus unit, which lets D3 unit mismatches surface.

**Grounding gate.** Cited page first, trust later: before proposed value v is admitted, the harness confirms its presence on the cited page. It computes grounding score G against that page's token stream S:

$$G(v, S) = [\ |T(v) \cap T_p(v)(S)| / |T(v)|\ ] \times I[\ B(v) \subseteq B_page(v)\ ]$$

where T(v) is the multiset of subword tokens in the value, T_p(v)(S) is the token stream of the cited page, B(v) is the value's bounding box, B_page(v) is the extent of the cited page, and I[·] equals 1 if the box lies within the page and 0 otherwise. The first term gives the share of the value's text found on the cited page; the second tests position on that page. Multiplied, they punish both failure modes: outside the cited page scores zero despite matching text, and a token-less value scores zero. Accepted: values with $G \geq 0.50$. The rest are excluded, logged and queued in quarantine for human review.

The gate has two limitations. Its 0.50 threshold lives in code yet awaits characterisation across values (Section 4.8). A correct value lifted from the wrong cell of the right table still slips through, and how often this happens has not been measured.

### *4.4.3 Stage 2 (Source of truth): one record per fact and source precedence*

Several places report the same fact. One assay result can surface in the certificate of analysis, batch analysis, first stability time point and process validation summary. Stage 2 merges these repeats into a canonical knowledge graph, storing each fact once with links to every printed location.

**Consolidation.** Values sharing batch, test, time point and attribute are grouped by the entity-matching sub-agent, which also maps aliases for one product — trade name, label description, internal formulation code — onto a single identifier. A fixed rule then governs each group:

- **Printings agree.** One record survives, listing every place the value is printed.
- **Printings disagree.** Stored on the record as a conflict, the disagreement travels to Stage 3. Conflicting values are never averaged or settled by majority vote.

**Structure of the graph.** Quality entities form the nodes: product, batch, analytical method, calibration point, stability time point, monograph limit and source span. Typed relations form the edges, and each node links to all source page regions it was read from. Three domain subgraphs — method validation, stability, process validation — share a core of product, batch, method and monograph-limit nodes. A batch identifier drifting between domains is thus caught at that shared core.

Embedding-based retrieval(38) was left out, for three reasons. Headings part from cell values. Near-identical strings read as similar — 99.8% versus 89.8%, or B-24-0117 versus B-24-01l7 — precisely when the harness must notice the gap. Cross-document references go unfollowed.

**Source precedence.** A fixed order, highest (L1) to lowest (L5), settles conflicts:

- **L1.** Raw instrument signals (chromatographic integration, balance printouts).
- **L2.** Sequence and run logs.
- **L3.** Executed validation and study reports.
- **L4.** The canonical knowledge graph.
- **L5.** The compiled dossier tables.

When sources disagree, the higher-level value stands and any correction targets the lower level. That reading follows 21 CFR 211.194, which bases quality release on laboratory raw records (43). Rewriting a raw peak area to fit a summary table would be manipulation. Fixed by design, the order is directly inspectable. Here, L1 and L2 exist only where dossiers include raw-data annexes (Section 4.1).

#### *4.4.4 Stage 3 (Identify): deterministic checks*

Record type routes each record to its check; the orchestrator does the routing with no model input. Checks come in two kinds: recalculation of derived statistics, and consistency rules across documents. Passing everything applicable earns “verified”. Any failure sends the value to Stage 4.

**Recalculation of derived statistics.** From graph raw values, five compiled calculation modules (C/NumPy, exact decimal arithmetic) recompute each derived statistic and test the reported figure against a preset tolerance(47,50):

- **Linearity.** Tolerance first: relative slope difference ≤ 0.10%. Ordinary least-squares slope m, intercept c and coefficient of determination $r^2$ come from the n calibration pairs ($x_i$, $y_i$), with intercept and $r^2$ recalculated alongside.
- **Precision.** Whereas linearity uses a relative bound, here the absolute difference must be ≤ 0.01 percentage points. Mean, sample standard deviation s and %RSD = 100 s / $\bar{x}$ over n ≥ 6 replicate injections.
- **Accuracy.** Recovery = 100 × recovered / spiked per level, with mean recovery and per-level %RSD. Absolute difference ≤ 0.05 percentage points.
- **Detection and quantitation limits.** LOD = 3.3 σ / S and LOQ = 10 σ / S, from residual standard deviation σ and calibration slope S(47). Since no method reports below what it can quantify, the declared reporting threshold must sit at or above the recalculated LOQ.
- **Dissolution similarity.** Absolute difference ≤ 0.1. That budget covers $f_2 = 50 \log_{10} \{ [1 + (1/n) \Sigma (R_t - T_t)^2]^{-0.5} \times 100 \}$, taken over n paired time points [71, 72, 74].

Supplementary Table S5 repeats the formulas and tolerances.

**Consistency rules.** Four rules sweep the graph:

- **Batch identity.** One package, one batch identifier. Several identifiers differing by a single character (Levenshtein edit distance of 1, e.g. letter l for digit 1) flag as likely typing errors, with every conflicting spot cited.
- **Method revision.** Summary-cited revision must equal the revision in raw sequence-log headers from the testing run.
- **Chronology.** Manufacture, sampling, testing then stability pull — in that order — with each pull inside its protocol window(44).
- **Specification limits.** Each result must sit within its test's acceptance limits. Limits themselves must mirror the governing pharmacopoeial monograph, exposing altered limits (D8). Impurity reporting thresholds follow the ICH Q3B(R2) order: reporting threshold no higher than identification and qualification thresholds(51).

Dissolution (USP <711>) and uniformity of dosage units (USP <905>) proceed in stages(52,53): a few units first, more only if stage one misses its criteria. The harness therefore tests each stage alone. Short of unit counts or stated statistics, the check stalls. Such cases read "insufficient evidence", with the gap raised separately.

QA verification gets every out-of-limit result. Confirmed out-of-specification (OOS) it is not, since only laboratory investigation confers that verdict(8,43).

### *4.4.5 Stages 4 and 5 (Fix and Compile): corrections and assembly*

**Stage 4: findings and proposed corrections.** Every check failure becomes a finding, with a correction offered solely where one correct value stands out. The path runs from the discrepant entry to the highest-ranked raw record on hand (Section 4.4.3), from which the value is recomputed; when recomputed and reported values diverge beyond tolerance, a directive goes out stating present value, proposed value, source level and evidence location. While source records stay untouched, a directive takes effect following reviewer acceptance alone. Three situations draw no proposal:

- **Out-of-specification results.** These require investigation under 21 CFR 211.192(54).
- **Missing values.** The documents cannot show whether the test was performed.
- **Conflicts between sources at the same level.** The precedence order cannot decide between them.

Each case closes under a single ledger entry — discrepancy detected, no discrepancy detected, requires QA verification, or insufficient evidence — with scoring of those entries governed by Section 4.6.

**Stage 5: assembly.** Target ACTD sections are built from canonical-graph records without reopening source documents, the required-element list selecting which records apply and each value passing over unchanged. Absence of a matching record triggers a missing result (D5) or missing document (D9) report. Accepted corrections enter with their audit note; pending cases carry the original value plus its open finding. Narrative prose, summaries and regulatory judgements lie outside this stage.

**Reviewer decision.** Accept, query with QC, or escalate — that choice, for every finding, belongs to the reviewer, whose closure and sign-off complete the record in the audit ledger (Section 4.4.6).

### *4.4.6 Context, traceability and learning*

**Context scoping.** Sub-agents see just the slice needed for their section: its immediate neighbours plus the shared product–batch–method–monograph core. Facts elsewhere must be

asked for by name, arrive for one check, then leave once recorded. Bulk intake from unrelated sections is refused, keeping that material away from the model.

**Traceability.** From assembled value back to graph record, then from graph record to each printed source region: that two-hop chain holds for every value, under one of four labels:

- **Line-verified.** A value tied to an exact line and position on its source page.
- **Derived.** Indicates a figure recomputed from raw inputs at known locations.
- **Ambiguous.** Applies where several identical pages could have supplied the match.
- **Blocked.** Means the grounding gate rejected the value, hence its exclusion.

**Audit ledger.** Extraction, consolidation, tool invocation, finding, correction and reviewer action all land in append-only storage, each entry carrying time, event identifier, agent identifier, tool signature, payload hash, verdict and prior-entry SHA-256. Editing history therefore breaks the chain visibly. Sealing happens at run end, after which scoring uses the sealed copy (Section 4.6).

**Learning.** Layout navigation is the sole persistent gain: reviewer layout-reading fixes return as YAML or JSON navigation notes reused on later packages from the same origin. Calculation modules, acceptance criteria, grounding threshold, sub-agent order and model weights remain unchanged in service.

### 4.5 Deployment and data handling

The page reader (PaddleOCR-VL), the grounding gate, consolidation, the calculation modules, the orchestrator and the audit ledger ran on a single local workstation with one NVIDIA A10G GPU (24 GB) and an AMD EPYC 7763 processor. The two language models, Gemini 3.8 Flash and Mistral Large 3, were fixed to specific versions. They were chosen as the lowest-cost models that could handle the most difficult extraction task.

Neither model fitted the local workstation. Gemini exists solely through hosted service, while Mistral Large 3 — weights openly released — exceeds a single 24 GB GPU. Hosted endpoints therefore served both, leaving the installation network-connected. Sub-agents themselves held no network, file-system or shell rights; the orchestrator alone forwarded page content. Identifiers, API versions, endpoint regions, data-retention and training-use terms, plus access dates, appear in Supplementary Table S6.

Consequently, material from three authentic submissions passed through third-party services under the Supplementary Table S6 terms and the written permission of each holder (Declarations).

### 4.6 Scoring

Every dossier ran once. Comparison of sealed-ledger findings against the injection register followed each run, performed outside the harness and hidden from it, placing each finding in one of three groups:

- **True positive.** Correct document, page and field together with correct class defined a hit, regardless of closing outcome — a correctly placed and classed item counted even when closed as "requires QA verification" or "insufficient evidence". Specification classes (D2, D8) need this proviso, since out-of-limit results there are raised for QA verification rather than confirmed (Section 4.4.4). Counts by closing outcome were not tabulated.
- **Partial match.** For findings right in one axis but not the other — location correct, class wrong, or the reverse — recording without credit as true positives was the handling.
- **Outside the register.** Findings matching nothing in the register sit unclassified as false positives until the census of pre-existing source defects finishes (Section 4.8).

### 4.7 Statistical analysis

Sensitivity to injected discrepancies — true positives over all injected items — forms the primary measure. Requiring only true positives and false negatives, it avoids the true-negative and false-positive counts that specificity and precision would demand and which were not collected. Per-dossier values carry Wilson score 95% confidence intervals(55). Pooling across dossiers used random-effects meta-analysis of logit-transformed proportions, DerSimonian–Laird between-dossier variance(56), Hartung–Knapp interval widening, advisable with few units(57). Heterogeneity appears as Cochran's Q and $I^2$(58), though three dossiers give the test scant power. An unadjusted pooled proportion with Wilson interval is shown for reference. Nesting of discrepancies within records goes unadjusted; cluster bootstrap by record(59) awaits discrepancy-to-record allocation (Supplementary Table S1). Class- and function-level figures pool counts across dossiers with Wilson intervals and remain exploratory. Further inferential testing was absent.

## 4.8 Prospective evaluation protocol

What the three runs omit, the protocol below supplies (Figure 3), with present-study status collected in Section 6.4.

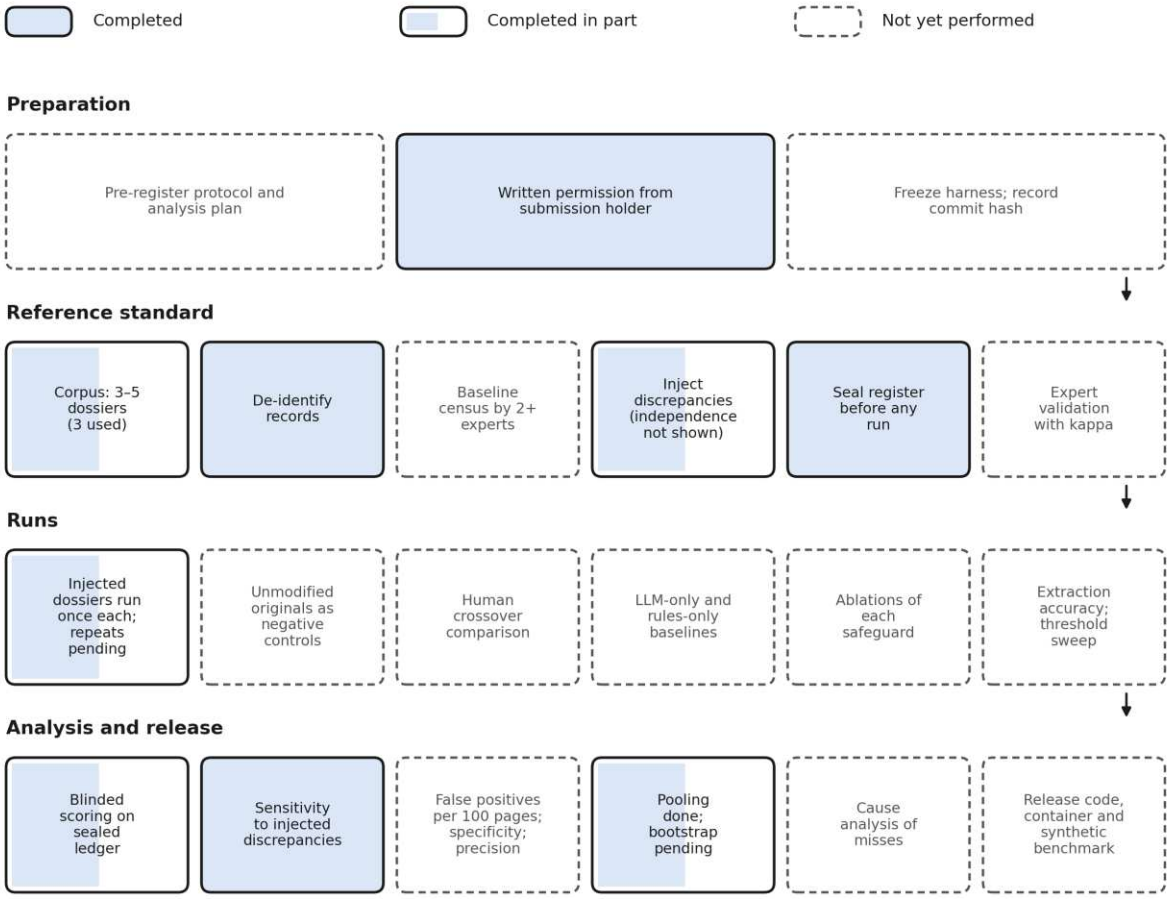



**Figure 3.** Evaluation protocol. Filled boxes show steps completed in the present study, half-filled boxes steps completed in part, and outlined boxes steps not yet performed. The protocol is to be pre-registered before the next evaluation.

**Registration and freezing.** Pre-registration of protocol, outcomes and analysis plan will use the Open Science Framework. Ahead of any injection, the harness freezes as a container image whose commit hash, with timestamp, opens the hash-chained ledger. Subsequent edits to code, prompts, thresholds or model versions open a new version for separate evaluation.

**Corpus.** Three to five dossiers or pre-submission QC packages at minimum, drawn from different manufacturers, with scanned records included and at least one non-oral-solid form, will make up the next corpus. Against that bar — one manufacturer, all born-digital, all oral solids (two film-coated tablets and a capsule) — the current three dossiers leave manufacturer, scanned-record and dosage-form coverage unmet. Each holder's written permission will be secured, and hosted-model processing will occur strictly within that permission.

**Baseline census.** Prior to injection, two or more independent QA or QC experts will log every pre-existing source-record anomaly by document, page, field and class, sorting genuine defects from conversion artefacts. Genuine defects enter the ground truth.

**Independent injection and validation.** Personnel outside the harness team will inject discrepancies at densities aligned with genuine deficiency rates. Two or more independent experts will then verify each register line — altered value wrong, cited record contradictory, untouched records clean — and score realism on a five-point scale. Raw agreement alongside Cohen's or Fleiss' kappa with 95% confidence intervals(60,61) will be published, a third expert settling disputes. Document-only reachability adjudication falls to the same experts.

**Negative controls.** Injected copies and their untouched originals both undergo processing. Findings on an injected copy, absent from its original and unmatched in the register, count as injection-process false positives. Findings on originals beyond the census count as harness false positives. Reported for every finding: true positive, outside-register, "requires QA verification" or "insufficient evidence" status, plus total findings raised.

**Repeatability.** Three to five repetitions per case, under locked decoding settings (temperature, top-p, seed) and locked model identifiers, will test stability. Reported: concordance of finding identity, cited evidence and class across runs, with detection-rate spread.

**Comparators.** Three head-to-heads are planned. Blinded QA pharmacists, at least three, will examine cases under time limits in a crossover with and without harness output. The same models, prompted end to end on the same documents, form an LLM-only baseline. Deterministic rules over conventional optical-character-recognition text, minus any LLM, form a rules-only baseline.

**Ablations.** Stripped variants will run: grounding check removed, consolidation removed, model arithmetic replacing calculation modules, and full-document context replacing section scoping.

**Reading accuracy and the grounding threshold.** Hand annotation of 300 to 500 cells will fix field- and cell-level extraction accuracy, wrong-cell events included despite their invisibility to the grounding check. Threshold behaviour follows from receiver operating characteristic analysis across at least 50 to 100 fabricated and wrong-cell values.

**Outcomes.** Sensitivity to injected discrepancies and false positives per 100 pages stand as primary. After census completion, specificity, precision and F1 join as secondary, along with

class- and function-specific sensitivity, crossover reviewer time and detection, repeatability, extraction accuracy, and root cause of each miss — extraction, normalisation, consolidation, specification retrieval, comparison or formatting.

**Sample size and analysis.** For 85% sensitivity within ±5 percentage points (95% confidence), $n = 1.96^2 \times 0.85 \times 0.15 / 0.05^2 \approx 196$ reachable discrepancies under independence(62). Averaging m discrepancies per record with intraclass correlation ρ inflates this by the design effect $1 + (m - 1)\rho$; at m = 9.0 (332 discrepancies in 37 records) with assumed ρ = 0.05 the design effect is 1.40, or about 274 discrepancies. Class-level work within ±10 percentage points wants roughly 50 discrepancies per class, about 450 over nine classes. Together the three dossiers hold 332 injected discrepancies — enough overall under the stated assumptions — but 17 to 53 per class, leaving only class D1 at target. Cluster bootstrap over records will supply confidence intervals(59), while mixed-effects logistic regression with a random intercept for record estimates class and function effects.

**Release.** Source code plus a runnable container will publish the frozen harness. Where authentic material must stay closed, a synthetic or openly licensed dossier with its injection register will go out instead, preserving independent reproduction.

# 5. Results

## 5.1 Composition of the three runs

Each dossier was processed once under identical conditions. The three injection registers contained 332 discrepancies in total: 112 in dossier A, 100 in dossier B and 120 in dossier C, with 5 to 19 per class in each dossier (Table 3). In dossier A, 5 discrepancies (all D1) had no contradicting record in the package under the document-based definition, and the remaining 107 are provisionally classed as reachable (Section 4.3).

## 5.2 Sensitivity to injected discrepancies

The harness detected 95 of 112 injected discrepancies in dossier A (sensitivity 84.8%; 95% Wilson CI 77.0–90.3), 80 of 100 in dossier B (80.0%; 71.1–86.7) and 105 of 120 in dossier C (87.5%; 80.4–92.3) (Table 3, Figure 4). The random-effects pooled sensitivity was 84.1% (95% CI 72.4–91.5 with the Hartung–Knapp adjustment; 79.4–88.0 without it), with heterogeneity between dossiers low and not statistically detectable ($Q = 2.32$, $df = 2$, $p = 0.31$; $I^2 = 14\%$). The unadjusted pooled proportion was 280 of 332 (84.3%; 80.0–87.9). Sensitivity held when

restricted. In dossier A, sensitivity against the 107 provisionally reachable discrepancies was 88.8% (81.4–93.5) (Supplementary Table S11).

At least one discrepancy of every class was detected in every dossier. Grouped by the function that reaches each class and pooled across the three dossiers, the harness detected 201 of 241 discrepancies in the error-detection classes (D1, D3–D7; 83.4%, 78.2–87.6), 65 of 74 in the specification classes (D2, D8; 87.8%, 78.5–93.5) and 14 of 17 missing documents (D9; 82.4%, 59.0–93.8). Pooled class-level sensitivity ranged from 77.4% for numerical transcription (D1) to 90.9% for batch mismatch (D4). Intervals overlap widely. The classes therefore cannot be ranked.

**Table 3.** Injected discrepancies detected, by class and dossier, grouped by the function that reaches each class. Cells show discrepancies detected / injected. Denominators are all injected discrepancies. Dossier-level intervals are Wilson 95% intervals; the pooled total uses random-effects meta-analysis with the Hartung–Knapp adjustment. Class and function rows give pooled counts with Wilson intervals, are not adjusted for clustering within records or dossiers, and are exploratory.

| Class | Dossier A | Dossier B | Dossier C | Pooled | Pooled sensitivity, % (95% CI) |
|---|---|---|---|---|---|
| **Error detection** | | | | | |
| D1 Numerical transcription | 12/18 | 12/16 | 17/19 | 41/53 | 77.4 (64.5–86.5) |
| D3 Unit mismatch | 11/13 | 10/12 | 12/14 | 33/39 | 84.6 (70.3–92.8) |
| D4 Batch mismatch | 11/11 | 8/10 | 11/12 | 30/33 | 90.9 (76.4–96.9) |
| D5 Missing value | 11/13 | 10/12 | 12/14 | 33/39 | 84.6 (70.3–92.8) |
| D6 Date inconsistency | 12/13 | 10/12 | 12/14 | 34/39 | 87.2 (73.3–94.4) |
| D7 Cross-document contradiction | 10/13 | 8/11 | 12/14 | 30/38 | 78.9 (63.7–88.9) |
| **Subtotal, error detection** | **67/81** | **58/73** | **76/87** | **201/241** | **83.4 (78.2–87.6)** |
| **Specification cross-checking** | | | | | |
| D2 Specification violation | 13/13 | 10/12 | 12/14 | 35/39 | 89.7 (76.4–95.9) |
| D8 Incorrect specification | 10/12 | 8/10 | 12/13 | 30/35 | 85.7 (70.6–93.7) |
| **Subtotal, specification cross-checking** | **23/25** | **18/22** | **24/27** | **65/74** | **87.8 (78.5–93.5)** |
| **Documentation assembly** | | | | | |
| D9 Missing supporting document | 5/6 | 4/5 | 5/6 | 14/17 | 82.4 (59.0–93.8) |
| **Total, all injected** | **95/112** | **80/100** | **105/120** | **280/332** | **84.1 (72.4–91.5)** |
| Sensitivity, % (95% CI) | 84.8 (77.0–90.3) | 80.0 (71.1–86.7) | 87.5 (80.4–92.3) | 84.3 (80.0–87.9), unadjusted | — |

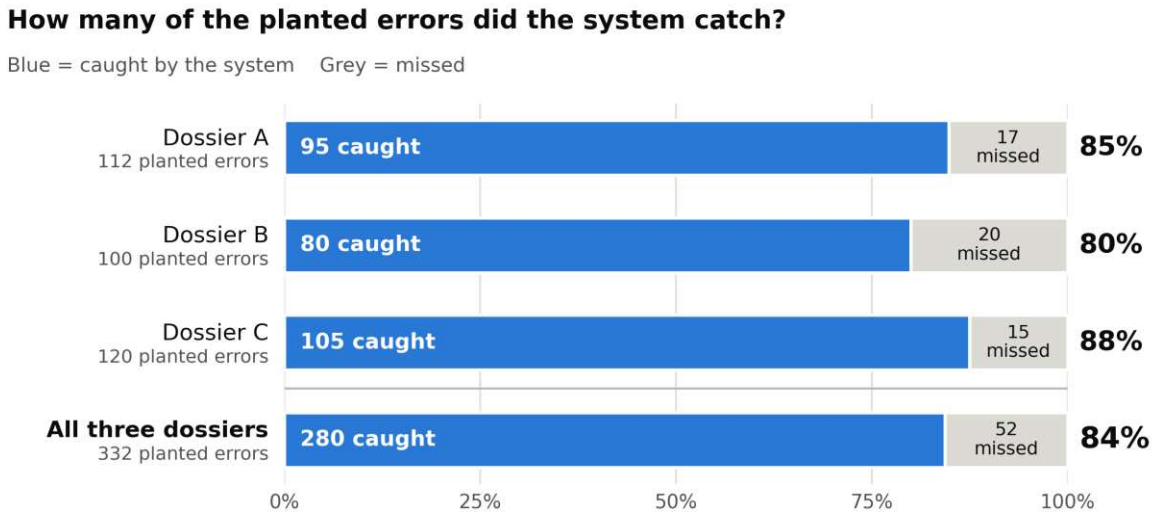



**Figure 4.** Injected discrepancies caught (blue) and missed (grey) by the system in each dossier and in all three combined. Percentages are the proportion caught, rounded to the nearest whole number; the combined figure is the unadjusted pooled proportion (280/332). Class-level counts and confidence intervals are given in Table 3.

### 5.3 Undetected discrepancies

Across the three dossiers, 52 injected discrepancies were not detected: 17 in dossier A, 20 in dossier B and 15 in dossier C. By class, the misses were D1(14), D7(8), D3(7), D5(7), D6(7), D8(6), D2(6), D4(6) and D9(6). In dossier A, 5 of the 17 misses had no contradicting record in the package, each a D1 alteration that remained within its limit, was not restated in another document and did not enter any printed statistic, so that only the raw instrument record would contradict it. The other 12 were provisionally reachable: D1(2), D3(1), D5(1), D6(2), D7(6), D8(1) and D9(2). Reachability in dossiers B and C is still unclassified. No cause analysis was done. Whether the misses reflect reading errors, rule gaps or scoring decisions is not known in any dossier.

### 5.4 Findings outside the register

The total number of findings raised in each run, and the number outside the register, were not tabulated for this analysis. That omission matters because, without these counts and the baseline census, findings outside the register cannot be classed as false positives or as genuine pre-existing defects. No false-positive rate is therefore reported.

### 5.5 An illustrative finding

A separate 48-page demonstration package was traced end to end with the answer key available. Its results show how the mechanisms operate but are not evidence of performance,

and they are reported in Supplementary Section S2. One finding shows a defect invisible to one-at-a-time checks. The specification declared a reporting threshold of 0.05% for impurities, while the validation report established a limit of quantitation of 0.08%. Each value looks plausible on its own page, yet together they commit the laboratory to reporting impurities at a level its method cannot quantify, which is not supportable under ICH Q3B(R2) and Q2(R2)(47,51). The harness bound both values to the same attribute and raised the conflict for QA verification. This single unblinded example does not establish detection frequency.

# 6. Discussion

## 6.1 What the three runs show

The three runs show that an architecture which keeps LLMs out of arithmetic, acceptance decisions and document authorship detected a similar share of discrepancies in three authentic dossiers: 80.0% to 87.5%, with a pooled estimate of 84.1% and no detectable heterogeneity between dossiers. Detection did not depend on one dossier's layout. That reassurance is narrow, however, because with only three dossiers the heterogeneity test has little power and the pooled interval (72.4–91.5%) is wide. The results do not show usefulness. A system that raised a finding on every value would reach 100% sensitivity, so 84.1% cannot be interpreted until the total findings count, and the false share, are known. At that sensitivity with an unknown false-alarm rate, the system could add work while missing one planted discrepancy in six. The practical comparison, against experienced QA reviewers under realistic time pressure, has not been made, and with no published detection rate for this task there is no alternative to compare against.

Several further features of the design limit what the figures can mean. The same team defined the taxonomy, built the harness and injected the discrepancies in all three dossiers, and there is no timestamped evidence that the harness was frozen before the registers were written, so circularity cannot be excluded. Injection density ran at 7.1 to 12.0 discrepancies per record, roughly an order of magnitude above real submissions, which may make discrepancies easier or harder to spot and would distort any false-alarm rate. Each dossier was processed once. Run-to-run variation is therefore unobserved, and agreement between dossiers shows transfer across dossiers rather than repeatability. Finally, a correctly located and classified finding counted as a detection even when its case closed as “requires QA verification” or “insufficient evidence”, and the number of such detections was not tabulated.

## 6.2 Contribution relative to prior work

Delegating computation from a language model to code is established(19–21,25), and the architecture here applies that pattern. The contribution lies elsewhere. First, the protocol in Section 4.8 defines how a reference corpus for pharmaceutical dossier verification should be built and used: a censused baseline, independent injection against a cited taxonomy, reachability defined from documents alone, expert validation with an agreement statistic, unmodified negative controls, comparators and ablations. Such a corpus, if completed and released, would be the first benchmark for this task. Second, the architecture maps each constraint onto an existing regulatory expectation: arithmetic removed from the model's tools, a source-location requirement before any value is admitted, disagreements preserved rather than resolved by frequency, and corrections running only from raw data toward summaries(43). Third, unlike table and claim verification benchmarks that check one statement against one source(27,30–32), or within-record consistency checking(33), dossier verification reconciles the same quantity across several documents under fixed source precedence and applies acceptance rules with regulatory force.

## 6.3 Regulatory and validation considerations

The separation of roles is intended to make the checking layer qualifiable by established methods. A risk-based approach under ICH Q9(R1), within the pharmaceutical quality system described in ICH Q10, would set how much assurance each component needs(63,64). The calculation modules are small, deterministic, custom (GAMP 5 Category 5) components with enumerable behaviour, the acceptance criteria are fixed data, and the control flow can be listed in advance. None of this has yet been validated. The reading layer carries different weight. Its outputs feed every check, so extraction accuracy is the main residual risk and should be treated as an AI component under the GAMP AI guidance(22) and the credibility-assessment frameworks of the FDA(23), with its error rate measured rather than assumed. Keeping the model out of decisions aligns with that intent, but whether a model that only extracts values counts as part of a critical application remains a question for the final text and for inspectors. Hosted models were used (Section 4.5). Their data-transfer, retention and confidentiality conditions, with submission-holder permission, are prerequisites for regulated use.

## 6.4 Limitations

The status of every protocol element and its consequence for interpretation is given in Table 4. The most consequential limitations are these. There is no false-positive estimate and no count of total findings, and no comparison with reviewers, baselines or ablated versions. The ground truth was written by the harness developers without independent validation. Each dossier was processed once, so repeatability is unmeasured, and intervals are not adjusted for clustering within records. The remaining gaps are narrower but still material: the test material consists of submitted dossiers rather than raw QC packages, injection density is unrealistic, and the three dossiers come from one manufacturer, are all born-digital and cover only oral solid dosage forms.

**Table 4.** Status of each element of the evaluation protocol in the present study, and what its absence means for interpretation.

| Protocol element | Status in present study | Consequence for interpretation |
|---|---|---|
| Sensitivity to injected discrepancies | Reported for three dossiers, one run each; per-dossier Wilson intervals; random-effects pooling | Within-record clustering not accounted for; class estimates exploratory |
| Total findings; findings outside the register | Not tabulated (Section 5.4) | Detection cannot be weighed against the number of alerts raised |
| Negative-control run on unmodified originals | Not performed | No false-positive estimate |
| Baseline census of pre-existing defects | Not performed | Specificity, precision and F1 cannot be computed |
| Harness frozen and hashed before injection | Not evidenced | Circularity between harness and register cannot be excluded |
| Injection by personnel independent of development | Not performed | Discrepancies may reflect the developers' expectations |
| Expert validation of the register, with kappa | Not performed | Ground truth is the developers' answer key |
| Reachability adjudicated from documents alone | Redefined for dossier A; pending for dossiers B and C | Reachable denominator provisional and reported for dossier A only |
| Repeatability, three to five runs of the same cases | Not performed (one run per dossier) | Run-to-run stability unknown |
| Human-reviewer crossover comparison | Not performed | No evidence on workload, time or relative detection |
| LLM-only and rules-only baselines | Not performed | Benefit of the architecture over simpler systems unknown |
| Ablation of grounding, consolidation and deterministic calculation | Not performed | Contribution of each safeguard untested |
| Cell-level extraction accuracy, including wrong-cell rate | Not measured | Main residual risk unquantified |
| Grounding-threshold sweep | Not performed | Threshold of 0.50 provisional |
| Cause analysis of the 52 misses | Not performed | Architectural and implementation causes cannot be separated |
| Multi-dossier corpus; realistic density; realism rating | Three dossiers from one manufacturer used; density and realism not addressed | Transfer across three dossiers observed; other manufacturers, scanned records and non-oral dosage forms untested |

| Protocol element | Status in present study | Consequence for interpretation |
|---|---|---|
| Deployment record and holder permission | Written permission obtained; model identifiers and service terms in Supplementary Table S6 | Reproducibility depends on continued availability of the hosted model versions |
| Code, container and synthetic benchmark release | Not yet released | Results cannot be reproduced independently |

## 7. Conclusion

Across three authentic dossiers, each processed once, a verification harness that confines language models to reading, and assigns all calculation and acceptance testing to deterministic code and all decisions to a qualified reviewer, detected 80.0% to 87.5% of injected discrepancies (pooled 84.1%; 95% CI 72.4–91.5), in every class in every dossier. Feasibility across dossiers is established. Usefulness is not. The false-alarm rate, repeatability, comparison with reviewers and simpler systems, and independent validation of the ground truth remain unmeasured, and the deployment and permission record must be completed before the work can be relied on. The principal contribution is the evaluation protocol, specifying a censused baseline, independent and validated injection, negative controls, comparators, ablations, repeated runs and cluster-adjusted analysis. Completing that protocol, with repeated runs, negative controls and comparators on dossiers from different manufacturers, would supply both a reference benchmark for dossier verification and the evidence needed to judge whether such systems belong in routine QA review.

# References


1. ASEAN Common Technical Dossier - ASEAN Main Portal [Internet]. [cited 2026 Sep 29]. Available from: https://asean.org/book/asean-common-technical-dossier/

2. ICH M4Q - The Common Technical Document for the registration of pharmaceuticals for human use - Quality - Scientific guideline | European Medicines Agency (EMA) [Internet]. 2003 [cited 2026 Sep 29]. Available from: https://www.ema.europa.eu/en/scientific-guidelines/ich-m4q-common-technical-document-registration-pharmaceuticals-human-use-quality

3. Haleem RM, Salem MY, Fatahallah FA, Abdelfattah LE. Quality in the pharmaceutical industry – A literature review. Saudi Pharm J. 2015 Oct 1;23(5):463–9. doi:10.1016/j.jsps.2013.11.004

4. Moorkoth S, Nayak R, B N S, Kunkalienkar S. Cognitive Factors Leading to Human Error: A Major Contributing Factor for Quality Deviations in Pharmaceutical Industry. Int J Pharm Qual Assur. 2024 Jun 25;15(02):1056–64. doi:10.25258/ijpqa.15.2.77

5. Miglani A, Saini C, Musyuni P, Aggarwal G. A review and analysis of product recall for pharmaceutical drug product. J Generic Med. 2022 Jun 1;18(2):72–81. doi:10.1177/17411343211033887

6. US Food and Drug Administration, Center for Drug Evaluation and Research. Data Integrity and Compliance With Drug CGMP: Questions and Answers [Internet]. FDA; 2022 [cited 2026 Sep 29]. Available from: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/data-integrity-and-compliance-drug-cgmp-questions-and-answers

7. Charoo NA, Khan MA, Rahman Z. Data integrity issues in pharmaceutical industry: Common observations, challenges and mitigations strategies. Int J Pharm. 2023 Jan 25;631:122503. doi:10.1016/j.ijpharm.2022.122503

8. US Food and Drug Administration.Investigating Out-of-Specification (OOS) Test Results for Pharmaceutical Production*: Guidance for Industry.* US Food and Drug Administration; May 2022..

9. Scientific and Regulatory Reasons for Delay and Denial of FDA Approval of Initial Applications for New Drugs, 2000-2012 | Clinical Pharmacy and Pharmacology | JAMA | JAMA Network [Internet]. [cited 2026 Sep 29]. Available from: https://jamanetwork.com/journals/jama/fullarticle/1817795

10. Kulkarni SB, Gaikwad VL. Common chemistry, manufacturing, and control deficiencies in abbreviated new drug applications assessed by the US Food and drug administration: Hurdle to access cost-effective medicines. J Pharmacol Toxicol Methods. 2023 Sep 1;123:107295. doi:10.1016/j.vascn.2023.107295

11. Hernán Pérez de la Ossa D, Haas F, Bream RN, Kotzagiorgis E, Tiitso K, Jekerle V. Learning from the EMA experience: common CMC deficiencies in marketing authorisation applications over the past decade. Drug Discov Today. 2025 Sep 1;30(9):104444. doi:10.1016/j.drudis.2025.104444

12. Nagy B, Galata DL, Farkas A, Nagy ZK. Application of Artificial Neural Networks in the Process Analytical Technology of Pharmaceutical Manufacturing—a Review. AAPS J. 2022 Jun 14;24(4):74. doi:10.1208/s12248-022-00706-0

13. Gaikwad AS, Pawar SB. Artificial intelligence assisted pharmaceutical analysis for sustainable quality monitoring and continuous manufacturing applications. Discov Artif Intell. 2026 Aug 16;6(1):820. doi:10.1007/s44163-026-01927-4

14. Chen T, Mo H, Wang T, Jiang C, Liu Z, Hou F, et al. A lightweight large language model for regulatory affairs translation in pharmaceutical industry. Sci Rep. 2025 Oct 30;15(1):37992. doi:10.1038/s41598-025-21867-w

15. Huanbutta K, Burapapadh K, Kraisit P, Sriamornsak P, Ganokratanaa T, Suwanpitak K, et al. Artificial intelligence-driven pharmaceutical industry: A paradigm shift in drug discovery, formulation development, manufacturing, quality control, and post-market surveillance. Eur J Pharm Sci. 2024 Dec 1;203:106938. doi:10.1016/j.ejps.2024.106938

16. Nakka G, Gupta S, Reddy KK. Artificial intelligence in pharmaceutical manufacturing: Applications, case studies, and GxP implementation considerations. J Pharm Sci. 2026 May 1;115(5). doi:10.1016/j.xphs.2026.104240 PubMed PMID: 41796681.

17. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of Hallucination in Natural Language Generation. ACM Comput Surv. 2023 Mar 3;55(12):248:1-248:38. doi:10.1145/3571730

18. 21 CFR Part 11 -- Electronic Records; Electronic Signatures [Internet]. [cited 2026 Sep 29]. Available from: https://www.ecfr.gov/current/title-21/part-11

19. Gao L, Madaan A, Zhou S, Alon U, Liu P, Yang Y, et al. PAL: Program-aided Language Models [Internet]. arXiv; 2023 [cited 2026 Sep 29]. Available from: http://arxiv.org/abs/2211.10435 doi:10.48550/arXiv.2211.10435

20. Chen W, Ma X, Wang X, Cohen WW. Program of Thoughts Prompting: Disentangling Computation from Reasoning for Numerical Reasoning Tasks. arXiv; 2023 [cited 2026 Sep 29]. Available from: http://arxiv.org/abs/2211.12588 doi:10.48550/arXiv.2211.12588

21. Yao S, Zhao J, Yu D, Du N, Shafran I, Narasimhan K, et al. ReAct: Synergizing Reasoning and Acting in Language Models. arXiv; 2023 [cited 2026 Sep 29]. Available from: http://arxiv.org/abs/2210.03629 doi:10.48550/arXiv.2210.03629

22. ISPE GAMP® Guide: Artificial Intelligence | ISPE | International Society for Pharmaceutical Engineering [Internet]. 2025 [cited 2026 Sep 29]. Available from: https://ispe.org/publications/guidance-documents/gamp-guide-artificial-intelligence

23. US Food and Drug Administration, Center for Drug Evaluation and Research. Considerations for the Use of Artificial Intelligence To Support Regulatory Decision-Making for Drug and Biological Products [Internet]. FDA; 2025 [cited 2026 Sep 29]. Available from: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/considerations-use-artificial-intelligence-support-regulatory-decision-making-drug-and-biological

24. Niazi SK. Regulatory Perspectives for AI/ML Implementation in Pharmaceutical GMP Environments. Pharmaceuticals. 2025 Jun;18(6):901. doi:10.3390/ph18060901

25. Schick T, Dwivedi-Yu J, Dessì R, Raileanu R, Lomeli M, Zettlemoyer L, et al. Toolformer: Language Models Can Teach Themselves to Use Tools [Internet]. arXiv; 2023 [cited 2026 Sep 29]. Available from: http://arxiv.org/abs/2302.04761 doi:10.48550/arXiv.2302.04761

26. Garcez A d'Avila, Lamb LC. Neurosymbolic AI: the 3rd wave. Artif Intell Rev. 2023 Nov 1;56(11):12387–406. doi:10.1007/s10462-023-10448-w

27. Thorne J, Vlachos A, Christodoulopoulos C, Mittal A. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In: Walker M, Ji H, Stent A, editors. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers) [Internet]. New Orleans, Louisiana: Association for Computational Linguistics; 2018 [cited 2026 Sep 29]. p. 809–19. Available from: https://aclanthology.org/N18-1074/ doi:10.18653/v1/N18-1074

28. Wadden D, Lin S, Lo K, Wang LL, van Zuylen M, Cohan A, et al. Fact or Fiction: Verifying Scientific Claims. In: Webber B, Cohn T, He Y, Liu Y, editors. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) [Internet]. Online: Association for Computational Linguistics; 2020 [cited 2026 Sep 29]. p. 7534–50. Available from: https://aclanthology.org/2020.emnlp-main.609/ doi:10.18653/v1/2020.emnlp-main.609

29. Chen W, Wang H, Chen J, Zhang Y, Wang H, Li S, et al. TabFact: A Large-scale Dataset for Table-based Fact Verification [Internet]. arXiv; 2020 [cited 2026 Sep 29]. Available from: http://arxiv.org/abs/1909.02164 doi:10.48550/arXiv.1909.02164

30. Aly R, Guo Z, Schlichtkrull M, Thorne J, Vlachos A, Christodoulopoulos C, et al. FEVEROUS: Fact Extraction and VERification Over Unstructured and Structured information [Internet]. arXiv; 2021 [cited 2026 Sep 29]. Available from: http://arxiv.org/abs/2106.05707 doi:10.48550/arXiv.2106.05707

31. Chen Z, Chen W, Smiley C, Shah S, Borova I, Langdon D, et al. FinQA: A Dataset of Numerical Reasoning over Financial Data. In: Moens MF, Huang X, Specia L, Yih SW tau, editors. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing [Internet]. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics; 2021 [cited 2026 Sep 29]. p. 3697–711. Available from: https://aclanthology.org/2021.emnlp-main.300/ doi:10.18653/v1/2021.emnlp-main.300

32. Zhu F, Lei W, Huang Y, Wang C, Zhang S, Lv J, et al. TAT-QA: A Question Answering Benchmark on a Hybrid of Tabular and Textual Content in Finance. In: Zong C, Xia F, Li W, Navigli R, editors. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) [Internet]. Online: Association for Computational Linguistics; 2021 [cited 2026 Sep 29]. p. 3277–87. Available from: https://aclanthology.org/2021.acl-long.254/ doi:10.18653/v1/2021.acl-long.254

33. proceedings.com [Internet]. [cited 2026 Sep 29]. EHRCon: Dataset for Checking Consistency between Unstructured Notes and Structured Tables in Electronic Health Records. Available from: https://www.proceedings.com/079017-2835.html

34. Xu Y, Li M, Cui L, Huang S, Wei F, Zhou M. LayoutLM: Pre-training of Text and Layout for Document Image Understanding. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining [Internet]. New York, NY, USA: Association for Computing Machinery; 2020 [cited 2026 Sep 29]. p. 1192–200. (KDD '20). Available from: https://dl.acm.org/doi/10.1145/3394486.3403172 doi:10.1145/3394486.3403172

35. Xu Y, Xu Y, Lv T, Cui L, Wei F, Wang G, et al. LayoutLMv2: Multi-modal Pre-training for Visually-rich Document Understanding. In: Zong C, Xia F, Li W, Navigli R, editors. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) [Internet]. Online: Association for Computational Linguistics; 2021 [cited 2026 Sep 29]. p. 2579–91. Available from: https://aclanthology.org/2021.acl-long.201/ doi:10.18653/v1/2021.acl-long.201

36. Zhong X, ShafieiBavani E, Jimeno Yepes A. Image-Based Table Recognition: Data, Model, and Evaluation. In: Vedaldi A, Bischof H, Brox T, Frahm JM, editors. Computer Vision – ECCV 2020. Cham: Springer International Publishing; 2020. p. 564–80. doi:10.1007/978-3-030-58589-1_34

37. Mathew M, Karatzas D, Jawahar CV. DocVQA: A Dataset for VQA on Document Images. In: 2021 IEEE Winter Conference on Applications of Computer Vision (WACV) [Internet]. 2021 [cited 2026 Sep 29]. p. 2199–208. Available from: https://ieeexplore.ieee.org/document/9423358 doi:10.1109/WACV48630.2021.00225

38. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks [Internet]. arXiv; 2021 [cited 2026 Sep 29]. Available from: http://arxiv.org/abs/2005.11401 doi:10.48550/arXiv.2005.11401

39. Es S, James J, Espinosa Anke L, Schockaert S. RAGAs: Automated Evaluation of Retrieval Augmented Generation. In: Aletras N, De Clercq O, editors. Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations [Internet]. St. Julians, Malta: Association for Computational Linguistics; 2024 [cited 2026 Sep 29]. p. 150–8. Available from: https://aclanthology.org/2024.eacl-demo.16/ doi:10.18653/v1/2024.eacl-demo.16

40. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021 Mar 29;372:n71. doi:10.1136/bmj.n71 PubMed PMID: 33782057.

41. Yu LX. Pharmaceutical Quality by Design: Product and Process Development, Understanding, and Control. Pharm Res. 2008 Apr 1;25(4):781–91. doi:10.1007/s11095-007-9511-1

42. ICH Q6A specifications: test procedures and acceptance criteria for new drug substances and new drug products: chemical substances - Scientific guideline | European Medicines

Agency (EMA) [Internet]. 1999 [cited 2026 Sep 30]. Available from: https://www.ema.europa.eu/en/scientific-guidelines/ich-q6a-specifications-test-procedures-acceptance-criteria-new-drug-substances-new-drug-products-chemical-substances

43. 21 CFR 211.194 -- Laboratory records. [Internet]. [cited 2026 Sep 30]. Available from: https://www.ecfr.gov/current/title-21/part-211/section-211.194

44. ICH Q1A (R2) Stability testing of new drug substances and drug products - Scientific guideline | European Medicines Agency (EMA) [Internet]. 2003 [cited 2026 Sep 30]. Available from: https://www.ema.europa.eu/en/scientific-guidelines/ich-q1a-r2-stability-testing-new-drug-substances-drug-products

45. GOV.UK [Internet]. 2021 [cited 2026 Sep 29]. Guidance on GxP data integrity. Available from: https://www.gov.uk/government/publications/guidance-on-gxp-data-integrity

46. PIC/S GOOD PRACTICES FOR DATA MANAGEMENT AND INTEGRITY IN REGULATED GMP/GDP ENVIRONMENTS (PI 041-1) - ECA Academy [Internet]. [cited 2026 Sep 30]. Available from: https://www.gmp-compliance.org/guidelines/gmp-guideline/pic-s-good-practices-for-data-management-and-integrity-in-regulated-gmp-gdp-environments-pi-041-1

47. ICH Q2(R2) Validation of analytical procedures - Scientific guideline | European Medicines Agency (EMA) [Internet]. 1994 [cited 2026 Sep 30]. Available from: https://www.ema.europa.eu/en/scientific-guidelines/ich-q2r2-validation-analytical-procedures

48. ICH Q1E Evaluation of stability data - Scientific guideline | European Medicines Agency (EMA) [Internet]. 2003 [cited 2026 Sep 30]. Available from: https://www.ema.europa.eu/en/scientific-guidelines/ich-q1e-evaluation-stability-data

49. US Food and Drug Administration, Center for Drug Evaluation and Research. Process Validation: General Principles and Practices [Internet]. FDA; 2020 [cited 2026 Sep 29]. Available from: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/process-validation-general-principles-and-practices

50. Shabir GA. Validation of high-performance liquid chromatography methods for pharmaceutical analysis: Understanding the differences and similarities between validation requirements of the US Food and Drug Administration, the US Pharmacopeia and the International Conference on Harmonization. J Chromatogr A. 2003 Feb 14;26th International Symposium on High Performance Liquid Phase Separations and Related Techniques987(1):57–66. doi:10.1016/S0021-9673(02)01536-4

51. ICH Q3B (R2) Impurities in new drug products - Scientific guideline | European Medicines Agency (EMA) [Internet]. 2006 [cited 2026 Sep 30]. Available from: https://www.ema.europa.eu/en/scientific-guidelines/ich-q3b-r2-impurities-new-drug-products

52. General Chapter <711> Dissolution. In: United States Pharmacopeia and National Formulary [Internet]. Rockville, MD: United States Pharmacopeial Convention. Available from: https://www.uspnf.com/

53. <905> Uniformity of Dosage Units. In: United States Pharmacopeia and National Formulary [Internet]. Rockville, MD: United States Pharmacopeial Convention. Available from: https://www.uspnf.com/

54. 21 CFR 211.192 -- Production record review. [Internet]. [cited 2026 Sep 30]. Available from: https://www.ecfr.gov/current/title-21/part-211/section-211.192

55. Probable Inference, the Law of Succession, and Statistical Inference: Journal of the American Statistical Association: Vol 22, No 158 [Internet]. [cited 2026 Sep 29]. Available from: https://www.tandfonline.com/doi/abs/10.1080/01621459.1927.10502953

56. DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials. 1986 Sep 1;7(3):177–88. doi:10.1016/0197-2456(86)90046-2

57. Hartung J, Knapp G. A refined method for the meta-analysis of controlled clinical trials with binary outcome. Stat Med. 2001;20(24):3875–89. doi:10.1002/sim.1009

58. Quantifying heterogeneity in a meta-analysis - Higgins - 2002 - Statistics in Medicine - Wiley Online Library [Internet]. [cited 2026 Sep 29]. Available from: https://onlinelibrary.wiley.com/doi/10.1002/sim.1186

59. Field CA, Welsh AH. Bootstrapping Clustered Data. J R Stat Soc Ser B Stat Methodol. 2007 Jun 1;69(3):369–90. doi:10.1111/j.1467-9868.2007.00593.x

60. A Coefficient of Agreement for Nominal Scales - Jacob Cohen, 1960 [Internet]. [cited 2026 Sep 29]. Available from: https://journals.sagepub.com/doi/10.1177/001316446002000104

61. Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull. 1971;76(5):378–82. doi:10.1037/h0031619

62. Buderer NMF. Statistical Methodology: I. Incorporating the Prevalence of Disease into the Sample Size Calculation for Sensitivity and Specificity. Acad Emerg Med. 1996;3(9):895–900. doi:10.1111/j.1553-2712.1996.tb03538.x

63. US Food and Drug Administration, Center for Drug Evaluation and Research.Q9(R1) Quality Risk Management [Internet]. FDA; 2026 [cited 2026 Sep 30]. Available from: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/q9r1-quality-risk-management

64. Q10 Pharmaceutical Quality System | FDA [Internet]. [cited 2026 Sep 30]. Available from: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/q10-pharmaceutical-quality-system

914

915

916

917

918

919

920

921

922

923

924

925

926

927

928

929

930

## Declarations

**Data availability.** The source dossiers are confidential to their holders and cannot be shared publicly. The de-identified records and the injection registers are available from the corresponding author on reasonable request, subject to written approval from each holder, an independent review of residual identifiability, and completion of the baseline census and expert validation (Section 4.8). A synthetic benchmark dossier with its injection register and the evaluation scripts will be deposited on Zenodo under a CC BY 4.0 licence, and the DOI will be provided before publication.

**Code availability.** The source code of the harness, calculation modules, grounding check and evaluation scripts will be released on GitHub under the Apache License 2.0 and archived on Zenodo. The repository address, DOI, commit hash and container image digest of the evaluated version will be provided before publication.

**Pre-registration.** The prospective evaluation (Section 4.8) will be registered on the Open Science Framework before the next evaluation begins.

**Permissions and ethics.** No human participants were involved and no personal data were processed; named individuals in the source records were replaced with role tokens before processing. Written permission was obtained from the holder of each source submission, covering the processing of its de-identified pages by third-party hosted language models.

**Funding.** This research received no external funding; the authors funded the work themselves.

**Competing interests.** All authors intend to commercialise the verification harness described in this paper. No author has any other competing interest to declare.

**Author contributions.** Anil Bhusal: Conceptualization, Methodology. Diwash Kunwar: Software. Puskar Niroula: Methodology, Investigation, Formal analysis. Shashwot Rimal: Writing – original draft, Writing – review and editing.

**Informed consent.** The subject-matter expert involved in this study, the Head of Quality Assurance of the company holding the source dossiers, gave written informed consent to his involvement and to publication. His role was advisory validation and was limited to reviewing (i) the problem statement, that is, whether the manual QC-to-QA transfer and checking of dossier data is a real problem for industry; (ii) the proposed approach; and (iii) the results of the approach, with a view to whether such an approach could reduce the labour

and human hours spent on dossier verification. He did not inject discrepancies, score detections or perform the statistical analysis, which were carried out by the authors.

**Use of artificial intelligence in preparing this manuscript.** The authors used generative AI tools to generate Figures 1, 2 and 4 and the first version of Figure 3. Claude (Anthropic) was used to review the manuscript, revise its text and redraw Figure 3. The authors checked all text and every figure against the underlying data and take full responsibility for the content of the publication.

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AppendixALiteraturesearchprotocol.docx
- AppendixBThecontrolledreferencedataset.docx
- fullAppendixLiteraturesearchprotocol.docx
- Algorithms.docx
- SupplementaryMaterial.docx