METARIGOR: FROM AGENTIC AUTOMATION TO VERIFIABLE END-TO-END META-ANALYSIS THROUGH SCIENTIFIC DECISION BOUNDARIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Automated meta-analysis requires not only accurate numerical extraction but also the consistent preservation of result identity across stages, including the associated study, comparison, population, outcome, time point, and analysis conditions. We present MetaRigor, an LLM workflow framework for verifiable end-to-end evidence synthesis. MetaRigor organizes study results as Study $\times$ Target units and operationalizes scientific decision boundaries at the task level through Task Contracts, programmatic rules, and source binding. The model performs semantic interpretation that depends on study context, whereas programs execute formalizable rules, computations, validation, and source verification. We conducted controlled evaluations across five stages: evidence acquisition, data extraction, risk of bias assessment, certainty of evidence assessment, and manuscript generation. MetaRigor achieved an F_1 score of 86.29% for complete result delivery and 80.00% agreement for domain-level certainty of evidence judgments. Generated manuscripts received scores of 94.63 and 88.25 under DeepSeek and GPT-5.6 Luna evaluation, respectively, both exceeding the generic CLI baseline. Cross-system tracing of archived outputs from three systems further showed that local mismatches or omissions in result identity could alter statistical inputs and the objects subjected to risk of bias assessment, with these errors subsequently propagating into the final report. Ablation experiments showed that task constraints and programmatic rules played distinct roles across different stages. These findings indicate that, in LLM-driven evidence synthesis workflows, consistently maintaining explicit result identity, decision boundaries, and provenance relationships is critical to improving end-to-end verifiability. Source code is available at `https://anonymous.4open.science/status/Meta-Rigor-8C62`.

1 INTRODUCTION

Meta-analysis typically synthesizes quantitatively comparable results from independent studies within a systematic review to estimate an overall effect and characterize between study heterogeneity and uncertainty. A complete meta-analysis workflow extends beyond statistical synthesis to encompass multiple interdependent stages, including research question formulation, literature retrieval and screening, data extraction, risk of bias assessment, certainty of evidence assessment, and result reporting, as described by (Higgins et al., 2024; Page et al., 2021). Both the Cochrane Handbook and PRISMA 2020 regard study identification, selection, appraisal, and synthesis as core components of systematic reviews. In recent years, large language models have advanced evidence synthesis toward multistage automation. Manalyzer organizes analysis through a multiagent workflow (Xu et al., 2025), TrialMind connects question-guided retrieval, screening, and extraction (Wang et al., 2025b), and AutoSynthesis further integrates structured retrieval, result extraction, and statistical synthesis (Taherinezhad et al., 2026). Figure 1 summarizes this progression from human-led workflows to AI-assisted systems and general-purpose agents. As automation increases, a fundamental problem becomes increasingly prominent: how can a system ensure that every stage consistently operates on the correct study object, uses only the facts admissible to the task, and follows the methodological rules applicable to the decision being made?

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

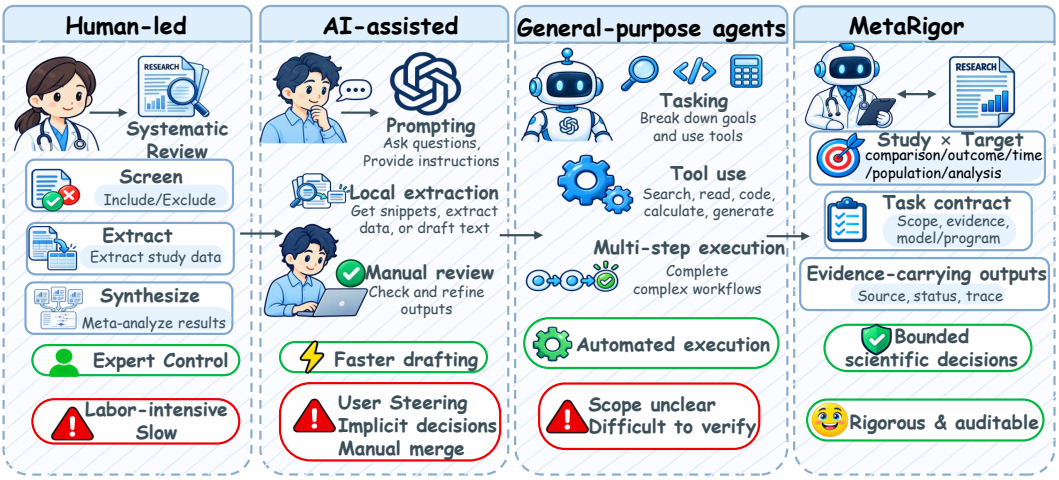


Figure 1: Evolution of meta-analysis paradigms. As meta-analysis evolves from human-led workflows to AI-assisted and general-purpose agents, automation increases while scientific decision boundaries may remain implicit. MetaRigor makes study identity, task constraints, and evidence provenance explicit to support rigorous and verifiable end-to-end evidence synthesis.

A central challenge in maintaining such consistency is that reports, studies, and target results do not have a one-to-one correspondence. Information from a single study may be distributed across the primary report, supplementary materials, trial registries, and follow-up publications. Consequently, deduplicating reports cannot substitute for correctly linking multiple reports to the same underlying study, as discussed by Li et al. (2024). The Cochrane Handbook explicitly treats the study, rather than the report, as the fundamental unit of a systematic review and requires multiple reports from the same study to be appropriately linked. A single outcome may also be reported under different comparisons, populations, time points, and analytical conditions. Extracting the correct numerical value therefore does not necessarily mean that the result required by the Protocol has been identified, as emphasized by McKenzie et al. (2024). Relevant methodological guidance likewise requires comparisons, outcomes, time points, and procedures for selecting among multiple eligible results to be prespecified. Once result identity is misaligned, downstream statistical synthesis, methodological assessment, and interpretation may diverge from the original research question.

We collectively refer to the object of judgment, the admissible facts, and the decision authority assigned to models and programs as the Scientific Decision Boundary. This concept defines the constraints required to maintain consistency across stages. For example, determining whether a comparison, population, or time point matches the target requires a model to interpret the study context. By contrast, candidate enumeration, numerical computation, source verification, and formalizable methodological rules are executed programmatically. These boundaries must also persist into downstream stages to prevent the generation process from altering established result identities or outputs produced by rule-based procedures.

Based on this principle, we propose MetaRigor, which operationalizes the Scientific Decision Boundary through structured representations and execution constraints. The system begins with a researcher-confirmed Protocol. It uses a Study Packet to consolidate reports from multiple sources that correspond to the same study, and represents each target result as a Study $\times$ Target pair defined by the comparison, population, outcome, time point, and analytical conditions. A Task Contract specifies the object of judgment, admissible facts, and delivery requirements for each task, while allocating responsibilities between models and programs. MetaRigor further binds evidence to retrievable locations in the original source, allowing correspondence among results, judgments, and supporting evidence to be preserved and verified across five stages: evidence acquisition, data extraction, risk of bias assessment, certainty assessment, and manuscript generation.

We evaluate MetaRigor at both stage and cross-stage propagation levels. A controlled five-stage evaluation compares recovery of study materials, result delivery, methodological judgments, and manuscript quality under fixed upstream inputs. Ablation experiments examine the stage-specific

contributions of key constraints, while source verifiability is evaluated independently. We further analyze continuous pipeline execution and archived artifacts from three systems to assess preservation of result identity and state across stages, and trace how local mismatches or omissions propagate into downstream statistical inputs, risk of bias assessment targets, and final reporting.

The main contributions of this work are as follows:

- We introduce a result representation centered on Study $\times$ Target, which preserves result identity together with its defining study conditions across material organization, result selection, methodological assessment, and report generation.
- We operationalize Scientific Decision Boundaries as task constraints and execution mechanisms that explicitly divide responsibilities between models and programs, while supporting cross-stage evidence verification through source binding.
- We establish a controlled five-stage evaluation framework with end-to-end artifact analysis. Comparative and ablation studies quantify key constraint contributions, while propagation analysis reveals how result identity deviations affect downstream evidence synthesis.

2 Related Work

Automated Evidence Synthesis Systems. Existing approaches to evidence synthesis automation have expanded from literature screening to retrieval, study identification, and structured information extraction. ASReview uses active learning to improve the efficiency of literature screening, while Trialstreamer automatically identifies clinical trials and extracts PICO elements, sample sizes, and key results Van De Schoot et al. (2021); Marshall et al. (2020). More recently, TrialMind has integrated research question guided retrieval, eligibility assessment, and target result extraction, while LEADS further decomposes literature mining into specialized tasks spanning study search, eligibility assessment, and structured data extraction Wang et al. (2025b;c). These systems primarily improve the efficiency and accuracy of information acquisition for individual tasks, but generally do not explicitly represent multiple reports from the same study and their multiple candidate results as persistent objects across the workflow. MetaRigor consolidates study materials through the Study Packet and fixes result identity through Study $\times$ Target, allowing that identity to continuously constrain downstream extraction, methodological assessment, and report generation.

Scientific Reasoning and Evidence Verification. Automated scientific workflows require not only information extraction but also methodological judgments that depend on study context and explicit verification of supporting evidence. ROBOTO2 combines document retrieval, responses to RoB 2 signaling questions, and human review to support risk of bias assessment. EviSearch, ALCE, and related studies further examine the correspondence between generated outputs and supporting evidence, while distinguishing content correctness from citation quality (Hevia et al., 2025; Ahuja et al., 2026; Gao et al., 2023). Scientific judgments, however, involve different decisions. Matching study semantics requires model-based interpretation, whereas rule execution, computation, and source verification can be performed programmatically. Without an explicit separation of the object of judgment, admissible facts, and decision authority, individual steps may be correct in isolation while remaining inconsistent across stages. MetaRigor addresses this issue through the Scientific Decision Boundary and Task Contract, which assign responsibilities to models and programs, while source binding preserves traceability between outputs and the original evidence.

General-Purpose Agent Frameworks. In recent years, general-purpose agent frameworks have extended the ability of large language models to perform complex tasks through tool use, code execution, and multistep planning (Yao et al., 2022; Yang et al., 2024). Systems such as Codex CLI and OpenHands provide execution environments for software engineering and general-purpose tasks, enabling agents to invoke external tools and complete long-horizon tasks Wang et al. (2025a). These frameworks, however, are primarily designed to improve task execution capability, whereas scientific workflows additionally require explicit study objects, evidence boundaries, and methodological rules. The execution capabilities of general-purpose agents alone are therefore insufficient to ensure the reliability and validity of scientific outputs. MetaRigor adds a constraint layer for scientific tasks on top of general agent capabilities. It organizes agent behavior through structured study representations, Task Contracts, and verification mechanisms, while supporting integration with external agents through a CLI interface for practical scientific workflows.

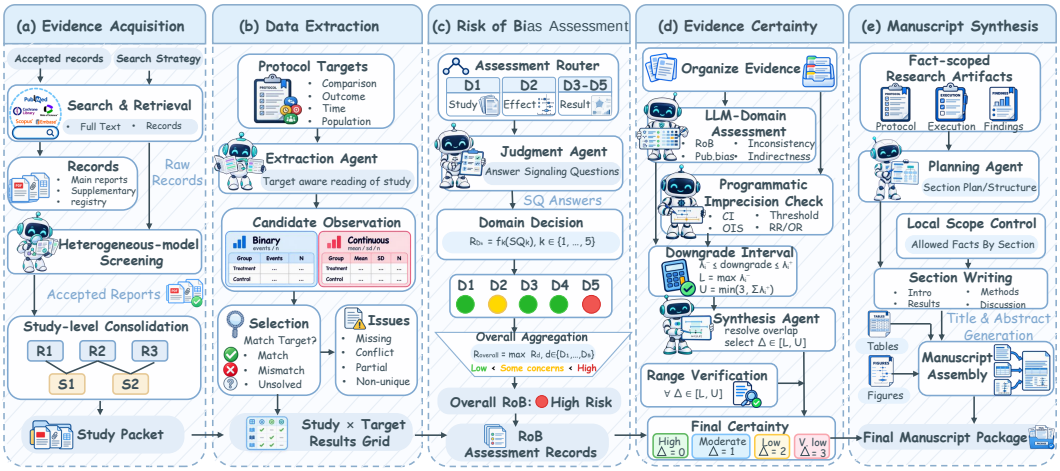


Figure 2: End-to-end evidence synthesis workflow of MetaRigor. (a) Evidence acquisition retrieves and consolidates eligible reports into Study Packets. (b) Data extraction maps observations to protocol-defined Study × Target results. (c) Risk of bias assessment produces structured domain judgments and overall ratings. (d) Certainty assessment integrates domain judgments into final certainty ratings. (e) Manuscript generation composes verified artifacts under explicit fact scopes.

3 METHOD

As illustrated in Figure 2, MetaRigor begins with a researcher-confirmed Protocol and organizes five stages of evidence synthesis, including evidence acquisition (EA), data extraction (DE), risk of bias assessment (RoB), certainty of evidence assessment (EC), and manuscript generation (MS), by maintaining a unified result identity, partitioning decision responsibilities, and binding outputs to source evidence. The system implements a domain-specific execution harness for evidence synthesis that centrally manages Task Contracts, model invocation, programmatic validation, and the storage of stage-specific research artifacts. MetaRigor also provides a command-line interface (CLI) that allows external agents to invoke the capabilities of individual stages. Throughout execution, the system continuously maintains study objects, source relationships, and unresolved issues, ensuring that research artifacts produced across stages remain consistent and traceable.

3.1 RESULT REPRESENTATION AND THE SCIENTIFIC DECISION BOUNDARY

MetaRigor first defines the research objects shared across downstream stages. The system uses the study as the unit for organizing materials and identifies each result by a combination of study and target. Given a study set $\mathcal{S}$ and a target set $\mathcal{T}$, the primary report, supplementary materials, registry records, and follow-up reports associated with the same study are consolidated into a Study Packet. Reports and studies are allowed to have many-to-many correspondences. A target $t = (\kappa, o, \tau, p, a)$ specifies the comparison, outcome, time point, population, and analytical conditions, respectively. The Cartesian product $\mathcal{G} = \mathcal{S} \times \mathcal{T}$ defines the result index space. This representation permits partial results, missingness, and ambiguity, without assuming that complete data exist for every Study × Target combination. Study consolidation preserves report identities and the evidence supporting their linkage. Studies with unresolved identities or insufficient materials remain unresolved.

On top of this unified result identity, MetaRigor implements contract-driven execution through the Task Contract. Each contract specifies the object of judgment, admissible facts, decision authority assigned to models and programs, delivery requirements, and applicable validation rules. Formalizable constraints are checked programmatically. Models handle factual interpretation and semantic selection that depend on understanding the study context. Programs execute identifier validation, numerical computation, source verification, and formalized methodological rules. Accordingly, DE preserves the identities of the study, target, and group roles. RoB constrains the scope of assessment. MS distinguishes among the prespecified protocol, actual study conduct, and study findings.

To support provenance-aware execution, the evidence interface binds task outputs to retrievable locations in original sources. Each output is stored together with its source identifier, source location, retrieved text, and associated hashes, enabling downstream inspection and tracing. Let $\mathcal{D}_j$ denote the set of source identifiers permitted for task j, and let c denote a citation submitted by the model. Parse resolves the source d and location ℓ, while Read retrieves the corresponding source text q:

$$B_{\mathcal{D}_j}(c) = \begin{cases} (d, \ell, q), & \text{Parse}(c) = (d, \ell),\ d \in \mathcal{D}_j,\ q = \text{Read}(d, \ell) \neq \perp \\ \perp, & \text{otherwise} \end{cases}, \tag{1}$$

where $\perp$ denotes failure. Parsing failure, violation of the permitted source boundary, and retrieval failure are recorded as issues. The source location, retrieved text, and associated hashes are preserved with the output. The Verifier checks formatting requirements and deterministic constraints. The Reviewer Agent provides semantic feedback for Human Decision without modifying artifacts or labels, and was not used in experiments reported in this work. We therefore distinguish technical source closure from evidence-level support. Technical source closure verifies whether an output is linked to a valid and retrievable source location, whereas reference recall evaluates whether the retrieved evidence overlaps with manually identified supporting evidence spans. Figure 3 illustrates how these structured artifacts are transformed and propagated across evidence synthesis stages. The Study $\times$ Target representation provides a shared interface linking result alignment, statistical synthesis, risk-of-bias assessment, and certainty assessment while preserving provenance.

3.2 Evidence Acquisition and Result Construction

This stage performs two successive identity mappings in sequence. EA links dispersed reports to their corresponding studies, after which DE maps candidate observations within each study to the predefined Study $\times$ Target combinations specified by the Protocol.

EA constructs the analysis materials through screening and study consolidation. FAST and DEEP use heterogeneous model configurations and make separate judgments at both the title and abstract stage and the full-text stage. This design gives the two screening pathways different inductive biases and reduces the likelihood of highly homogeneous judgments. At the title and abstract stage, a record proceeds to full-text assessment only when both pathways complete successfully and explicitly judge it eligible. At the full-text stage, FAST and DEEP again make separate judgments. Concordant judgments determine the decision, whereas disagreements are referred to an independent adjudicator. The organizational unit then shifts from the report to the study. Primary reports, supplementary materials, registry records, and follow-up materials are linked to the corresponding Study Packet, while reports whose study identity cannot be established remain unresolved.

DE represents observations reported in the source separately from Protocol-directed result selection. An observation records the comparison, outcome, time point, population, analytical conditions, group roles, numerical values, and provenance actually reported in the source. Selection determines whether a given observation is compatible with the specified Target and records the rationale for acceptance or rejection. For binary outcomes, the delivered result contains the event count and denominator for each group. For continuous outcomes, it contains the mean, standard deviation, and sample size for each group. When information is insufficient, sources conflict, or a unique candidate cannot be identified, the system preserves partial results and ambiguity rather than combining information across time points or analysis populations.

3.3 Methodological Assessment under Rule Constraints

RoB separates evidence interpretation from the formal execution of predefined risk rules. The system uses the August 22, 2019 version of RoB 2 to assess the effect of assignment to intervention in individually randomized parallel group trials Sterne et al. (2019). D1 is organized at the study level, D2 is bound to the effect of intervention under assessment, and D3 through D5 are bound to the specific result. This binding allows RoB judgments to retain the result identities defined above rather than creating a new assessment object detached from Study $\times$ Target at the risk of bias stage. The model answers signaling questions and selects supporting evidence, while programs execute the applicable decision branches and domain rules.

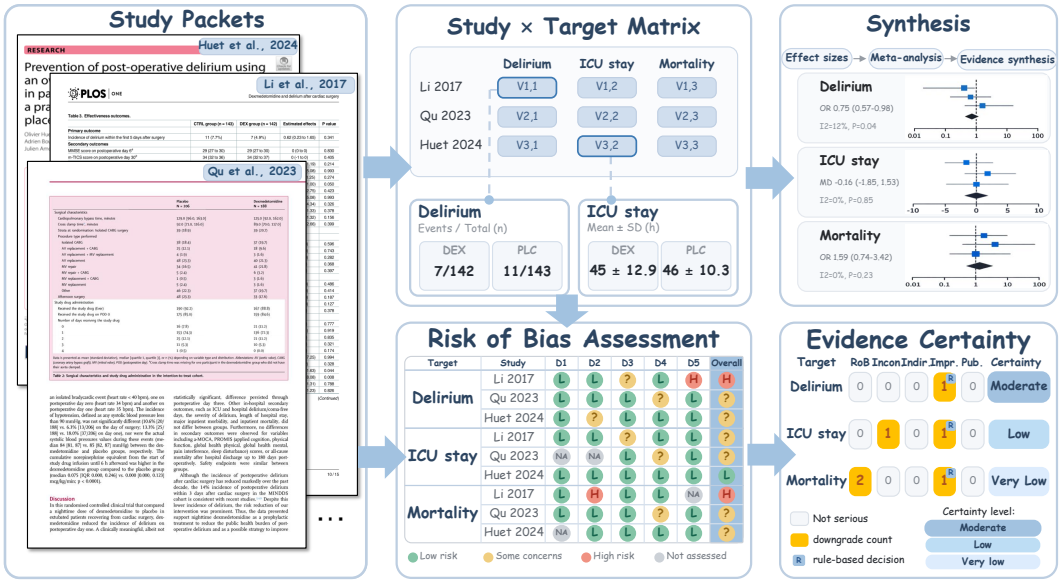


Figure 3: Organization and provenance-aware transformation of evidence synthesis artifacts in MetaRigor. Source materials are aligned into Study × Target results and propagated to statistical synthesis, risk-of-bias assessment, and certainty assessment. The figure shows (left) study-level evidence consolidation, (middle) Study × Target result alignment and structured judgments, and (right) downstream synthesis with preserved evidence relationships.

Overall risk is aggregated according to the programmatic rules implemented in the current system. If any domain is rated High risk, the overall rating is High risk. If all domains are rated Low risk, the overall rating is Low risk. All other cases are rated Some concerns. The program does not automatically resolve the contextual judgment that multiple Some concerns ratings may collectively warrant an overall rating of High risk. Signaling question responses, domain judgments, and the overall judgment are delivered together with their supporting sources.

EC applies GRADE across five domains to a body of evidence defined by a fixed comparison, outcome, and prespecified protocol. Independent model tasks propose downgrading judgments for risk of bias, inconsistency, indirectness, and publication bias. Imprecision is determined programmatically using predefined rules. The current imprecision branch is designed for binary outcomes expressed as RR or OR and applies rules based on the confidence interval, the relevant certainty thresholds, and the optimal information size, or OIS Guyatt et al. (2025).

Overall certainty is generated through semantic synthesis under explicit interval constraints. After judgments have been completed for all five domains, the program maps each domain label to a corresponding downgrading interval $[\lambda_i^-, \lambda_i^+]$ and defines

$$L = \max_{1 \le i \le 5} \lambda_i^-, \qquad U = \min\left(3, \sum_{i=1}^{5} \lambda_i^+\right), \qquad \Delta \in [L, U] \cap \mathbb{Z}. \tag{2}$$

The lower bound preserves the largest minimum downgrading requirement among the five domains. The upper bound aggregates the maximum downgrading allowed across domains and truncates the total at three levels. This interval enables the program to enforce minimum certainty requirements and admissible range constraints in a principled manner without deciding in advance whether judgments from multiple domains reflect the same underlying issue.

The EC synthesis model uses the domain-level rationales to identify potential duplicate downgrading and proposes an integer Δ within the admissible range. The program then verifies that the proposed value falls within the permitted interval. Starting from High certainty, $\Delta = 0, 1, 2, 3$ map to High, Moderate, Low, and Very low, respectively. The interval constraint does not replace semantic identification of duplicate downgrading by the model. The EC synthesis model participates directly in

the formal generation workflow and is distinct from the Reviewer Agent. Both domain judgments and the overall certainty synthesis preserve provenance relationships throughout the workflow to the study materials, RoB outputs, and statistical artifacts.

3.4 Manuscript Generation within Explicit Fact Scopes

MS does not redetermine scientific facts established in upstream stages. Instead, it organizes existing research artifacts within stratified fact scopes and generates the report in accordance with PRISMA 2020 Page et al. (2021). The model first develops a manuscript level plan, then drafts the Introduction, Methods, Results, and Discussion using the Protocol, execution records, and study findings as appropriate, and finally generates the title and abstract.

4 Experiment

4.1 Experimental Setup

We conducted a controlled evaluation across five stages to assess the effects of MetaRigor on study material recovery, complete target result delivery, methodological judgment, manuscript fidelity, and source verifiability. Reference annotations were independently produced by two researchers and reconciled through discussion. Each stage was evaluated using fixed upstream inputs to distinguish differences attributable to the stage under evaluation from errors propagated from earlier stages. Input materials and interannotator agreement are provided in Appendix A.

Under identical research questions, upstream inputs, and admissible materials, the comparative experiments evaluated MetaRigor against Codex CLI and OpenHands CLI. DE, RoB, EC, and MS used deepseek-v4-flash across all systems, whereas EA retained the standard execution configuration of each system. Failed cases were not rerun, outputs were not manually repaired, and no results were selectively retained. Complete configurations and execution constraints are reported in Appendix E.

For the ablation experiments, EA compared three configurations: w/o DEEP, w/o FAST, and Dual DeepSeek. For the remaining stages, we evaluated the w/o Task Contract and w/o Code Logic configurations. The former removed the explicit Task Contract, whereas the latter transferred selected formalizable rules from programmatic execution to model-based judgment. The evaluation separately assessed artifact correctness, manuscript quality, and source verifiability. Metric definitions, denominators, and scoring procedures are provided in Appendix D.

4.2 Comparative Experiments

Table 1: Comparative performance across evidence acquisition, data extraction, risk-of-bias assessment, and evidence certainty assessment under the same evaluation setting. Best and second-best results are indicated in bold and underlined, respectively.

	Data Extraction (DE)				Risk of Bias Assessment (RoB)			Evidence Certainty (EC)		
Method	**Recall↑**	**F_1↑**	**Field Acc.↑**	**Precision↑**	**Domain↑**	**Judgment↑**	**High Risk↑**	**Domain↑**	**Downgrade↑**	**MAE↓**
Codex CLI	<u>84.62%</u>	<u>84.62%</u>	<u>85.99%</u>	<u>84.62%</u>	<u>65.00%</u>	<u>50.00%</u>	<u>7.69%</u>	<u>68.00%</u>	<u>15.00%</u>	<u>1.000</u>
OpenHands CLI	78.46%	79.07%	81.40%	79.69%	62.50%	31.25%	0.00%	61.00%	<u>15.00%</u>	1.053
Ours	**87.18%**	**86.29%**	**88.89%**	**85.43%**	**68.13%**	**59.38%**	**69.23%**	**80.00%**	**50.00%**	**0.579**

	Title/Abstract Stage (EA)				Full-Text Stage (EA)				Overall (EA)	
Method	**Precision↑**	**F_1↑**	**F_2↑**	**NNR↓**	**Precision↑**	**F_1↑**	**F_2↑**	**NNR↓**	**Family Rec.↑**	**Comp. Rec.↑**
Codex CLI	<u>8.50%</u>	<u>15.67%</u>	<u>31.73%</u>	<u>11.76</u>	18.90%	31.58%	52.86%	5.29	32.00%	<u>79.31%</u>
OpenHands CLI	6.94%	12.99%	27.17%	14.40	<u>32.47%</u>	<u>49.02%</u>	<u>70.62%</u>	<u>3.08</u>	<u>36.00%</u>	34.48%
Ours	**52.17%**	**67.61%**	**82.19%**	**1.92**	**63.16%**	**76.19%**	**86.96%**	**1.58**	**88.00%**	**93.10%**

Across all five stages, MetaRigor consistently achieved stronger performance than both CLI baselines (Table 1). In data extraction, MetaRigor obtained an F_1 of 86.29% and field accuracy of 88.89%, improving over Codex CLI by 1.67 and 2.90 percentage points, respectively, while achieving the highest recall (87.18%). The improvements were more pronounced in methodological assessment. MetaRigor reached 68.13% domain agreement and 59.38% overall judgment agreement

for RoB, with high-risk recall of 69.23%, compared with 7.69% for Codex CLI and 0.00% for OpenHands CLI. For EC, MetaRigor achieved 80.00% domain agreement and 50.00% downgrade-count agreement, while reducing downgrade MAE to 0.579.

In evidence acquisition, MetaRigor substantially improved both screening accuracy and study recovery. At the title/abstract stage, it achieved an F1 of 67.61% and an F2 of 82.19%, substantially exceeding Codex CLI (15.67% and 31.73%) and OpenHands CLI (12.99% and 27.17%). At the full-text stage, MetaRigor maintained the highest F1 (76.19%) and F2 (86.96%) while achieving the lowest screening burden among the compared systems (NNR=1.58). These improvements resulted in recovery of 88.00% of study families and 93.10% of companion reports.

For manuscript generation, evaluations by human experts, Luna, and DeepSeek showed a consistent overall advantage for MetaRigor, as shown in Figure 4. MetaRigor achieved a composite score of 89.63, compared with 58.25 for Codex CLI and 73.25 for OpenHands CLI. The dimension-level patterns varied across evaluator types, but the largest improvements generally involved factual fidelity, protocol and execution fidelity, material coverage, and source traceability. Structural organization was not consistently improved; in particular, the human reviewers assigned Full MetaRigor a lower organization score than OpenHands CLI. Because manuscript generation used fixed upstream inputs, these results reflect differences in the generation stage itself rather than accumulated gains from upstream stages under the same controlled setting.

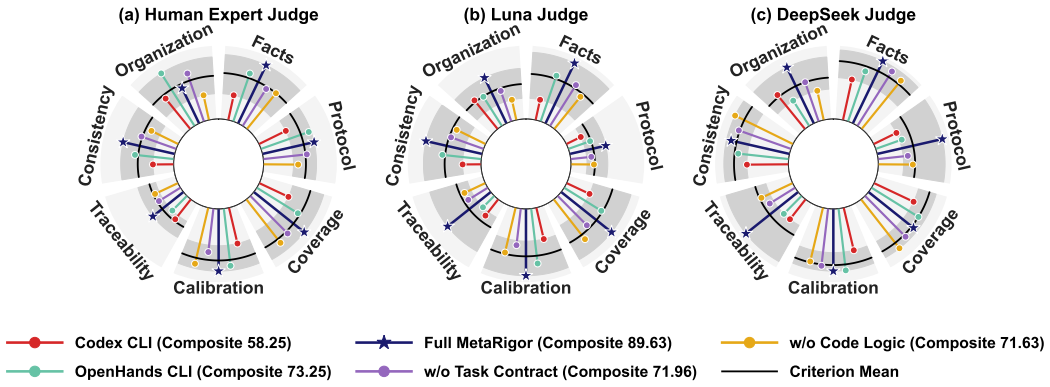


Figure 4: Evaluation of manuscript generation quality under comparative and ablation settings. Human experts, Luna, and DeepSeek assess generated manuscripts across seven criteria, including factuality, protocol adherence, coverage, and traceability.

4.3 Ablation Experiments

Ablation results show that both explicit task constraints and programmatic rules contribute to performance across multiple stages, as shown in Table 2 and Figure 4. Removing the Task Contract reduced DE F1 from 86.29% to 81.61% and field accuracy from 88.89% to 86.84%. It also decreased RoB domain agreement from 68.13% to 62.50% and EC downgrade-count agreement from 50.00% to 25.00%. Removing Code Logic produced a larger degradation in rule-dependent decisions, reducing RoB high-risk recall from 69.23% to 15.38% and increasing EC downgrade MAE from 0.579 to 0.900 across these evaluated rule-dependent settings.

In EA, different screening configurations showed stage-dependent trade-offs. Removing DEEP improved full-text precision (66.67%) and reduced NNR (1.50), whereas removing FAST or DEEP reduced title/abstract screening performance. The full configuration achieved the highest title/abstract F1 (67.61%), full-text F2 (86.96%), and overall study recovery, indicating that heterogeneous screening paths provide complementary benefits. These results show that the contributions of individual components depend on the specific stage and evaluation objective rather than uniformly improving every metric across all evaluated screening settings.

For manuscript generation, both ablations substantially reduced the final manuscript quality scores. Removing the Task Contract reduced the composite score from 89.63 to 71.96, while removing Code Logic resulted in a score of 71.63. The largest decreases were observed in dimensions related

to factual fidelity, protocol execution fidelity, source traceability, and scientific organization, indicating that structured intermediate artifacts and rule-constrained execution contribute to maintaining consistency during report generation across these dimensions.

Table 2: Ablation analysis of key design components in MetaRigor. The table evaluates the contribution of task contracts, code logic, and stage-specific components across evidence acquisition (EA), data extraction (DE), risk-of-bias assessment (RoB), and evidence certainty assessment (EC).

	Data Extraction (DE)				Risk of Bias Assessment (RoB)			Evidence Certainty (EC)		
Method	**Recall↑**	**F_1↑**	**Field Acc.↑**	**Precision↑**	**Domain↑**	**Judgment↑**	**High Risk↑**	**Domain↑**	**Downgrade↑**	**MAE↓**
w/o Task Contract	83.08%	81.61%	86.84%	80.20%	62.50%	56.25%	**69.23%**	67.00%	25.00%	0.824
w/o Code Logic	N/A	N/A	N/A	N/A	56.25%	53.13%	15.38%	67.00%	20.00%	0.900
Ours	**87.18%**	**86.29%**	**88.89%**	**85.43%**	**68.13%**	**59.38%**	**69.23%**	**80.00%**	**50.00%**	**0.579**

	Title/Abstract Stage (EA)				Full-Text Stage (EA)				Overall (EA)	
Method	**Precision↑**	**F_1↑**	**F_2↑**	**NNR↓**	**Precision↑**	**F_1↑**	**F_2↑**	**NNR↓**	**Family Rec.↑**	**Comp. Rec.↑**
w/o DEEP	10.43%	18.82%	36.36%	9.58	42.86%	59.26%	76.92%	2.33	**88.00%**	86.21%
w/o FAST	13.66%	24.04%	44.17%	7.32	**66.67%**	75.86%	82.71%	**1.50**	84.00%	**93.10%**
Dual DeepSeek	12.95%	22.94%	42.66%	7.72	44.00%	58.67%	73.33%	2.27	84.00%	86.21%
Ours	**52.17%**	**67.61%**	**82.19%**	**1.92**	63.16%	**76.19%**	**86.96%**	1.58	**88.00%**	**93.10%**

4.4 End-to-End Integration and Result Identity Propagation

We further evaluated whether result identities and provenance relationships were preserved across continuous execution and downstream stages. In two complete MetaRigor runs, Lin and Turalde, the generated artifacts maintained Study × Target identities, source bindings, and the states of unresolved items, issues, and missing information throughout execution. We then performed a read-only trace analysis of archived outputs from Full MetaRigor, Codex CLI, and OpenHands CLI for the Lin 2025 case. The analysis examined whether local result identity deviations could propagate into downstream statistical inputs, risk-of-bias assessment targets, and final reporting artifacts.

A representative discrepancy involved the SELECT2 mortality target. The predefined target required death from any cause within 90 days. While OpenHands CLI associated a source-supported mortality result with this target, the extracted outcome and follow-up definition corresponded to in-hospital mortality rather than the target 90-day mortality outcome. Full MetaRigor selected the result matching the predefined target identity. This example demonstrates that evidence synthesis errors may arise even when extracted values are supported by the source, if the scientific identity of the result is not preserved. Detailed artifact comparisons and offline analyses are provided in Appendix G. Overall, the tracing results show that maintaining explicit result identities and provenance relationships is important for preventing local mismatches from propagating across evidence synthesis stages.

5 Conclusion

In this work, we presented MetaRigor, a framework for verifiable end-to-end evidence synthesis built around explicit Scientific Decision Boundaries. MetaRigor maintains Study × Target identities across stages, uses Task Contracts to separate model interpretation from programmatic rule execution, and binds outputs to source evidence. Controlled evaluations across five stages show improvements over two general-purpose CLI baselines in evidence acquisition, complete result delivery, methodological assessment, and manuscript fidelity. Ablations identify distinct, stage-dependent contributions from task constraints and programmatic rules. Tracing archived outputs reveals why cross-stage consistency matters: correctly extracted values can enter final reports under the wrong outcome or time point, compromising downstream synthesis. This finding highlights result identity as an essential part of correctness in evidence synthesis. While broader validation remains necessary, our results motivate scientific agents that explicitly preserve the relationships among evidence, decisions, and results throughout execution.

AI USE STATEMENT

We used generative AI tools solely for manuscript refinement, including language translation, grammar correction, and improving clarity and readability. AI tools were not used for research ideation, hypothesis generation, experimental design, data analysis, result generation, dataset creation, mathematical derivations, proofs, or scientific claims. All AI-assisted suggestions were reviewed and verified by the authors, who take full responsibility for the final content.

ETHICS STATEMENT

MetaRigor is designed to support, rather than replace, human judgment in evidence synthesis and is not intended to provide clinical recommendations. Automated evidence acquisition, data extraction, methodological assessment, and manuscript generation may introduce errors or unsupported interpretations; therefore, system outputs should be independently verified by researchers. The experiments use existing scholarly materials and research artifacts and do not involve the recruitment of human participants or the collection of new personally identifiable information. Responsibility for study inclusion, methodological judgments, evidence interpretation, and downstream scientific or clinical claims remains with human researchers.

REPRODUCIBILITY STATEMENT

Source code and evaluation artifacts are available at `https://anonymous.4open.science/status/Meta-Rigor-8C62`.

REFERENCES

Naman Ahuja, Saniya Mulla, Muhammad Ali Khan, Zaryab Bin Riaz, Kaneez Zahra Rubab Khakwani, Mohamad Bassam Sonbol, Irbaz Bin Riaz, and Vivek Gupta. Evisearch: A human in the loop system for extracting and auditing clinical evidence for systematic reviews. *arXiv preprint arXiv:2604.14165*, 2026.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6465–6488, 2023.

Gordon Guyatt, Linan Zeng, Romina Brignardello-Petersen, Manya Prasad, Hans De Beer, M Hassan Murad, Alfonso Iorio, Arnav Agarwal, Liang Yao, Thomas Agoritsas, et al. Core grade 2: choosing the target of certainty rating and assessing imprecision. *Bmj*, 389, 2025.

Anthony Hevia, Sanjana Chintalapati, Veronica Ka Wai Lai, Nguyen Thanh Tam, Wai-Tat Wong, Terry P Klassen, and Lucy Lu Wang. Roboto2: An interactive system and dataset for llm-assisted clinical trial risk of bias assessment. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 12–25, 2025.

Julian P. T. Higgins, James Thomas, Jacqueline Chandler, Miranda Cumpston, Tianjing Li, Matthew J. Page, and Vivian A. Welch. Cochrane handbook for systematic reviews of interventions, version 6.5, 2024. URL `https://www.cochrane.org/handbook`. Updated August 2024.

Tianjing Li, Julian P. T. Higgins, and Jonathan J. Deeks. Chapter 5: Collecting data. In Julian P. T. Higgins, James Thomas, Jacqueline Chandler, Miranda Cumpston, Tianjing Li, Matthew J. Page, and Vivian A. Welch (eds.), *Cochrane Handbook for Systematic Reviews of Interventions*. Cochrane, version 6.5 edition, 2024. URL `https://training.cochrane.org/handbook/current/chapter-05`. Chapter last updated October 2019.

Iain J Marshall, Benjamin Nye, Joël Kuiper, Anna Noel-Storr, Rachel Marshall, Rory Maclean, Frank Soboczenski, Ani Nenkova, James Thomas, and Byron C Wallace. Trialstreamer: A living, automatically updated database of clinical trial reports. *Journal of the American Medical Informatics Association*, 27(12):1903–1912, 2020.

Joanne E. McKenzie, Sue E. Brennan, Rebecca E. Ryan, Hilary J. Thomson, Renea V. Johnston, and James Thomas. Chapter 3: Defining the criteria for including studies and how they will be grouped for the synthesis. In Julian P. T. Higgins, James Thomas, Jacqueline Chandler, Miranda Cumpston, Tianjing Li, Matthew J. Page, and Vivian A. Welch (eds.), *Cochrane Handbook for Systematic Reviews of Interventions*. Cochrane, version 6.5 edition, 2024. URL `https://training.cochrane.org/handbook/current/chapter-03`. Chapter last updated August 2023.

Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, et al. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *bmj*, 372, 2021.

Amrou Sarraj, Ameer E Hassan, Michael G Abraham, Santiago Ortega-Gutierrez, Scott E Kasner, M Shazam Hussain, Michael Chen, Spiros Blackburn, Clark W Sitton, Leonid Churilov, et al. Trial of endovascular thrombectomy for large ischemic strokes. *New England Journal of Medicine*, 388(14):1259–1271, 2023.

Jonathan AC Sterne, Jelena Savović, Matthew J Page, Roy G Elbers, Natalie S Blencowe, Isabelle Boutron, Christopher J Cates, Hung-Yuan Cheng, Mark S Corbett, Sandra M Eldridge, et al. Rob 2: a revised tool for assessing risk of bias in randomised trials. *bmj*, 366, 2019.

Moein Taherinezhad, Sebastian Maier, Gerardo Vitagliano, Francesco Pierri, and Stefan Feuerriegel. Autosynthesis: An agentic system for automated meta-analysis. *arXiv preprint arXiv:2607.15247*, 2026.

Rens Van De Schoot, Jonathan De Bruin, Raoul Schram, Parisa Zahedi, Jan De Boer, Felix Weijdema, Bianca Kramer, Martijn Huijts, Maarten Hoogerwerf, Gerbrich Ferdinands, et al. An open source machine learning framework for efficient and transparent systematic reviews. *Nature machine intelligence*, 3(2):125–133, 2021.

Xingyao Wang, Boxuan Li, Yufan Song, Frank F Xu, Xiangru Tang, Mingchen Zhuge, Jiayi Pan, Yueqi Song, Bowen Li, Jaskirat Singh, et al. Openhands: An open platform for ai software developers as generalist agents. In *International Conference on Learning Representations*, volume 2025, pp. 65882–65919, 2025a.

Zifeng Wang, Lang Cao, Benjamin Danek, Qiao Jin, Zhiyong Lu, and Jimeng Sun. Accelerating clinical evidence synthesis with large language models. *NPJ Digital Medicine*, 8(1):509, 2025b.

Zifeng Wang, Lang Cao, Qiao Jin, Joey Chan, Nicholas Wan, Behdad Afzali, Hyun-Jin Cho, Chang-In Choi, Mehdi Emamverdi, Manjot K Gill, et al. A foundation model for human-ai collaboration in medical literature mining. *Nature communications*, 16(1):8361, 2025c.

Wanghan Xu, Wenlong Zhang, Fenghua Ling, Ben Fei, Yusong Hu, Runmin Ma, Bo Zhang, Fangxuan Ren, Jintai Lin, Wanli Ouyang, et al. Manalyzer: End-to-end automated meta-analysis with multi-agent system. *arXiv preprint arXiv:2505.20310*, 2025.

John Yang, Carlos Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik Narasimhan, and Ofir Press. Swe-agent: Agent-computer interfaces enable automated software engineering. *Advances in Neural Information Processing Systems*, 37:50528–50652, 2024.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

A Reference Sets and Stage Inputs

A.1 Construction of Reference Answers

Reference answers were independently annotated by two researchers, with disagreements resolved through discussion. EA counts followed the report- and study-grouping conventions of the original reviews and are not listed separately in the agreement table. DE was evaluated at the Study $\times$ Target unit and numerical-field levels; RoB annotations covered signaling questions, domain

judgments, and overall judgments; and EC annotations covered judgments across five domains and overall judgments. Table 3 presents the annotation counts and pre-adjudication agreement at each evaluation level under the same evaluation protocol.

Table 3: Pre-adjudication agreement of reference annotations.

Stage	Evaluation unit	Annotation units	Initial matches	Initial agreement	Discussed
DE	Study × Target	405	393	97.0%	12
	Numerical fields	828	816	98.6%	12
RoB	Signaling questions / domains / overall	896	858	95.8%	38
EC	Five domains / overall	120	115	95.8%	5

A.2 Stage Inputs and Evaluation References

The stage−controlled evaluations in Sections 4.2 and 4.3 of the main text use fixed upstream inputs to distinguish differences within the current stage from upstream error propagation. Table 4 lists the task-visible materials and evaluation references for each stage under the same evaluation protocol. Inputs do not include the final answers to be predicted at the current stage; upstream results required by downstream tasks are provided only as permitted facts. Reviews reused across stages are not counted again as new, independent review samples.

Table 4: Inputs and references for stage-controlled evaluation.

Stage	Task-visible materials	Evaluation reference
EA	Protocol, bibliographic records or full texts, and report relationships	Reference study set and known report families
DE	Protocol, frozen source representations, and Study × Target list	Study × Target numerical results and statuses
RoB	Study materials, result identity, and signaling questions	Signaling-question, domain, and overall labels
EC	Comparisons, outcomes, RoB results, and statistical artifacts	Five-domain judgments and overall downgrade counts
MS	Protocol, execution records, and research findings	Seven-dimensional scoring rubric and permitted facts

A.3 Task Scale and Scope of Supplementary Analyses

Table 5 summarizes the scale of the five-stage controlled evaluation across all stage-specific tasks. Case counts are reported by stage-specific task; cases reused across stages are not summed and interpreted as the total number of independent reviews.

Table 5: Task scale of the five-stage controlled evaluation.

Stage	Cases	Evaluation scale
EA	5	822 title–abstract records; 300 full texts; 25 reference studies and 29 companion reports
DE	10	405 Study × Target units; 195 reported results; 828 numerical fields
RoB	7	32 study–result units; 160 domain judgments; 704 signaling questions
EC	20	20 fixed analyses; 100 domain judgments; 20 overall downgrade judgments
MS	10	10 manuscripts per condition; 7 scoring criteria per manuscript

Appendix F provides tool-call traces from the EC ablation run, and Appendix G documents continuous-execution states, archived artifacts from the three systems, and downstream artifact displays. Case identifiers and audit details are provided only in the corresponding appendices. These

supplementary analyses are neither pooled with the stage-controlled evaluations to calculate accuracy nor counted as additional independent review samples.

B Result Alignment and the Full Evaluation Space

B.1 Study × Target Alignment

Before DE evaluation, two researchers independently aligned the results delivered by each system to the predefined Study × Target grid. Alignment was based on study identity, comparison roles, outcome definitions, follow-up time points, populations, analysis conditions, and result types. The researchers aligned only the content already delivered by each system; they did not supplement missing system outputs with numbers from the original sources. Partial results, missing results, and unresolved ambiguities retained their original statuses and were scored uniformly by the program.

B.2 DE Evaluation Space

The DE evaluation covers 405 Study × Target units. The reference set contains 195 reported results: 171 binary results and 24 continuous results. Binary results contain event counts and denominators for both groups; continuous results contain means, standard deviations, and sample sizes for both groups. These yield a total of 828 numerical fields: $171 \times 4 + 24 \times 6 = 828$. A result is considered completely correct only when all four fields of a binary result or all six fields of a continuous result match the reference. Reported results that are not delivered in full remain in the fixed reference denominator. Complete candidates for ordinary non-positive targets and reference-conflict targets enter the candidate set but do not contribute to the correct-result numerator. Complete candidates must also satisfy the claim-closure requirements in Appendix D.2.

B.3 Evaluation Conditions

DE results are reported for Full MetaRigor, Codex CLI, OpenHands CLI, and w/o Task Contract. The w/o Code Logic ablation does not apply to DE, and all corresponding entries are marked N/A (not applicable). EA configuration comparisons and mechanism ablations for the other stages are described in Appendix E under the same evaluation framework.

C Representative Task Contract and Baseline Prompt

C.1 Example DE Task Contract

Task object. Determine whether an observation matches the specified Study × Target and whether its reported fields can be adopted for that target.

Permitted facts. The study, comparison, outcome, population, time point, and analysis conditions specified in the Protocol; original reports and Source Spans in the current Study Packet; fields already parsed from the current observation; and the predefined schema and numerical constraints.

Prohibited facts. Results from other Studies or Targets; results from other time points, populations, or comparison groups; model memory or inferences without supplied source support; and numbers absent from the current materials that are inferred backward from review conclusions.

Delivery requirements. Return match, mismatch, or unresolved, explain each field judgment, and provide a Source Anchor for every adopted field. Retain unresolved for insufficient evidence; do not fill in missing numerical values or modify deterministic fields generated by the program.

C.2 Example CLI Baseline Prompt

Complete task using the supplied research question, Protocol, and study materials. Use only facts explicitly provided in the input files. Do not add studies, numbers, time points, comparison groups, or analysis populations. When evidence is insufficient, return unresolved and explain why. Follow the specified output format strictly, and do not combine results from other studies or time points.

This example illustrates only the core constraints of the baseline task; it does not replace the complete stage-specific prompts, schemas, and model configurations.

D EC METRICS, INTERVAL CONSTRAINTS, AND VALID-OUTPUT SCOPE

This appendix consolidates the metric names, evaluation levels, and supplementary results for Tables 1 and 2 and the main-text MS scoring figure. Percentage metrics are rounded to two decimal places by default; reference-annotation agreement retains the one-decimal precision used in Table 3; and O-MAE and descriptive risk ratios are rounded to three decimal places. N/A indicates that an ablation is not applicable, rather than an unreported result, a failed run, or a score of zero.

D.1 EA METRICS

Precision, F_1, F_2, and NNR are reported separately for title–abstract (TA) and full-text (FT) screening. Each stage uses micro-aggregation: the numbers of correctly included records (TP), system-predicted inclusions (C), and reference inclusions (N) are first summed across cases, and metrics are then calculated from these aggregate counts. Case-level scores are not calculated first and then averaged. The metrics are defined as follows

$$\text{Precision} = \frac{TP}{C}, \qquad F_1 = \frac{2\,TP}{N + C}, \qquad F_2 = \frac{5\,TP}{4N + C}, \qquad NNR = \frac{C}{TP}. \tag{3}$$

NNR is the number of system-included records per correctly included record and equals the reciprocal of precision. It expresses the screening burden in practical review settings and is not treated as performance evidence independent of precision.

When the denominator of any formula is zero, the evaluation script records the metric as null rather than zero. This treatment differs from N/A for inapplicable conditions and does not alter the micro-aggregation convention based on aggregate counts.

Study-material recovery is reported as the complete study-family recovery rate and the companion-report recovery rate, with reference denominators of 25 study families and 29 companion reports, respectively. The column labels Family Rec. and Comp. Rec. in the main-text tables correspond to these two rates. For the full configuration, 88.00% corresponds to 22/25 study families, and 93.10% corresponds to 27/29 companion reports. Screening and recovery results are presented in Tables 1 and 2 of the main text for direct cross-system comparison.

D.2 DE METRICS AND SUPPLEMENTARY RESULTS

DE uses the claim closure only evaluation convention. Claim-closure filtering is performed at the Item level: an Item is retained only if every delivered numerical field has a valid source attachment. If any delivered numerical field lacks a valid source attachment, the Item is excluded from the complete-candidate set, and none of its fields contributes to the correct-field count. Missing fields are not filled to obtain source attachments; partial results passing the filter may still be retained.

Within the filtered set, let $N = 195$ be the number of reference reported results, E the number of results that satisfy claim closure and are numerically complete and correct, and C the number of complete candidates that satisfy claim closure and can be uniquely encoded. Then:

$$\text{Recall} = \frac{E}{N}, \qquad \text{Precision} = \frac{E}{C}, \qquad F_1 = \frac{2E}{N + C}. \tag{4}$$

A complete result requires all four binary fields or all six continuous fields to be correct. Field accuracy uses the fixed set of 828 reference numerical fields as its denominator. Let M be the number of fields that match the reference values after Item-level closure filtering. Then:

$$\text{Field Acc.} = \frac{M}{828}. \tag{5}$$

The corresponding DE performance across systems and configurations is summarized in Table 6.

Table 6: DE performance at the complete result and field levels (%)

System / configuration	Complete result recall	Precision	F_1	Field accuracy
Codex CLI	84.62	84.62	84.62	85.99
OpenHands CLI	78.46	79.69	79.07	81.40
Full MetaRigor	87.18	85.43	86.29	88.89
w/o Task Contract	83.08	80.20	81.61	86.84
w/o Code Logic	N/A	N/A	N/A	N/A

D.3 RoB Metrics

RoB reports domain agreement, overall-judgment agreement, and high risk recall, corresponding to Domain, Judgment, and High Risk in the main-text tables. The denominators are 160 domain judgments for domain agreement, 32 study–result overall judgments for overall-judgment agreement, and 13 objects labeled as overall high risk in the reference for high risk recall. Missing or invalid judgments are not counted as correct matches and are not removed from the fixed denominators.

The current program uses a simplified overall aggregation rule: if any domain is High risk, the overall judgment is High risk; if all domains are Low risk, the overall judgment is Low risk; otherwise, it is Some concerns. Context-dependent judgments in which multiple Some concerns domains jointly warrant an overall High risk judgment are not yet included in the automated aggregation. Overall errors may therefore arise both from domain judgments and from the coverage of the aggregation rules. Results for each configuration are presented in Tables 1 and 2 of the main text.

D.4 EC Metrics and Valid-Output Scope

EC reports domain agreement, overall downgrade-count agreement, and mean absolute error in the overall downgrade count, corresponding to Domain, Downgrade, and MAE in the main-text tables. Domain agreement uses 100 domain judgments as the denominator under exact matching, with boundary labels treated as distinct valid categories for exact matching. Overall downgrade-count agreement uses all 20 analyses as the denominator and compares only explicitly reported overall downgrade counts. Missing or invalid outputs count as non-matches and are not imputed from the final certainty level or the sum of domain downgrades.

The interval constraint used in overall certainty synthesis is based on the mapping from domain labels to permitted downgrade intervals shown in Table 7.

Table 7: Mapping from EC domain labels to permitted downgrade intervals.

Domain label	Permitted downgrade interval
NOT_SERIOUS	$[0, 0]$
BORDERLINE_NOT_SERIOUS_SERIOUS	$[0, 1]$
SERIOUS	$[1, 1]$
BORDERLINE_SERIOUS_VERY_SERIOUS	$[1, 2]$
VERY_SERIOUS	$[2, 2]$

The same mapping is applied to all five EC domains. The overall lower bound is the maximum of the five domain-level lower bounds, whereas the overall upper bound is the sum of the five domain-level upper bounds, capped at three. Borderline labels preserve uncertainty between two adjacent downgrade levels; they do not determine the final overall downgrade by themselves.

The interval constraint limits the admissible numerical range of the final downgrade count, but it does not determine whether domain-level limitations are independent, overlapping, or caused by the same underlying issue. Resolving potential duplicate downgrading remains a semantic synthesis decision performed by the EC synthesis model. We therefore do not report a separate synthesis-only accuracy estimate for this step within the current evaluation framework.

The mean absolute error in the overall downgrade count (O-MAE, corresponding to MAE in the main text) is calculated only over the paired set V in which both the system and the reference provide valid overall downgrade counts:

$$O\text{-MAE} = \frac{1}{|V|} \sum_{i \in V} \left| \hat{\Delta}_i - \Delta_i \right|. \tag{6}$$

where Δ_i and $\hat{\Delta}_i$ denote the reference and system overall downgrade counts, respectively. The number of valid pairs is reported alongside this metric, which does not replace overall downgrade-count agreement with its fixed denominator of 20 analyses. Because the number of valid pairs differs across configurations, O-MAE should be interpreted together with output completeness.

The corresponding EC judgment performance and valid-pair counts are summarized in Table 8.

Table 8: EC judgment performance and valid-pair counts for O-MAE

System / configuration	Domain agreement	Overall downgrade-count agreement	O-MAE	Valid pairs
Codex CLI	68.00	15.00	1.000	20/20
OpenHands CLI	61.00	15.00	1.053	19/20
Full MetaRigor	80.00	50.00	0.579	19/20
w/o Task Contract	67.00	25.00	0.824	17/20
w/o Code Logic	67.00	20.00	0.900	20/20

D.5 MS Scoring and Complete Dimension-Level Results

D.5.1 Scoring Criteria and Review Procedure

MS uses seven scoring criteria, each rated on a 0–4 scale, with weights listed in Table 9. The scoring anchors are: 4, fully satisfies the criterion with no substantive issues; 3, largely satisfies the criterion with localized issues; 2, partially satisfies the criterion but contains substantive omissions requiring manual correction; 1, fails to satisfy most requirements and may mislead readers; and 0, a critical failure, fabricated facts, or a major deviation from the Protocol.

Table 9: MS scoring dimensions, figure labels, and weights

Dimension	Figure label	Weight
Factual fidelity	Facts	25%
Protocol and execution fidelity	Protocol	15%
Material coverage	Coverage	15%
Interpretive calibration	Calibration	15%
Source traceability	Traceability	10%
Cross-section consistency	Consistency	10%
Scientific organization	Organization	10%

DeepSeek, Luna, and the human reviewers assessed manuscripts from the same ten cases, with mutually independent scoring contexts. Both human reviewers had research experience in biomedical artificial intelligence, medical image analysis, and meta-analysis. They scored independently using the same materials, weights, and scoring anchors, without viewing each other's scores, and did not participate in system execution, manuscript generation, or evaluation of the other stages. Human review used anonymous configuration labels and randomized configuration order; system, model, and experimental-condition identities were not disclosed to the reviewers.

For each manuscript and dimension, the initial score was the arithmetic mean of the two reviewers' scores, rounded to the prespecified precision. When the difference between their scores exceeded 0.4 points on any dimension, they discussed the scoring rationale for that dimension only and adopted a mutually agreed final score. Discussion did not alter the manuscript or scores on other dimensions. The number of reviewers, criteria, weights, disagreement threshold, and adjudication rules were fixed before the formal evaluation to ensure procedural consistency.

D.5.2 Evaluator-Specific and Composite Scoring

Let r_{srjk} denote the final score for manuscript j from system s, assigned by evaluator category r on dimension k, and let w_k denote the dimension weight. The manuscript score on a 100-point scale and the system-level score for that evaluator category are, respectively:

$$\mathrm{Score}_{s,r,j} = 100 \sum_{k=1}^{7} w_k \frac{r_{s,r,j,k}}{4}, \quad \overline{\mathrm{Score}}_{s,r} = \frac{1}{10} \sum_{j=1}^{10} \mathrm{Score}_{s,r,j}. \tag{7}$$

Weights are expressed as proportions; for example, 25% is entered as 0.25. The evaluator categories r are DeepSeek, Luna, and human review. The human category uses the final scores aggregated from the two reviewers as described above; the two human reviewers are not treated as two independent evaluator categories. Tables 10, 11, 12, and 13 provide the complete dimension-level results. Each dimension is scored from 0 to 4, and total scores are expressed on a 100-point scale.

The MS scoring figure in the main text reports a cross-evaluator composite score as Overall, calculated as the equally weighted mean of the human, Luna, and DeepSeek evaluation results:

$$\mathrm{Overall}_s = \frac{\overline{\mathrm{Score}}_{s,\mathrm{Human}} + \overline{\mathrm{Score}}_{s,\mathrm{Luna}} + \overline{\mathrm{Score}}_{s,\mathrm{DeepSeek}}}{3}. \tag{8}$$

System-level scores for each evaluator category and Overall are aggregated from unrounded scores and rounded to two decimal places only for final presentation. Overall is 89.63 for Full MetaRigor and 71.63 for w/o Code Logic. The composite score does not replace evaluator-specific results; the three category-specific totals and the composite score are reported side by side in Table 13.

Table 10: Manuscript evaluation by DeepSeek.

System /configuration	Factual fidelity	Protocol execution fidelity	Material coverage	Interpretive calibration	Source traceability	Cross-section consistency	Scientific organization	Total
Codex CLI	2.80	2.10	2.90	2.90	2.10	2.80	2.90	66.63
OpenHands CLI	3.30	2.20	3.40	3.70	2.10	3.20	2.30	74.50
Full MetaRigor	4.00	3.90	3.50	3.70	3.90	3.60	3.70	94.63
w/o Task Contract	3.80	2.30	3.50	3.50	2.30	3.40	2.80	79.88
w/o Code Logic	3.70	2.50	3.70	3.40	2.50	3.80	2.30	80.63

Table 11: Manuscript evaluation by Luna.

System /configuration	Factual fidelity	Protocol execution fidelity	Material coverage	Interpretive calibration	Source traceability	Cross-section consistency	Scientific organization	Total
Codex CLI	1.90	1.70	2.10	2.40	1.90	1.80	2.60	50.88
OpenHands CLI	3.10	2.00	2.90	3.40	1.70	2.70	2.50	67.75
Full MetaRigor	3.90	2.60	3.80	3.90	3.40	3.50	3.20	88.25
w/o Task Contract	3.10	1.90	2.80	2.60	2.00	2.50	2.40	64.00
w/o Code Logic	2.80	2.00	3.20	3.00	2.00	2.40	1.90	64.00

Table 12: Manuscript evaluation by human reviewers.

System /configuration	Factual fidelity	Protocol execution fidelity	Material coverage	Interpretive calibration	Source traceability	Cross-section consistency	Scientific organization	Total
Codex CLI	2.10	2.30	2.40	2.60	2.10	1.90	2.70	57.25
OpenHands CLI	3.20	3.20	3.10	3.50	1.90	2.70	3.70	77.50
Full MetaRigor	3.80	3.30	3.80	3.70	2.70	3.30	2.70	86.00
w/o Task Contract	2.90	2.90	3.30	2.90	2.10	2.60	3.20	72.00
w/o Code Logic	3.00	2.50	3.40	3.50	2.10	2.30	2.10	70.25

D.6 Claim Closure and Reference Recall

Source verifiability is evaluated using two complementary metrics: Claim Closure and Reference Recall, which capture technical binding validity and evidence recovery, respectively. Claim Closure measures whether a delivered output is attached to a permitted and retrievable source location. Reference Recall measures whether the system recovers manually identified evidence spans that support the corresponding output. These metrics use different evaluation units and denominators and are

Table 13: Evaluator-specific totals and cross-evaluator composite scores.

System / configuration	DeepSeek	Luna	Human review	Overall
Codex CLI	66.63	50.88	57.25	58.25
OpenHands CLI	74.50	67.75	77.50	73.25
Full MetaRigor	94.63	88.25	86.00	89.63
w/o Task Contract	79.88	64.00	72.00	71.96
w/o Code Logic	80.63	64.00	70.25	71.63

therefore reported separately. Neither metric establishes that the retrieved text semantically supports the corresponding judgment, nor do they replace task-level correctness metrics. The corresponding source-verifiability results are summarized in Table 14.

Table 14: Source-verifiability metrics under internal ablations (%).

Configuration	DE		RoB		EC	
	Claim Closure ↑	Reference Recall ↑	Claim Closure ↑	Reference Recall ↑	Claim Closure ↑	Reference Recall ↑
Ours	95.70	88.21	74.67	50.85	100.00	82.98
w/o Task Contract	95.38	86.67	76.68	49.15	68.07	69.50
w/o Code Logic	N/A	N/A	76.84	52.78	97.63	90.78

D.6.1 Claim Closure

Let n denote the number of delivered units submitted to source verification and v the number with valid source attachments. Claim Closure is defined as:

$$\text{Claim Closure} = \frac{v}{n}. \tag{9}$$

For DE, the verification unit is a delivered numerical field in the unfiltered system output. Full MetaRigor achieves 891/931 = 95.70%, while w/o Task Contract achieves 908/952 = 95.38%. Undelivered fields are not included in this denominator. This evaluation precedes the Item-level closure filtering used for the task metrics in Appendix D.2, where an Item is retained only when all delivered numerical fields have valid source attachments. Consequently, the retained DE evaluation set has 100% closure by construction in this analysis.

RoB applies the same source-attachment validity criterion as DE to unfiltered delivered RoB evidence. EC Claim Closure likewise measures the technical binding of delivered evidence to valid source attachments before task-level judgment scoring.

D.6.2 Reference Recall

Let H denote the number of matched reference evidence units and N_ref the number of predefined reference units. Reference Recall is defined as:

$$\text{Reference Recall} = \frac{H}{N_{\text{ref}}}. \tag{10}$$

For DE, the denominator is 195 reference reported-result Items. An Item is counted once if an accepted reference span is recovered; repeated matches do not increase the numerator. Full MetaRigor recovers 172/195 = 88.21%, and w/o Task Contract recovers 169/195 = 86.67%. This metric assesses recovery of reference evidence and differs from complete result recall in Appendix D.2.

For RoB, Reference Recall is evaluated against 413 reference signaling questions (SQs), with repeated citations to the same reference unit counted only once. This reference denominator is distinct from the 704 task signaling questions reported in Appendix A. For EC, the reference set contains 141 evidence links after deduplication by case, domain, and evidence ID; each deduplicated reference unit contributes at most once to the numerator.

Claim Closure, Reference Recall, and task correctness need not change in the same direction. For example, w/o Code Logic obtains higher Reference Recall than the full configuration in both RoB and EC, while its task-level judgment performance is lower. Source-verifiability metrics are therefore reported as complementary diagnostics rather than substitutes for task accuracy.

E Execution Configurations, Baselines, and Ablation Conditions

E.1 Stage-Controlled Comparisons

Sections 4.2–4.3 of the main text present system-level comparisons using the same research questions, permitted materials, fixed upstream inputs, and task requirements. DE, RoB, EC, and MS all use deepseek-v4-flash; EA retains each system's formal execution configuration. Raw outputs, tool returns, and failure states were retained, without rerunning failed items, applying manual repairs, or selecting the best result for any evaluated system.

Codex CLI and OpenHands CLI retain their native execution interfaces and tool-orchestration mechanisms and serve as system-level agent baselines, rather than isolated model or prompt baselines. The comparison concerns workflow organization, including result identity, task contracts, programmatic rules, source binding, and delivery requirements. Because EA uses each system's formal configuration, its comparison includes the associated configuration differences; between-system differences are not attributed exclusively to any one component. Cross-stage artifact analyses are interpreted separately within the scope specified in Appendix G.

In the full EA configuration, FAST uses deepseek-v4-flash, DEEP uses gpt-5.6-luna, and full-text disagreement adjudication uses gpt-5.6-terra. The two model-evaluator categories for MS are DeepSeek and Luna. The recorded software versions are Codex 0.142.4 and the OpenHands 1.16.0 series. Specific roles are listed in Table 15.

Table 15: Model roles and CLI versions

Task / role	Model or software identifier
Generation model for DE, RoB, EC, and MS	deepseek-v4-flash
EA FAST	deepseek-v4-flash
EA DEEP	gpt-5.6-luna
EA full-text disagreement adjudication	gpt-5.6-terra
MS evaluator (DeepSeek)	deepseek-v4-flash
MS evaluator (Luna)	gpt-5.6-luna
Codex CLI	0.142.4
OpenHands CLI	1.16.0 series

E.2 Configuration Comparisons and Applicable Stages

The full EA configuration uses FAST, DEEP, and disagreement-adjudication routing and is compared with w/o DEEP, w/o FAST, and Dual DeepSeek. The w/o DEEP configuration retains only the FAST screening path, whereas w/o FAST retains only the DEEP screening path. Dual DeepSeek retains both paths, with DeepSeek covering all model-based evaluation roles in this EA comparison, including full-text disagreement adjudication.

The w/o Task Contract ablation applies to DE, RoB, EC, and MS as appropriate. The w/o Code Logic ablation applies to RoB, EC, and MS, but not DE. The former removes explicit task contracts; the latter transfers formalizable rules from program execution to model judgment and does not mean all program code is removed. All other inputs and formal execution procedures remain unchanged.

The applicable stages and configuration changes are summarized in Table 16.

F Execution Traces and Error Cases

F.1 Overwriting of Tool Outputs During Subsequent Generation

This section analyzes all 8 calculator calls across 20 EC cases under w/o Task Contract. Tool calls were selected autonomously by the model at runtime. These 8 calls constitute all calculator calls in that run, rather than a post hoc sample, and originated from 8 distinct cases. For each call, the calculator input, calculator output, and corresponding final output were retained for analysis.

Table 16: Comparison conditions and scope of applicability

Condition	Applicable stages	Core configuration or change
Full MetaRigor	EA, DE, RoB, EC, MS	Full workflow: result identity, Task Contract, programmatic rules, and source binding
Codex CLI	EA, DE, RoB, EC, MS	Native CLI agent execution
OpenHands CLI	EA, DE, RoB, EC, MS	Native agent workflow
w/o DEEP	EA	FAST screening path only
w/o FAST	EA	DEEP screening path only
Dual DeepSeek	EA	DeepSeek covers all evaluation roles in this EA comparison, including full-text adjudication
w/o Task Contract	DE, RoB, EC, MS	Explicit Task Contract removed
w/o Code Logic	RoB, EC, MS	Selected formalized rules delegated to model judgment; not applicable to DE

The input arguments for all 8 calls were consistent with the frozen upstream inputs. All labels returned by calculator matched the reference and were included in subsequent model inputs, but 4 were overwritten during final generation with labels that did not match the reference. For example, in the short-term stroke rehospitalization case, SERIOUS was changed to NOT_SERIOUS solely because the confidence interval did not cross the null value, while the information-size criterion was omitted. The proportion 4/8 describes overwriting among the tool calls that occurred and is not an overall error rate for all 20 EC analyses.

F.2 Diagnostic Recalculation

In a post hoc recalculation that held all other fields fixed, restoring only these 4 tool-returned labels increased domain matches from 67/100 to 71/100. This diagnostic recalculation examines the effect of overwriting tool results; it does not replace the formal experimental results and differs from the direct observation of actual archived artifacts in Appendix G.

G End-to-End Artifacts, Result Identity Propagation, and Downstream Displays

G.1 Scope of Analysis

The end-to-end audit comprises two parts: state preservation during continuous execution and result propagation across three systems. The specific cases are reported together in this appendix. The first part examines the sequential artifacts produced by Full MetaRigor in the Lin and Turalde cases, checking preservation of Study $\times$ Target, source bindings, UNRESOLVED, Issue, and missing statuses. These states were preserved in both cases; this check is not used to estimate overall scientific accuracy. The three-system analysis below retains the case identifier Lin-2025.

The three-system analysis uses the actual archived artifacts from Full MetaRigor, Codex CLI, and OpenHands CLI for Lin-2025. DE results are aligned to the same 42 Study × Target units, and the corresponding RoB artifacts and facts in the final reports are traced. The analysis is read-only: models are not called again, failed items are not rerun, results absent from system outputs are not supplemented from the original sources, and historical artifacts are not modified.

G.2 DE Differences and Identity Mismatches

Table 17 compares the delivery statuses and four binary fields—event counts and denominators for both groups—across the three systems. A unit is counted as differing if its status or any of these fields differs. These counts describe differences between delivered sets; Full MetaRigor is not treated as the gold standard for all units, and difference counts are not interpreted as error counts.

Table 17: Unit-level differences among the three systems' DE artifacts.

System	Study × Target units	REPORTED units	Units differing from Full in status or any of the four fields
Full MetaRigor	42	30	—
Codex CLI	42	24	6
OpenHands CLI	42	26	5

The Protocol target for SELECT2 × result.death.day90 is "Death from any cause within 90 days." The original SELECT2 report lists both this outcome and "In-hospital death." The former is 68/177 vs 71/171 (Table 3 of the original report), whereas the latter is 42/178 vs 44/174 (Table 2 of the original report) (Sarraj et al., 2023). Full MetaRigor selected the former, whereas OpenHands CLI bound the latter to the same 90-day mortality Target. This is an outcome/follow-up identity mismatch, rather than the generation of numbers absent from the source. The corresponding identity verification is summarized in Table 18.

Table 18: Identity verification for the SELECT2 target result.

Verification item	Full MetaRigor	Codex CLI	OpenHands CLI
Delivered values (EVT vs MED)	68/177 vs 71/171	Not reported	42/178 vs 44/174
Outcome actually represented by values	90-day all-cause mortality	—	In-hospital death
Relationship to target identity	Outcome and time point match	Target result not delivered	Outcome / time-point mismatch

G.3 Statistical Inputs, RoB Binding, and Final Reporting

Statistical inputs. For each system's reported result.death.day90 results, event counts and sample sizes are summed separately for the two groups. Let S_s be the set of contributing studies for system s, and let $e_{i,g}$ and $n_{i,g}$ be the event count and sample size, respectively, for group g in study i, where g is EVT or MED. The descriptive risk ratio based on aggregate event counts is:

$$\mathrm{RR}_{desc,s} = \frac{\sum_{i\in S_s} e_{i,\mathrm{EVT}} \Big/ \sum_{i\in S_s} n_{i,\mathrm{EVT}}}{\sum_{i\in S_s} e_{i,\mathrm{MED}} \Big/ \sum_{i\in S_s} n_{i,\mathrm{MED}}}. \tag{11}$$

This calculation is used only to transparently examine differences in the delivered sets and numerical values. It does not replace the prespecified meta-analysis model and is not used to report a formal pooled effect, confidence interval, or significance test. The corresponding descriptive statistical inputs are summarized in Table 19.

Table 19: Descriptive statistical inputs for 90-day mortality.

System	Contributing studies	EVT events / N	MED events / N	Descriptive risk ratio
Full MetaRigor	6	295/937	343/932	0.855
Codex CLI	5	227/760	272/761	0.836
OpenHands CLI	5	212/780	225/771	0.931

RoB binding. Codex CLI did not deliver this target result for SELECT2, and its corresponding RoB status was NOT_ASSESSED. OpenHands CLI assigned LASTE the status NOT_ASSESSED and delivered overall labels for RESCUE-Japan-LIMIT and TENSION that differed from those of Full MetaRigor. Table 20 retains the statuses of all six studies for direct cross-system comparison. NOT_ASSESSED denotes absence of assessment, not a risk level.

Table 20: Overall RoB statuses for 90-day mortality.

Study	Full MetaRigor	Codex CLI	OpenHands CLI
ANGEL-ASPECT	SOME_CONCERNS	SOME_CONCERNS	SOME_CONCERNS
LASTE	SOME_CONCERNS	SOME_CONCERNS	NOT_ASSESSED
RESCUE-Japan-LIMIT	SOME_CONCERNS	SOME_CONCERNS	LOW
SELECT2	HIGH	NOT_ASSESSED	HIGH
TENSION	SOME_CONCERNS	SOME_CONCERNS	LOW
TESLA	SOME_CONCERNS	SOME_CONCERNS	SOME_CONCERNS

Final reporting. OpenHands CLI's mismatched SELECT2 values, 42/178 vs 44/174, appeared in the final supplementary table alongside the target identity of 90-day mortality. This directly shows that genuine numbers from the wrong outcome or time point can pass beyond the DE stage and become incorrect facts in the final deliverable.

G.4 Interpretation and Offline Reproduction

The statistical differences reflect changes in both the contributing study sets and the numerical values; the entire shift in the risk ratio from 0.855 to 0.931 is not attributed solely to SELECT2. Missing objects and label differences in the RoB artifacts are also recorded separately, without attributing all label changes to a single DE mismatch. The shared observation supported by the three systems' artifacts is that local result mismatches or omissions can propagate into statistical inputs, the objects of risk-of-bias assessment, and final reports.

Offline calculations generated reports and machine-readable results from the fixed archived artifacts, with identical outputs across two executions. This reproduction check concerns the determinism of the offline reporting and calculation procedures; it does not represent rerunning the three systems' models. Historical artifacts remained unchanged throughout the analysis.