Journal of Evidence-Based Medical Research
e-ISSN 3107-1996 · jebmr.com
2026;2(3):105–114 · Article 11

[ORIGINAL RESEARCH] OPEN ACCESS

Artificial Intelligence in Pharmacovigilance: Evaluating Opportunities for Automation, Human Oversight Requirements, and Limitations of Autonomous Decision-Making

Harshavardhan Gangavarapu[1]

[1] United Therapeutics
Correspondence: hrshvrdhn7@gmail.com

Abstract

Background: Artificial intelligence (AI) may accelerate pharmacovigilance, but the point at which efficiency gains become incompatible with accountable safety decisions remains uncertain.

Objective: To compare automation, oversight intensity, decision quality, and residual risk across pharmacovigilance tasks in a matched scenario-based benchmark.

Methods: A structured dataset contained 4,096 evaluation records comprising 1,024 scenarios repeated under autonomous AI, human-on-the-loop, human-in-the-loop, and human-in-command arrangements. Sixteen tasks were grouped into five workflow stages. Final correctness, residual high-impact errors, predefined autonomy eligibility, policy adherence, time, and modeled cost were assessed at the scenario level.

Results: Final decision correctness was 77.05% (789/1,024) with autonomous AI, 87.50% (896/1,024) with human-on-the-loop oversight, 94.53% (968/1,024) with human-in-the-loop verification, and 96.00% (983/1,024) under human-in-command governance. Residual high-impact errors affected 12.30%, 6.15%, 3.12%, and 1.46% of scenarios, respectively. Aggregate time savings versus an assigned manual baseline ranged from 95.15% for autonomy to 17.68% for human-in-command. Only 13 scenarios (1.27%) were coded as eligible for autonomous use.

Conclusion: The benchmark reveals a pronounced efficiency–safety trade-off and supports risk-stratified automation of bounded processing tasks, with stronger human control for interpretive and consequential decisions. These scenario-based findings are not evidence of real-world effectiveness or regulatory compliance without external validation.

Keywords: pharmacovigilance, artificial intelligence, human oversight, drug safety, signal detection, automation governance, risk-stratified validation.

Received: 10 November 2026
Revised: 19 November 2026
Accepted: 20 November 2026
Published: 21 November 2026

Cite this article
Gangavarapu H. Artificial Intelligence in Pharmacovigilance: Evaluating Opportunities for Automation, Human Oversight Requirements, and Limitations of Autonomous Decision-Making. J Evid-Based Med Res. 2026;2(3):105–114.

© The Author(s) 2026. Published by Acadentium Research. This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0), which permits unrestricted use, distribution and reproduction in any medium, provided the original work is properly cited.

1. Introduction

Pharmacovigilance integrates the detection, assessment, understanding, and prevention of adverse effects and other drug-related problems. Its evidentiary infrastructure combines individual case safety reports (ICSRs), medical literature, electronic health records, registries, and spontaneous-reporting databases. These sources are heterogeneous, incomplete, and frequently affected by reporting delays. Under-reporting is a longstanding constraint, while statistical signal detection is intended to prioritize assessment rather than establish causation [1], [2]. Growing volumes of narrative reports make manual extraction and triage resource intensive; however, a faster process is not necessarily a safer one when errors involve serious events, emerging signals, or regulatory recommendations.

Artificial intelligence (AI) has been investigated for record retrieval, de-duplication, terminology coding, seriousness classification, narrative extraction, and the identification of drug–event relationships. These functions differ in how explicitly they can be specified and audited. A structured code assignment may be verifiable against a dictionary; a causal judgment demands integration of chronology, competing explanations, pharmacological plausibility, and evidence quality. Early work therefore advocated augmentation rather than wholesale substitution of pharmacovigilance specialists [3]–[5]. Advances in transformers and large language models have expanded the amount of unstructured information that can be processed, but performance on one extraction benchmark cannot be presumed to generalize to signal validation or benefit–risk decisions [6]–[9].

The methodological hazards of spontaneous-report analyses provide a further reason to separate signal

generation from action. Confounding, selective reporting, masking, inaccurate coding, and incomplete documentation can distort apparent drug–event associations. Even correctly executed disproportionality analyses are generally hypothesis-generating and do not replace patient-level clinical assessment [10]–[12]. Accordingly, decision quality should be assessed jointly with the severity of residual errors, traceability, escalation completion, and the allocation of final authority. An aggregate accuracy statistic can conceal unacceptable performance in the subset where patient safety is most vulnerable.

Recent empirical studies have demonstrated feasible automated extraction of case features, clinically documented adverse events, and safety-relevant narratives, while systematic reviews continue to emphasize inconsistent external validation and limited prospective assessment [13]–[22]. CIOMS has articulated the importance of risk-based governance, human oversight, validity, transparency, privacy, fairness, and accountability. Contemporary proposals likewise distinguish computable routine tasks from decisions that require contextual professional judgment [23], [24]. Yet comparative evidence across alternative oversight models remains less developed than task-level performance reporting. The present analysis addresses that operational question in a structured, matched scenario benchmark, without treating simulated or assigned workflow metrics as measured clinical outcomes.

2. Study Objectives and Research Questions

The primary objective was to estimate the difference in final decision correctness and residual high-impact errors between autonomous AI and three human oversight arrangements using the same pharmacovigilance scenarios. The secondary objectives were to compare stage-specific reliability, characterize correction and harmful-override patterns, quantify modeled time and cost trade-offs, and examine the relationship between preassigned autonomy eligibility and governance-policy adherence. The practical contribution is a decision-oriented framework that makes the conditions for automation explicit rather than inferring them from pooled accuracy alone.

Four research questions guided the analysis: (RQ1) How much does final accuracy change with oversight intensity on matched tasks? (RQ2) How does oversight affect residual high-impact errors, escalations, and erroneous human overrides? (RQ3) Which workflow stages and tasks exhibit the greatest reliability gradient? (RQ4) What efficiency is surrendered when final decision authority remains with qualified reviewers? The working expectation was that greater oversight would improve final correctness and reduce consequential errors while diminishing modeled throughput and financial savings.

3. Methods

3.1 Study design and analytical unit

This study was a retrospective, scenario-based secondary analysis of a structured pharmacovigilance evaluation workbook. The file contained 4,096 records and 55 variables, including scenario identifiers, task labels, proposed and final decisions, case attributes, oversight conditions, timing, costing, and governance flags. Each of 1,024 unique scenario identifiers appeared exactly four times, once under each oversight arrangement. The scenario, not the individual record, was therefore the independent matching unit. The dataset did not provide the provenance of individual reference decisions, a recruitment frame, a validated sampling protocol, or documentary evidence of prospective system deployment. Consequently, all outcomes are interpreted as benchmark characteristics rather than empirical clinical effectiveness estimates.

3.2 Workflow taxonomy and comparators

Sixteen tasks were organized into collection, structuring, synthesis, interpretation, and decision stages. Collection comprised ICSR intake, literature safety extraction, and real-world data screening; structuring covered MedDRA coding, duplicate detection, seriousness classification, and ICSR data-quality verification. Synthesis comprised signal prioritization, evidence synthesis, and aggregate trend assessment. Interpretation covered clinical context, drug–event causality, and signal validation. The decision stage included benefit–risk recommendations, regulatory action triage, and risk-minimization recommendations. Each task was represented by 64 unique scenarios, yielding 192 collection, 256 structuring, and 192 scenarios in each remaining stage (Table 1).

The four comparator labels encoded distinct governance designs: autonomous AI (no human review); human-on-the-loop (monitoring with exception review); human-in-the-loop (pre-action verification); and human-in-command (human final-authority assessment). The underlying AI proposal, confidence value, reference decision, and scenario characteristics were invariant across the four records within each scenario. Review actions, final decisions, and workflow resource estimates could differ. This matched structure isolates observed differences between dataset-defined oversight conditions but does not show that reviewers were independently randomized, blinded, or studied in a live organizational setting.

3.3 Outcomes and operational definitions

The primary effectiveness endpoint was final correctness, defined as agreement between the recorded final decision and the recorded reference decision. Initial AI correctness compared the AI proposal with that same reference. An AI error was corrected when an initially incorrect proposal became correct after oversight; a harmful override occurred when an initially correct proposal became incorrect. Residual high-impact error was determined from the dataset's binary flag, not retrospectively inferred from all incorrect responses. Severity analyses also used the assigned low, moderate, high, and critical safety-impact categories. Therefore, residual high-impact error is a subset-specific flag with a different denominator and interpretation from overall inaccuracy.

Secondary outcomes were human review and intervention, completed escalations among cases marked

as requiring them, preassigned eligibility for autonomous use, dataset-coded policy adherence, and evidence traceability. The time comparison used the supplied manual baseline seconds and total workflow seconds; cost comparisons used the provided manual and total workflow cost fields. Aggregate savings were calculated as one minus the ratio of summed workflow time or cost to summed baseline time or cost. These are modeled workflow estimates and omit downstream consequences of incorrect decisions, implementation, model maintenance, validation, and remediation costs. Confidence was examined descriptively using the supplied probability-like score; its derivation, calibration procedure, and relationship to an independently estimated probability were unavailable.

3.4 Data integrity and analysis

Scenario repetition, one-to-one representation of the four oversight conditions, and invariance of reference decisions and AI predictions were verified before aggregation. Agreement fields were cross-checked against their source decisions; no disagreement was observed for the binary correctness flags. Descriptive frequencies were computed per scenario and per governance model. Paired differences were expressed in percentage points relative to autonomous AI, without treating the 4,096 rows as 4,096 independent observations. Wilson intervals describe accuracy precision conditional on the benchmark; exploratory paired bootstrap intervals were generated by resampling the 1,024 scenario identifiers with replacement 3,500 times, preserving all four observations within each resampled scenario.

Because the selection process, data-generating mechanism, comparator implementation, and independence from scenario construction were not documented, these intervals were not interpreted as population-level confidence statements. No hypothesis-testing claim of clinical superiority was made. Results were stratified by workflow stage, impact tier, and routine-versus-stress-test subset. Visualizations and governance options were derived from the observed task taxonomy and previously published regulatory principles, not from a prospectively tested deployment policy. This distinction limits the permissible inference from any apparent performance ranking.

The analysis distinguished four types of evidence that are often combined in automation claims. First, technical correctness was measured against a recorded reference label; it did not quantify all potential errors in source interpretation. Second, the high-impact-error endpoint incorporated the workbook's severity designation rather than assigning equal clinical consequence to each mismatch. Third, governance adherence and escalation completion described conformity with the scenario's encoded workflow requirements, not independent compliance with external regulation. Fourth, time and cost metrics assessed the resources attributed to the defined processing path, excluding the broader costs of deploying or correcting an AI system. Keeping these constructs separate was necessary because a workflow might improve one without improving the others.

Comparative rates were calculated using equal scenario weights, reflecting the benchmark's balanced task construction rather than actual pharmacovigilance caseloads. Results were not extrapolated to national reporting volumes or estimated drug-safety outcomes. No post hoc threshold for an acceptable severe-error frequency was selected from the observed results, since such limits require clinical risk assessment and institutional authorization. In addition, a confidence score was never treated as a substitute for a verified evidence trail or a completed required escalation. These safeguards prevent descriptive comparisons from being converted into unsupported clinical recommendations.

3.5 Ethical, governance, and reporting considerations

The analytical file contained scenario identifiers and categorical evaluations rather than identifiable patient information. No intervention or patient recruitment was conducted as part of this secondary analysis. An institutional ethics determination or dataset-origin statement was not included with the source file. The analysis was informed by transparent reporting principles for AI evaluation and human–AI systems; guidance written for clinical prediction models or interventional trials was treated as relevant methodological context, not as an assertion that this benchmark met those study designs [25]–[27].

4. Results

4.1 Benchmark composition

The 1,024 scenarios were balanced evenly across the 16 named tasks. A total of 768 (75.0%) were labeled routine evaluations and 256 (25.0%) were stress tests. Each scenario generated four oversight records, so task and stage percentages refer to unique scenarios unless otherwise noted. Safety-impact tiers were low in 239 (23.3%), moderate in 346 (33.8%), high in 307 (30.0%), and critical in 132 (12.9%). Overall, 439/1,024 scenarios had high or critical impact classifications. The scenario set contained 201 (19.6%) emerging-signal indicators and 500 (48.8%) records per model requiring escalation. AI proposals were initially correct in 789/1,024 scenarios (77.05%) and incorrect in 235 (22.95%), before any human review.

4.2 Decision correctness and correction dynamics

Final correctness increased in the ordered comparison from autonomous AI (789/1,024; 77.05%) to human-on-the-loop (896/1,024; 87.50%), human-in-the-loop (968/1,024; 94.53%), and human-in-command (983/1,024; 96.00%) (Fig. 2; Table 2). Relative to autonomy, the paired accuracy differences were +10.45, +17.48, and +18.95 percentage points, respectively. Exploratory scenario-level paired-bootstrap intervals for these differences were approximately 8.50–12.50, 15.04–20.12, and 16.41–21.68 percentage points. These estimates compare recorded outcomes within a matched, non-representative scenario set and should not be interpreted as external clinical validation.

Among 235 initially erroneous AI proposals, human-on-the-loop corrected 116 (49.4%), human-in-the-loop

corrected 191 (81.3%), and human-in-command corrected 211 (89.8%). However, reviewers also converted some initially correct AI proposals into incorrect final decisions: nine, 12, and 17 under these three conditions, respectively. Thus, additional review reduced net error but did not eliminate review-associated harm. A policy of universal human review should not be equated with an infallible human reference standard. It needs training, evidence presentation, override justification, and its own quality assurance.

4.3 Consequential errors and stage heterogeneity

Residual high-impact errors occurred in 126 autonomous scenarios (12.30%), 63 human-on-the-loop scenarios (6.15%), 32 human-in-the-loop scenarios (3.12%), and 15 human-in-command scenarios (1.46%) (Table 2). Among the 439 high- or critical-impact scenarios, the corresponding counts were 126, 63, 32, and 15, giving conditional rates of 28.7%, 14.4%, 7.3%, and 3.4%. The absolute decrease between autonomous and human-in-command conditions was 10.84 percentage points over the full benchmark, with an exploratory paired-bootstrap interval of approximately 8.79–12.89 percentage points for the reduction.

Performance varied considerably by workflow stage (Fig. 4; Table 3). Autonomous accuracy was 89.06% for collection, 80.47% for structuring, 80.21% for synthesis, 70.83% for interpretation, and 63.54% for decision tasks. Under human-in-command governance, stage accuracies ranged from 94.27% to 97.40%. This pattern was more pronounced in particular consequential tasks: autonomous correctness was 35/64 (54.69%) for risk-minimization recommendation, 40/64 (62.50%) for regulatory action triage, and 41/64 (64.06%) for signal validation, compared with 60/64 (93.75%) for ICSR intake extraction and 59/64 (92.19%) for literature extraction. Such gradients argue against assigning a common autonomy threshold to all pharmacovigilance functions.

4.4 Autonomy eligibility, governance, and stress testing

Only 13/1,024 scenarios (1.27%) carried the dataset's affirmative autonomous-use eligibility flag; seven were collection scenarios and six were structuring scenarios. No synthesis, interpretation, or decision scenarios were flagged as eligible. The flag is a preassigned benchmark property rather than an independently validated clinical rule. Dataset-coded policy adherence was 1.27% in the autonomous condition, 37.99% with human-on-the-loop monitoring, 70.31% with human-in-the-loop review, and 90.72% with human final authority. Among 500 scenarios requiring escalation, completion was recorded in none under autonomy, 304 under human-on-the-loop, 469 under human-in-the-loop, and all 500 under human-in-command. These fields reflect the constructed governance criteria and cannot establish legal or regulatory compliance.

Autonomous correctness was 78.65% among routine scenarios and 72.27% among stress tests. Corresponding accuracies were 88.54% and 84.38% for human-on-the-loop, 94.40% and 94.92% for human-in-the-loop, and 95.70% and 96.88% for human-in-command. The relatively small stress-test decrement in the latter conditions supports a possible protective role for review within this benchmark, but stress-test design was not independently specified. AI confidence averaged 86.3% while initial accuracy was 77.1%; 89 of 823 predictions with recorded confidence of at least 80% were wrong. Confidence therefore cannot be treated as a sufficient authorization criterion without prospective calibration and high-consequence subgroup evaluation.

4.5 Modeled time and cost

The assigned manual baseline averaged 1,240 seconds per scenario. Mean total workflow times were 60 seconds for autonomous AI, 232 seconds for human-on-the-loop, 593 seconds for human-in-the-loop, and 1,021 seconds for human-in-command. Aggregate time reductions versus the same baseline were 95.15%, 81.31%, 52.16%, and 17.68%, respectively. Aggregate workflow cost savings were 99.80%, 84.05%, 54.09%, and 17.94% (Fig. 3; Table 2). These values reflect the workbook's specified processing duration, reviewer time, staff rates, and overhead; they are not observed budgetary savings and make no allowance for false-negative harm, unnecessary actions, liability, audit infrastructure, or system development. An autonomous workflow can appear inexpensive precisely because it excludes corrective review.

5. Discussion

5.1 Principal findings and contribution

This matched benchmark shows that the apparent value of AI depends on the outcome used to define value. Autonomy led on modeled throughput but ranked last on final accuracy and consequential-error suppression. Human-in-command improved final accuracy by almost 19 percentage points and reduced the residual high-impact error rate by about 11 points, but most of the modeled labor saving disappeared. The intermediate conditions represented meaningful, distinct compromises. The study's central contribution is not evidence that one model is universally optimal; it is an empirically grounded separation of technical correctness, safety impact, oversight-policy performance, and operating resources in a single task-level comparison. This prevents a favorable accuracy or cost statistic from concealing unacceptable decisions at the interpretive end of pharmacovigilance.

5.2 Automation should follow task computability, not model branding

Routine extraction, taxonomy coding, and data cleaning have comparatively clear targets and can be checked against source material. This accords with nationwide report-coding evaluations, clinical note extraction studies, and recent NLP applications that demonstrate concrete capabilities but also substantial variability across settings and labels [4]–[6], [16]–[20]. Conversely, causality attribution, signal validation, benefit–risk assessment, and risk minimization require assessment of alternative explanations and the consequences of acting on incomplete evidence. Wu and colleagues distinguish computable tasks from those that cannot yet be reduced to reliable algorithmic procedures; the present stage

gradient is compatible with this distinction but does not independently establish its universality [24]. Even for structured tasks, accuracy on internal or historical examples should not be mistaken for robust performance during shifts in drug classes, languages, coding ontologies, or reporting conventions.

5.3 The limitations of autonomous decisions are asymmetric

A false-positive administrative coding error may be reversible; a missed emerging safety signal or an unsupported recommendation may delay investigation or affect clinical communication. The benchmark's high- and critical-impact cases accounted for all flagged residual high-impact errors, underscoring the need to preserve explicit safety severity rather than report a single mean error rate. Fundamental limitations of spontaneous reporting also persist irrespective of the model architecture. Signal scores can be biased by reporting behavior, confounding, duplicate reports, and notoriety effects; they are not estimates of attributable clinical risk [2], [10]–[12]. Thus, an algorithm capable of quickly ranking signals may still lack the evidence required to validate causation or determine an appropriate regulatory response. Autonomy cannot correct a deficient epistemic basis merely by accelerating its processing.

5.4 Human oversight is not a binary control

The three oversight configurations performed differently because they assign different opportunities and authority for intervention. Exception monitoring preserved more of the automated time advantage but left comparatively more high-impact errors and incomplete escalations. Pre-action verification substantially improved final correctness while maintaining about half of the aggregate baseline time saving. Human final authority produced the lowest flagged residual risk, particularly in the high-consequence categories, at much higher modeled labor consumption. The benchmark therefore supports an escalation ladder in which task risk, uncertainty, source traceability, and potential consequences determine review intensity. The ladder is a design hypothesis, not an empirically validated selection algorithm (Fig. 1).

The recorded harmful overrides are equally instructive. Human review is a safety control only when the reviewer possesses an appropriate evidentiary record, enough time, relevant expertise, and an effective mechanism to challenge the system's output. Automation bias may make an apparently confident recommendation unusually persuasive; conversely, unstructured reviewer discretion can reverse a correct assessment. A sound governance process should document whether a proposal was accepted, modified, or rejected, with a rationale and a reproducible audit trail. It should monitor reviewer-specific error rates and escalation failures as well as model errors. Reviews of AI-enabled decision support underscore the gap between promising computational performance and demonstrated effects on human work and clinical care [28], [29].

5.5 Confidence, explainability, and the conditions for trust

The discrepancy between recorded confidence and observed correctness demonstrates why score thresholds require calibration. An average confidence of 86% does not guarantee that similarly scored cases are correct at that rate, and high confidence did not prevent mistakes in this benchmark. Selective automation would require calibration plots, subgroup-specific false-negative rates, uncertainty under distribution shift, and explicit safe-abstention procedures. The reported explainability and traceability fields in the workbook are not equivalent to independent proof that the conclusion is justified by clinical evidence. Hauben has questioned the appropriate role of explainability in pharmacovigilance, and work in clinical AI cautions that plausible explanations can produce misplaced trust without validation [30], [31].

An additional concern is unequal performance by source, therapeutic domain, language, and patient population. A reference decision may itself reflect historical reporting and classification biases; models optimized to mimic those labels can reproduce them at scale. General healthcare algorithm research has demonstrated that apparently reasonable surrogate outcomes may generate material inequities, a principle that warrants explicit fairness checks in pharmacovigilance systems [32]. Subgroup audits, differential missingness assessment, documented model changes, and post-deployment surveillance should therefore be prerequisites for expanding autonomous scope rather than optional interpretability enhancements.

5.6 Translation into governance and regulatory practice

A practical operating standard would require predeployment validation for each intended task and source, with locked versioning and acceptance criteria chosen by clinical safety leaders. For low-consequence, bounded tasks, validated AI outputs could be automatically accepted if source evidence is retrievable, data-quality rules pass, confidence is calibrated, no privacy exception arises, and periodic sampling remains active. Uncertain or incomplete cases should enter a specialist review queue. Clinical interpretation, signal validation, benefit–risk recommendation, or action triage should retain documented expert authority, ideally with second-level review for disagreements involving critical events. The autonomy-eligibility field in this dataset already reflects an exceptionally narrow set, but its accuracy and fairness cannot be certified from the workbook alone.

The EMA reflection paper situates AI within risk-proportionate pharmaceutical governance, and CIOMS recommends human oversight, validity, transparency, privacy, and accountability throughout the pharmacovigilance lifecycle [23], [33]. European AI legislation includes requirements for oversight of systems legally classified as high-risk, but it would be incorrect to presume that every pharmacovigilance application is automatically within that category; classification depends on actual intended use and applicable provisions [34]. WHO similarly advocates human-centered governance and protection of health-related rights [35]. Together these instruments support a governance objective: automate bounded activities while preserving accountable judgment where errors may materially affect safety decisions. They do not, by themselves, validate the specific thresholds suggested by this benchmark.

5.7 Methodological limitations

Several constraints materially narrow interpretation. First, the data-generating procedure was undocumented and the repeated matched structure, common AI proposals, preset timing values, and policy flags are consistent with a designed scenario evaluation; it is not possible to verify how realistic those scenarios are. Second, neither source documents nor clinical adjudication protocols were available to validate reference decisions, severity classifications, or the correction process. Third, oversight modes were preassigned labels rather than prospectively observed randomized interventions. Reviewer roles and processing times may encode assumptions that mechanically favor one arrangement. Consequently, causal attribution of improvement to human governance cannot be established.

Fourth, the benchmark contains one recorded model version and no evidence of external testing, calibration against actual case reports, temporal drift, multilingual performance, or operational failure modes. Fifth, sample balance across tasks is artificial relative to unknown real-world task frequencies, so pooled percentages cannot be used to forecast deployment accuracy without reweighting. Sixth, residual high-impact error and policy-adherence variables are dataset-specified outcomes, not independent safety-committee adjudications. Seventh, financial savings are incomplete by design: they do not include implementation, oversight infrastructure, error management, or the clinical and regulatory consequences of mistakes. The precision intervals reflect only resampling uncertainty inside the scenario set. These limitations preclude claims of regulatory conformity, comparative clinical effectiveness, or economic return on investment.

5.8 Implications and future research

A defensible validation pathway should begin with a preregistered, multi-institutional holdout dataset containing diverse drug classes, event types, adverse-reaction narratives, languages, and sources. At least two qualified drug-safety adjudicators should resolve reference decisions with explicit adjudication rules, and blinded reviewers should evaluate all four governance strategies under realistic workloads. Primary outcomes should include sensitivity for serious and emerging events, false-negative rates, time to validated signal, harmful overrides, escalation completion, evidence citation fidelity, reviewer workload, and downstream decisions. Workflows should be stress-tested under source missingness, rare events, distribution shift, and intentionally misleading narrative content; cost-effectiveness analyses should incorporate downstream harms and quality-control overhead. Prospective multisite studies, following relevant AI reporting principles, would be needed before considering autonomy expansion [25]–[27], [29].

Studies should also test whether a risk-weighted routing policy outperforms fixed governance categories. Such a policy could minimize reviewer burden subject to prespecified ceilings for severe false negatives, minimum review coverage of critical cases, and auditability constraints. Fairness and privacy evaluations should examine who is most exposed to missed or delayed signals, not merely aggregate model discrimination. Until these questions are addressed, success should be defined by reliable human–AI cooperation and better clinical safety decisions rather than the removal of human labor as an end in itself [21]–[23].

6. Conclusion

In 1,024 matched pharmacovigilance scenarios, increasing human oversight was associated with higher recorded final correctness and fewer residual high-impact errors, while reducing modeled time and cost savings. The largest reliability deficiencies under autonomy appeared in interpretive and decision-stage activities. A narrow subset of routine collection and structuring tasks was designated eligible for autonomous use, but those labels were not independently validated. The results support a risk-stratified architecture: automation for bounded and auditable processing, specialist verification for exceptions and uncertainty, and qualified human final authority for consequential judgments. Validation against independently adjudicated real-world cases and prospective operational evaluation are required before these benchmark comparisons can inform clinical or regulatory deployment.

Declarations

Ethics: No identifiable patient data or direct human intervention was involved in this secondary scenario-level analysis.

Consent: Not applicable to the analysis of the non-identifiable scenario-level records provided.

Funding: No funding

Competing interests: None.

References

1. L. Hazell and S. A. W. Shakir, “Under-reporting of adverse drug reactions: a systematic review,” Drug Saf, vol. 29, no. 5, pp. 385–396, 2006, doi:10.2165/00002018-200629050-00003.

2. O. Caster, K. Juhlin, S. Watson, and G. N. Norén, “Improved statistical signal detection in pharmacovigilance by combining multiple strength-of-evidence aspects in vigiRank,” Drug Saf, vol. 37, no. 8, pp. 617–628, 2014, doi:10.1007/s40264-014-0204-5.

3. J. K. Aronson, “Artificial intelligence in pharmacovigilance: an introduction to terms, concepts, applications, and limitations,” Drug Saf, vol. 45, no. 5, pp. 407–418, 2022, doi:10.1007/s40264-022-01156-5.

4. G. L. Martin, J. Jouganous, R. Savidan, et al., “Validation of artificial intelligence to support the automatic coding of patient adverse drug reaction reports, using nationwide pharmacovigilance data,” Drug Saf, vol. 45, no. 5, pp. 535–548, 2022, doi:10.1007/s40264-022-01153-8.

5. T. Botsis and K. Kreimeyer, “Improving drug safety with adverse event detection using natural language processing,” Expert Opin Drug Saf, vol. 22, no. 8, pp. 659–668, 2023, doi:10.1080/14740338.2023.2228197.

6. A. L. Silverman, M. Sushil, B. Bhasuran, et al., “Algorithmic identification of treatment-emergent adverse events from clinical notes using large language models: a pilot study in inflammatory bowel disease,” Clin Pharmacol Ther, vol. 115, no. 6, pp. 1391–1399, 2024, doi:10.1002/cpt.3226.

7. S. Modi, K. A. Kasmiran, N. M. Sharef, and M. Y. Sharum, “Extracting adverse drug events from clinical notes: a systematic review of approaches used,” J Biomed Inform, vol. 151, Art. no. 104603, 2024, doi:10.1016/j.jbi.2024.104603.

8. M. K. Desai, "Artificial intelligence in pharmacovigilance: opportunities and challenges," Perspect Clin Res, vol. 15, no. 3, pp. 116–121, 2024, doi:10.4103/picr.picr_290_23.

9. E. Kopacheva, A. Lincke, O. Björneld, and T. Hammar, "Detecting adverse drug events in clinical notes using large language models," Stud Health Technol Inform, vol. 327, pp. 892–893, 2025, doi:10.3233/SHTI250495.

10. P. M. Cutroneo, D. Sartori, M. Tuccori, et al., "Conducting and interpreting disproportionality analyses derived from spontaneous reporting systems," Front Drug Saf Regul, vol. 3, Art. no. 1323057, 2023, doi:10.3389/fdsfr.2023.1323057.

11. M. Fusaroli, D. Sartori, E. P. van Puijenbroek, and G. N. Norén, "Charting and sidestepping the pitfalls of disproportionality analysis," Drug Saf, vol. 49, no. 2, pp. 147–162, 2026, doi:10.1007/s40264-025-01604-y.

12. E. Scosyrev, S. Behr, D. Jain, A. Ponnuru, and C. Michel, "Disproportionality analysis and causal inference in drug safety," Pharmaceut Med, vol. 39, no. 2, pp. 97–107, 2025, doi:10.1007/s40290-024-00549-4.

13. N. Chalabianloo, F. Ahmadi, M. A. Omrani, et al., "Machine learning methods for predicting adverse drug events: a systematic review," Br J Clin Pharmacol, vol. 92, no. 2, pp. 422–444, 2026, doi:10.1002/bcp.70377.

14. A. Nagar, J. Gobburu, and A. Chakravarty, "Artificial intelligence in pharmacovigilance: advancing drug safety monitoring and regulatory integration," Ther Adv Drug Saf, vol. 16, Art. no. 20420986251361435, 2025, doi:10.1177/20420986251361435.

15. J. Warner, L. H. De Souza Teodoro, A. P. Jardim, and C. Albera, "Intelligent automation improved efficiency in pharmacovigilance safety signal assessment," Clin Pharmacol Ther, 2026, doi:10.1002/cpt.70409.

16. D. Ludwig, M. Wang, J. Buchanan, and T. Trinh, "A large language model for extracting post-marketing adverse drug events from clinical notes in the electronic health record," Drug Saf, vol. 49, no. 10, pp. 1181–1198, 2026, doi:10.1007/s40264-026-01682-6.

17. A. Franchini, R. Noseda, J. Cornelius, et al., "NLP in support of pharmacovigilance: QUality Adverse Drug Reaction AcTIve Control (QUADRATIC)," Clin Pharmacol Ther, vol. 119, no. 6, pp. 1597–1603, 2026, doi:10.1002/cpt.70250.

18. J. M. Plasek, Y. Li, M. G. Amato, et al., "Relationship extraction for adverse drug events in clinical notes using large language models: a retrospective cross-sectional study," EBioMedicine, vol. 133, Art. no. 106516, 2026, doi:10.1016/j.ebiom.2026.106516.

19. S. G. Lim and M. A. Ismail, "Automated multi-label adverse drug reaction extraction from Malaysia's national pharmacovigilance narratives," J Med Syst, vol. 50, Art. no. 120, 2026, doi:10.1007/s10916-026-02448-4.

20. A. Katz, A. Dhankar, G. Fransoo, et al., "Using natural language processing to identify adverse drug events characterized by medication replacement in primary care electronic medical records: algorithm and validation study," J Med Internet Res, vol. 28, Art. no. e86467, 2026, doi:10.2196/86467.

21. I. El Sayed, R. N. Bedwany, M. Salama, and A. S. H. Eltrass, "Integration of artificial intelligence into pharmacovigilance: a systematic review," Res Social Adm Pharm, 2026, doi:10.1016/j.sapharm.2026.09.001.

22. T. Y. Kim, W. S. Oh, and D. H. Jeong, "Large language models in adverse drug reaction detection and pharmacovigilance: a systematic review of current applications, challenges, and future directions," Diagnostics, vol. 16, no. 15, Art. no. 2435, 2026, doi:10.3390/diagnostics16152435.

23. G. N. Norén, J. Durand, V. Kara, et al., "Artificial intelligence in pharmacovigilance: guiding principles from the CIOMS Working Group XIV," Drug Saf, 2026, doi:10.1007/s40264-026-01708-z.

24. L. Wu, J. Xu, O. Dang, and R. Ball, "Does generative AI mean the "end of history" for pharmacovigilance automation? Towards a framework for the future of human–AI systems," Front Drug Saf Regul, vol. 6, Art. no. 1846339, 2026, doi:10.3389/fdsfr.2026.1846339.

25. G. S. Collins, K. G. M. Moons, P. Dhiman, et al., "TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods," BMJ, vol. 385, Art. no. e078378, 2024, doi:10.1136/bmj-2023-078378.

26. X. Liu, S. Cruz Rivera, D. Moher, M. J. Calvert, and A. K. Denniston, "Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension," Nat Med, vol. 26, no. 9, pp. 1364–1374, 2020, doi:10.1038/s41591-020-1034-x.

27. S. Cruz Rivera, X. Liu, A. W. Chan, et al., "Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension," Nat Med, vol. 26, no. 9, pp. 1351–1363, 2020, doi:10.1038/s41591-020-1037-7.

28. C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King, "Key challenges for delivering clinical impact with artificial intelligence," BMC Med, vol. 17, Art. no. 195, 2019, doi:10.1186/s12916-019-1426-2.

29. K. Lekadir, A. F. Frangi, A. R. Porras, et al., "FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare," BMJ, vol. 388, Art. no. e081554, 2025, doi:10.1136/bmj-2024-081554.

30. M. Hauben, "Artificial intelligence in pharmacovigilance: do we need explainability? Pharmacoepidemiol Drug Saf," 2022;31(12):1311–1316, doi:10.1002/pds.5501.

31. M. Ghassemi, L. Oakden-Rayner, and A. L. Beam, "The false hope of current approaches to explainable artificial intelligence in health care," Lancet Digit Health, vol. 3, no. 11, pp. e745–e750, 2021, doi:10.1016/S2589-7500(21)00208-9.

32. Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," Science, vol. 366, no. 6464, pp. 447–453, 2019, doi:10.1126/science.aax2342.

33. European Medicines Agency, "Reflection paper on the use of artificial intelligence (AI) in the medicinal product lifecycle," EMA/CHMP/CVMP/83833/2023. 2024. https://www.ema.europa.eu/en/scientific-guidelines/use-artificial-intelligence-ai-medicinal-product-lifecycle.

34. European Parliament and Council, "Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)," Off J Eur Union. 2024. https://eur-lex.europa.eu/eli/reg/2024/1689/oj.

35. World Health Organization, "Ethics and governance of artificial intelligence for health: WHO guidance," Geneva: WHO; 2021. ISBN 9789240029200. https://www.who.int/publications/i/item/9789240029200.

Table 1. MATCHED BENCHMARK DESIGN, SCENARIO ATTRIBUTES, AND ANALYTICAL ENDPOINTS

Characteristic	Unique scenarios / definition
Total evaluations	4,096 rows = 1,024 scenarios × 4 oversight configurations
Workflow tasks	16 (64 scenarios per task)
Workflow stages	Collect: 192; Structure: 256; Synthesize: 192; Interpret: 192; Decide: 192
Routine / stress-test	768 / 256
Impact tiers	Low: 239; moderate: 346; high: 307; critical: 132
Emerging-signal flag	201/1,024 (19.6%)
Cases requiring escalation	500/1,024 (48.8%)
Underlying AI version	PV-AI-1.0 (single recorded version)
Initial AI proposal correct	789/1,024 (77.05%)
Primary endpoint	Final decision matches reference decision
Safety endpoint	Binary residual high-impact error flag supplied for each oversight evaluation
Autonomous-use eligible	13/1,024 (1.27%) (assigned indicator)

Note. Counts are for distinct matched scenarios unless noted; governance records are not independent patient observations.

Table 2. FINAL DECISION PERFORMANCE, RESIDUAL RISK, GOVERNANCE, AND MODELED RESOURCE USE

Oversight	Final correct	High-impact errors	Interventions	Policy adherence	Time saved	Cost saved
Autonomous AI	789/1024 (77.05%)	126 (12.30%)	0	1.27%	95.15%	99.80%
Human-on-the-loop	896/1024 (87.50%)	63 (6.15%)	125	37.99%	81.31%	84.05%
Human-in-the-loop	968/1024 (94.53%)	32 (3.12%)	203	70.31%	52.16%	54.09%
Human-in-command	983/1024 (96.00%)	15 (1.46%)	228	90.72%	17.68%	17.94%

Note. Time and cost savings are calculated from assigned baseline and workflow values, not observed economic effects. Policy adherence is an encoded dataset label.

Table 3. WORKFLOW-STAGE DECISION ACCURACY AND RESIDUAL HIGH-IMPACT ERRORS

Stage	n	Autonomous	Human-on-loop	Human-in-loop	Human-in-command	Auto harm	Command harm
Collect	192	89.06%	94.79%	97.40%	96.88%	7	2
Structure	256	80.47%	90.23%	94.92%	94.92%	17	2
Synthesize	192	80.21%	88.02%	93.23%	97.40%	19	1
Interpret	192	70.83%	82.81%	93.75%	96.88%	38	4
Decide	192	63.54%	80.73%	93.23%	94.27%	45	6

Note. High-impact error columns show stage-specific counts for the autonomous and human-in-command conditions; denominators differ by stage.

Table 4. GOVERNANCE DIAGNOSTICS, ESCALATION, AND DECISION REVERSAL

Measure	Autonomous	Human-on-loop	Human-in-loop	Human-in-command
Initially wrong AI proposals	235	235	235	235
Wrong proposals corrected	0	116	191	211
Correct proposals changed to wrong	0	9	12	17
Human reviewed (n)	0	551	1,024	1,024
Required escalations completed	0/500	304/500	469/500	500/500
Autonomy-eligible scenarios	13	13	13	13
High/critical-tier residual errors	126/439	63/439	32/439	15/439

Note. Error corrections and harmful overrides were counted by comparing final correctness with the matched initial AI correctness status.

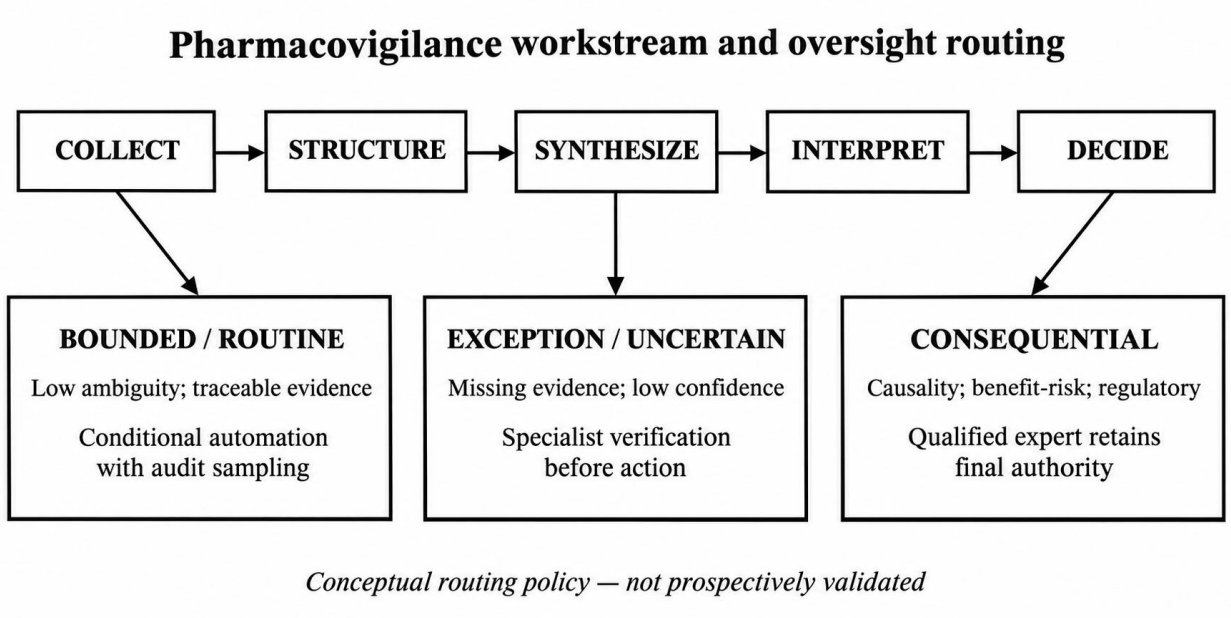


Figure 1. Proposed risk-stratified governance pathway across pharmacovigilance stages. The decision rules are conceptual and were not validated as an operational policy.

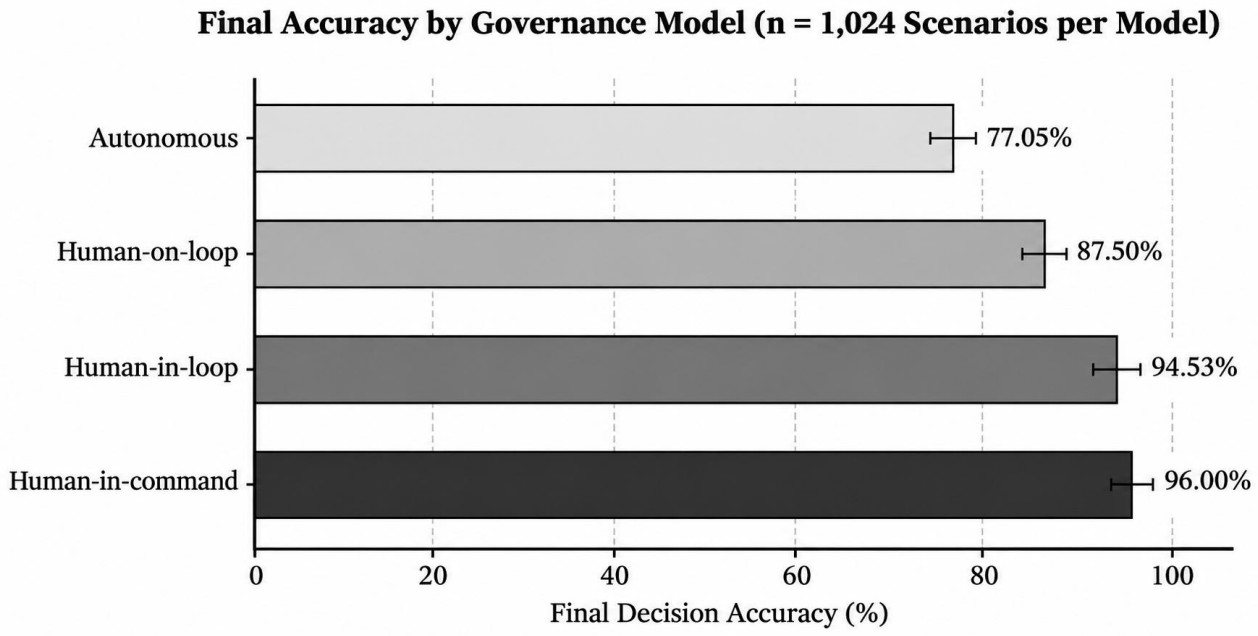


Figure 2. Recorded final decision accuracy by oversight configuration. Error bars are descriptive Wilson 95% intervals conditional on the 1,024-scenario benchmark.

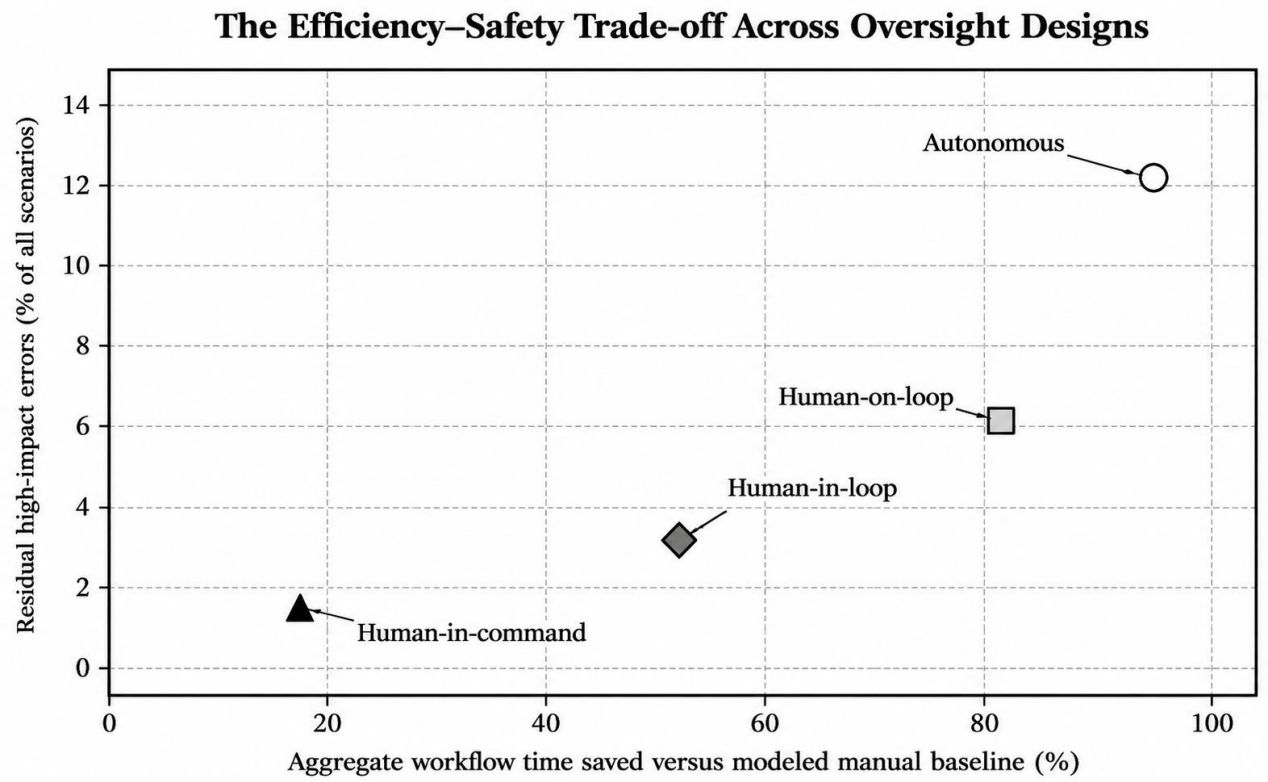


Figure 3. Aggregate modeled time savings versus residual high-impact-error frequency. Percentage savings exclude costs of downstream error management and system deployment.

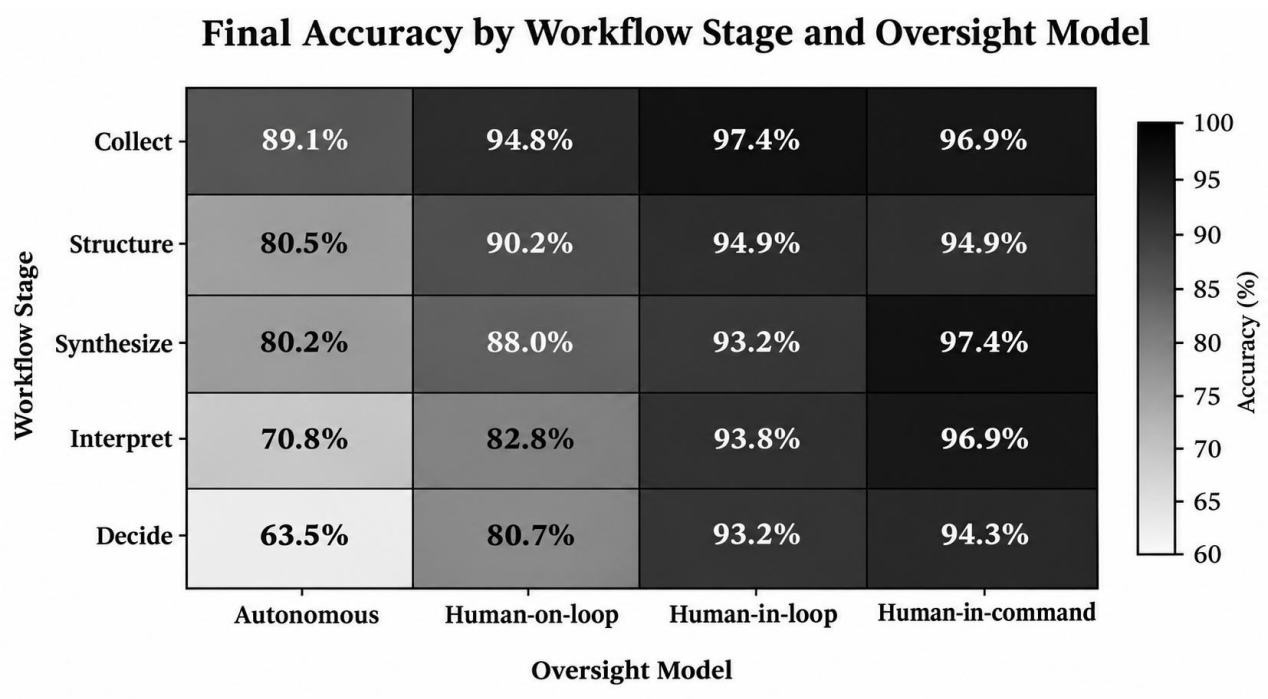


Figure 4. Stage-specific final decision accuracy across oversight models. Stage-specific sample sizes are 192 except for structuring (256).