AutoMETA: A Multi-Agent LLM System for Autonomous Meta-Analysis

Keeheon Lee
Yonsei University
Seoul, Republic of Korea
keeheon@yonsei.ac.kr

Kunhee Ryu
Yonsei University
Seoul, Republic of Korea
rgh00826@yonsei.ac.kr

ABSTRACT

Meta-analysis is a statistical method to draw a conclusion from synthesizing the scientific evidence of multiple similar studies on a common research question. However, its execution requires manual tasks including paper selection, data extraction, and statistical pooling. There has been automation efforts for executing partial or whole tasks in a divide-and-conquer manner by a single or multiple AI agents. Yet, meta-analysis using multiple agents in a decentralized and distributed manner under an explicit and auditable protocol is not examined. We introduce **AutoMETA**, a system of multiple LLM agents that assigns one agent per subject paper of a target meta-analysis. Each agent extracts page-anchored outcomes, critiques peer outputs, and revises entries before passing the finalized dataset to a non-agent statistical module implementing DerSimonian–Laird random-effects pooling with heterogeneity diagnostics. We compare three systems under identical conditions: a **Human reference** (published meta-analyses), a **Single-LLM** pipeline without coordination, and **AutoMETA** (multi-agent with inter-agent critique). As a case study, we selected meta-analyses papers on cardiology in 2004 (8 out of 15 papers meeting inclusion criteria). AutoMETA achieved a median relative effect-size error of 6.4% against human-authored references, compared to 34.0% for the single-agent baseline, while ablation analysis confirmed that removing critique or protocol enforcement progressively degraded accuracy (12.4%, 25.1%, 28.1%). These findings suggest that structured inter-agent critique and procedural enforcement can yield reliable, reproducible statistical reasoning without relying solely on model capability.

KEYWORDS

Multi-Agent Systems; Large Language Models; Meta-Analysis; Decentralized Decision Making; Inter-Agent Critique

ACM Reference Format:
Keeheon Lee and Kunhee Ryu. 2026. AutoMETA: A Multi-Agent LLM System for Autonomous Meta-Analysis. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 10 pages. https://doi.org/10.65109/HXKA2256

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/HXKA2256

1 INTRODUCTION

Meta-analysis integrates results from multiple independent studies to produce quantitative, generalizable evidence across domains such as medicine and social science [2, 5, 9, 15, 16, 39]. It enables researchers to move beyond individual findings through systematic aggregation of effect sizes and diagnostic performance. However, performing a valid meta-analysis remains labor-intensive and error-prone: reviewers must identify eligible studies, extract numerical data from heterogeneous tables and figures, manage missing or zero-valued entries, and apply consistent statistical models (e.g., DerSimonian–Laird random effects) [8, 14]. Even small deviations in extraction or protocol adherence can alter pooled estimates and undermine reproducibility [23, 27].

Recent large language models (LLMs) can summarize documents, extract structured data, and perform limited reasoning across long contexts [35, 37, 43]. Yet their use in statistically rigorous meta-analysis remains limited [23]. Most existing pipelines rely on a single model acting independently, without cross-verification or enforcement of statistical rules [1, 30, 37, 43]. As a result, they often conflate analysis units (e.g., patient vs. vessel), omit uncertain values, or deviate from inclusion criteria—producing unstable and irreproducible pooled estimates. This study asks whether multiple LLM agents, when coordinated under an explicit and auditable protocol, can perform meta-analysis with human-level reliability. We evaluate three systems under identical settings: (1) a **Human reference** (published meta-analyses), (2) a **Single-LLM** pipeline without coordination, and (3) **AutoMETA**, which incorporates inter-agent critique and protocol safeguards.

We propose **AutoMETA**, a multi-agent LLM system for *autonomous meta-analysis*. Each primary study is assigned to a *study-centered agent* that extracts page-anchored quantitative data following a fixed protocol. Agents exchange critiques to resolve inconsistencies in units, thresholds, or zero-cell corrections, while a *protocol engine* enforces consistency rules. After consensus is reached, a centralized *statistical synthesizer* performs random-effects pooling (DerSimonian–Laird; HSROC) and reports heterogeneity diagnostics (Q, τ^2, I^2). Coordination emerges from structured dialogue rather than fixed role hierarchies, forming a decentralized and auditable chain of statistical reasoning. To evaluate AutoMETA, we compare it against (a) a *human-authored* meta-analysis and (b) a *single-agent* LLM pipeline using identical statistical formulas as an external reference.

We employ a cardiology corpus—spanning treatment comparisons, diagnostic accuracy studies, and biomarker associations, with heterogeneous outcomes, multiple analysis units, and legacy reporting styles—as a stress test, while keeping the domain incidental

to our central question: *Can multiple LLM agents achieve protocol-faithful, reproducible pooling at scale?* Figure 1 illustrates these three paradigms.

Contributions. This paper makes the following key contributions. We present **AutoMETA**, a multi-agent LLM system that performs meta-analysis collaboratively under a unified and auditable protocol; we demonstrate that **structured inter-agent critique** and **rule-based procedural enforcement** jointly lead to human-comparable pooled estimates while reducing protocol violations; and we provide a **reproducible framework** that supports transparent evaluation of cooperative, protocol-faithful reasoning among large language models.

2 RELATED WORKS

2.1 Multi-Agent LLMs and Coordination

Recent frameworks such as AutoGen [44] and CAMEL [22] organize LLMs as cooperating agents for task decomposition and dialogue-based reasoning, and debate-style settings have shown that structured disagreement improves factual accuracy [10]. Multi-agent systems have also been applied in scientific domains including protein design [12], chemical synthesis [11], and materials discovery [26]. However, these frameworks target open-ended reasoning or domain experimentation rather than tasks requiring strict statistical protocols and verifiable consensus. Despite progress in agentic collaboration, few works explore coordination for reproducible evidence synthesis—a gap that motivates AutoMETA.

2.2 Tool-Augmented Reasoning and Protocol Following

Tool-augmented prompting such as ReAct [45] interleaves reasoning with actions like retrieval or computation, and later systems add verification layers to align reasoning with domain rules [6]. However, tool-based LLMs still act reactively without enforcing cross-document consistency [18, 31]. AutoMETA adopts controlled tool use and rule-based guards for unit, threshold, and continuity correction [19], shifting from reactive tool calls to cooperative protocol fidelity.

2.3 Self-Reflection and Test-Time Refinement

At the single-model level, methods such as Self-Consistency [42], Reflexion [36], Self-Refine [28], and process-level verification [13] improve reliability through iterative critique and revision of intermediate steps. However, these approaches process all documents within a single context window, causing fine-grained field-level information—such as specific units, diagnostic thresholds, and outcome time points—to be compressed or lost during reasoning. This limitation is critical for meta-analysis, where cross-study comparability depends on preserving exactly such details. AutoMETA addresses this by assigning each agent to a single study, preserving granular extraction fidelity; cross-study consistency is then enforced through structured inter-agent critique rather than within-model reflection, creating verifiable audit trails that single-model methods cannot provide.

2.4 AI-Driven Meta-Analysis and Evidence Synthesis

LLMs have assisted specific meta-analysis phases—screening [25, 41], tabular extraction [46], and summarization [32, 38]—while RAG-based systems improve retrieval quality [1]. Mutinda et al. [30] built a BioBERT pipeline for PICO extraction and numeric parsing, and earlier tools such as RobotReviewer [29] and ASReview [40] enhanced screening via machine learning. Datasets such as Evidence Inference [21] further highlight challenges in tracing extracted data to source text—a problem AutoMETA addresses through mandatory page anchoring. Yet all such approaches operate as single agents and omit end-to-end statistical pooling or cross-verification under an auditable protocol.

2.5 Statistical Foundations for Diagnostic Meta-Analysis

AutoMETA follows established standards: the DerSimonian–Laird random-effects estimator [8], bivariate and HSROC models for sensitivity and specificity [3], and heterogeneity metrics Q, τ^2, and I^2 [20]. These enable direct comparison between human and LLM-based analyses within a conventional statistical framework.

2.6 Positioning

Unlike prior multi-agent or self-improvement frameworks focused on abstract reasoning or planning, **AutoMETA** frames meta-analysis as a protocol-driven coordination problem. Each agent acts as an independent reviewer operating under identical procedural rules, contributing page-anchored evidence and verifying peers through structured dialogue. This design combines procedural rigor with cooperative reasoning: it yields statistically valid pooled estimates comparable to human analyses while improving rule compliance and reproducibility. By linking agentic coordination, constrained reasoning, and evidence synthesis, AutoMETA connects multi-agent research with the practical need for verifiable, transparent large-scale reasoning.

3 METHODS

3.1 Task and Corpus

We cast meta-analysis as a cooperative reasoning task in which a population of study-centered LLM agents jointly execute a fixed, human-authored synthesis protocol. Each agent is assigned to one primary study drawn from an existing meta-analysis, extracts page-anchored quantitative evidence, critiques the extractions of its peers, and participates in a shared decision on the final pooled estimate. To build a representative yet tractable evaluation corpus, we retrieved all cardiology-related meta-analyses published in 2004 from Scopus using the query: `(TITLE-ABS-KEY("cardiology") AND TITLE-ABS-KEY("meta-analysis")) AND (SUBJAREA = MEDI) AND (DOCTYPE = ar)`. This year was selected to balance sample size and experimental feasibility—providing enough studies for multi-agent evaluation while ensuring consistent reporting formats and full-text accessibility for reproducible analysis. Importantly, 2004 also marks the period when *meta-analysis became widely adopted as a*

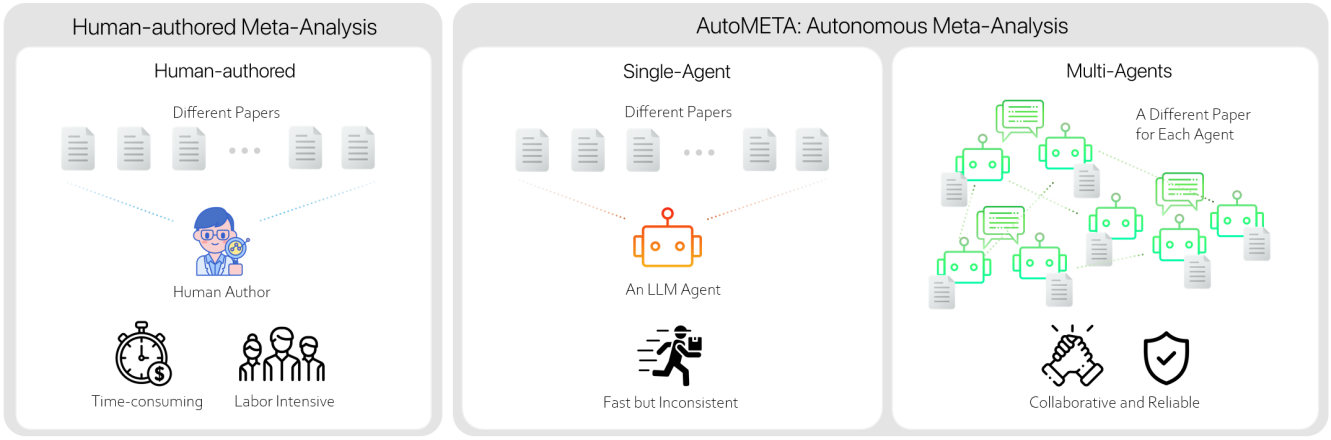

Figure 1: Overview of Meta-Analysis Paradigms. The figure contrasts three modes of evidence synthesis: (a) *Human-authored* meta-analysis, where a single human expert manually extracts and pools results across studies; (b) *Single-agent* automation, where one LLM processes all studies efficiently but lacks peer verification and cross-study consistency checks, making it vulnerable to uncorrected extraction errors; and (c) *AutoMETA (multi-agent)*, where study-centered agents handle distinct papers, exchange critiques, and reach a protocol-guided consensus. This progression illustrates the transition from labor-intensive but reliable human synthesis, to efficient yet verification-limited single-agent reasoning, and finally to collaborative, reproducible multi-agent analysis.

standard methodology in clinical and biomedical research, following the formalization of reporting and statistical practices in early systematic review frameworks.

Among the 15 retrieved papers, 7 were excluded due to irrelevance or inaccessible full texts, leaving 8 eligible meta-analyses comprising a total of 114 primary studies (with individual meta-analyses containing 7, 28, 25, 22, 8, 4, 11, and 9 studies, respectively). From each included paper, we compiled the set of primary studies analyzed by the authors; these studies constitute the input for AutoMETA, the single-agent baseline, and the human reference comparison. Each experiment was repeated 10 times to average stochastic variation and ensure statistical stability.

The resulting corpus spans diverse subdomains and reporting conventions, intentionally preserving heterogeneity in outcome measures, analysis units, and presentation formats. This variation provides a realistic stress test for evaluating whether multi-agent coordination and procedural reasoning can maintain protocol-faithful and reproducible pooling across inconsistent data sources. Although the studies originate from cardiology, the medical content itself is incidental—the central goal is to assess the reliability and generalizability of structured cooperation in complex evidence-synthesis settings. Figure 2 summarizes the end-to-end AutoMETA workflow from preprocessing through extraction, critique, revision, and pooling.

3.2 Harmonized Protocol

To ensure comparability across systems, all models follow the same explicitly specified and auditable meta-analysis protocol. The protocol defines inclusion and exclusion criteria, the format of extracted outcomes, and the statistical models to be applied. Only adult, original-data studies are included; reviews, case reports, and duplicates are excluded. Binary diagnostic accuracy outcomes are represented as 2×2 contingency tables (TP, FP, FN, TN), following the canonical thresholding criterion used in the evaluation corpus (typically 50% stenosis in cardiac imaging studies). Continuous outcomes (e.g., CNR, SNR, or visible length/diameter) are standardized as Hedges g to ensure comparability across heterogeneous measurement scales. Each analysis is restricted to a single unit type and handled with explicit subgroup definitions for contrast enhancement and reader identity. A continuity correction of 0.5 (a conventional practice in diagnostic test meta-analysis) is applied to any zero cell, and random-effects DerSimonian–Laird (DL-RE) pooling is used to compute Q, τ^2, and I^2, with HSROC optionally reported for ablation experiments. Every numeric entry is automatically linked to its source page, table, or figure reference; uncertain entries are left as `NULL` with an explanatory note. This harmonized protocol guarantees that any performance difference arises solely from differences in coordination, not in modeling or data preprocessing. Its statistical components adhere to canonical standards for diagnostic test accuracy meta-analyses [7, 17, 33], providing a transparent and reproducible foundation for evaluating LLM-driven reasoning.

3.3 Multi-Agent Coordination through Rounds

Rather than assigning fixed functional roles, AutoMETA structures collaboration into four iterative rounds, each representing a collective phase of reasoning among study agents. In Round 0, each agent independently reads its assigned paper, follows the standardized protocol, and produces a structured summary of extracted data. Round 1 initiates peer critique: agents exchange and review one another's initial summaries, offering concise comments on potential inconsistencies in analysis units, thresholds, or cited evidence. In Round 2, each agent revises its extraction table by incorporating valid peer feedback and verifying the supporting evidence, resulting

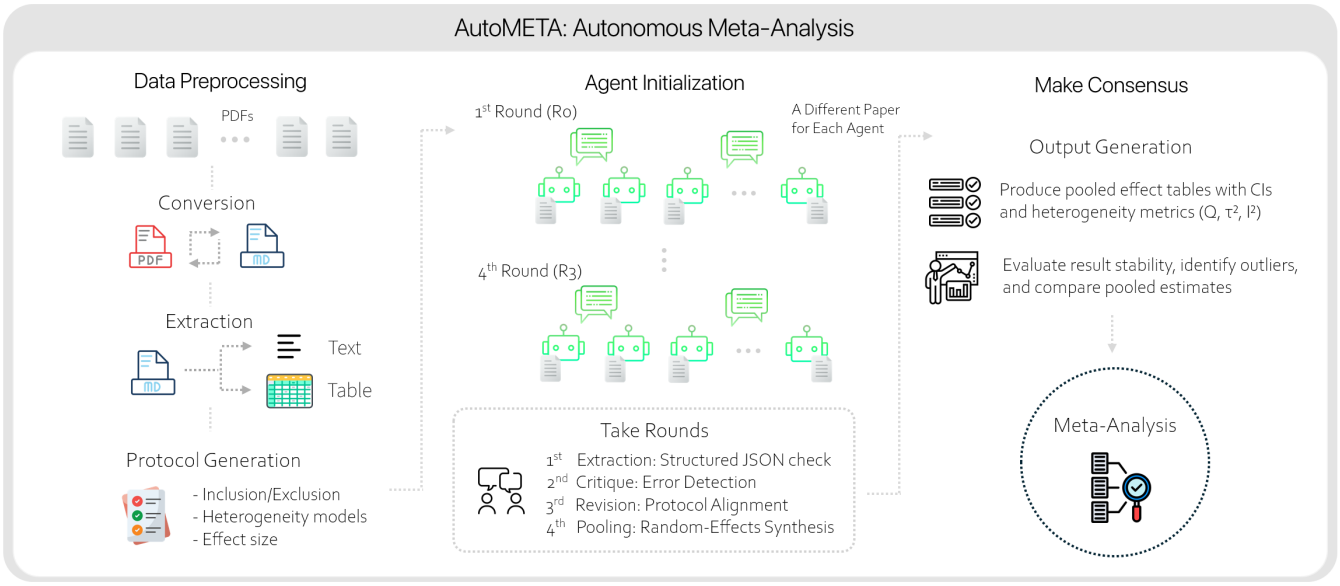

Figure 2: AutoMETA Framework: End-to-End Meta-Analysis Pipeline. This figure illustrates the overall methodological workflow of AutoMETA, showing how heterogeneous studies are converted, standardized, and synthesized into a reproducible meta-analytic consensus. The pipeline begins with *data preprocessing*, where PDFs are converted into text and tables, and a protocol is generated from either a reference paper or from heterogeneous studies specifying inclusion/exclusion criteria, heterogeneity models, and effect-size definitions. During *agent initialization*, each paper is assigned to a dedicated study agent that iteratively proceeds through four reasoning rounds: (1) *Extraction*—extracting study-level data and producing structured JSON summaries; (2) *Critique*—peer verification and error detection among agents; (3) *Revision*—protocol alignment and correction of inconsistencies; and (4) *Pooling*—random-effects synthesis and heterogeneity diagnostics (Q, τ^2, I^2). Finally, the *consensus module* generates pooled effect tables with confidence intervals and stability analysis, evaluating outliers and comparing the system's pooled estimates to human-authored meta-analyses. Together, these stages form a fully reproducible, protocol-faithful pipeline for autonomous evidence synthesis.

in a refined dataset. Finally, in Round 3, all revised summaries are merged and used for a DerSimonian–Laird random-effects pooling with 95% confidence intervals and heterogeneity diagnostics (Q, τ^2, I^2). During this final round, agents compare their individual study results to the collective pooled estimate and reflect on alignment or deviation. The outcome is a consensus dataset that reflects protocol-faithful, auditable agreement across all participating agents. Coordination in AutoMETA thus emerges organically from dialogue and revision, rather than from top-down orchestration or rigid role specialization.

3.4 Single-LLM and Human Baselines

To evaluate the contribution of inter-agent interaction, we compare AutoMETA to two baselines: a single-LLM pipeline and a human-authored meta-analysis. The single-LLM baseline executes all steps—data extraction and pooling—sequentially under the same protocol but without peer review or iterative revision. Uncertain extractions are left unfilled rather than reconciled. The human reference corresponds to a previously published meta-analysis on the same corpus, serving as an external benchmark and ground truth for pooled estimates.

3.5 Evaluation and Metrics

Each system produces a unified meta-analytic dataset consisting of study-level observations annotated by analysis unit and subgroup. Rows identified as non-includable are excluded from the pooling stage. We assess performance along three core dimensions central to coordinated evidence synthesis: (1) **protocol faithfulness**, measured by adherence to unit definitions, thresholding rules, and continuity-correction requirements, as well as by the number of critical violations per 100 extracted rows; (2) **extraction coverage and consistency**, quantified by the proportion of successfully filled entries, disagreement rates among agents, and the share of disagreements resolved through critique and revision; and (3) **statistical reliability**, defined as the deviation from human pooled estimates (absolute difference in pooled effect size and heterogeneity indices Q, τ^2, and I^2). Together, these metrics capture the balance between procedural rigor, extraction accuracy, and reproducibility across systems.

3.6 Ablations and Implementation

All experiments are implemented using the **TinyTroupe** multi-agent orchestration framework, which provides lightweight dialogue management, shared episodic memory, and reproducible

round-based coordination among agents [34]. The framework supports both single-agent and multi-agent configurations within a unified execution environment. In the single-agent baseline, one model processes all papers sequentially under a unified protocol without any peer discussion. In the multi-agent setting, each study is assigned to an independent agent that participates in four coordination rounds (extraction, critique, revision, and pooling). Ablation variants disable specific coordination or control mechanisms to isolate their contributions. The *No-Critique* condition removes the critique and revision rounds, preventing agents from challenging or correcting each other's extractions. The optional *No-Protocol* configuration omits rule-based safeguards such as unit validation, threshold checks, and continuity corrections, testing the effect of removing explicit procedural control. These comparisons evaluate how structured collaboration and rule enforcement contribute to the reliability and reproducibility of the resulting meta-analyses.

All runs employ GPT-class large language models with identical decoding parameters (temperature = 0) and restricted reasoning modes limited to document inspection and in-dialogue deliberation, without the use of external web search or computational tools. The model stack consists of `gpt-4.1-mini` for dialogue and extraction, `o3-mini` for deterministic reasoning and consensus verification, and `text-embedding-3-small` for embedding-based memory retrieval. Each multi-agent dialogue includes up to two critique rounds, and all transcripts, summaries, and pooled results are archived to ensure full transparency and reproducibility.

To facilitate independent verification, we release all protocol specifications, prompts, and round transcripts in our GitHub repository.[1] This repository enables complete replication of the reported results and serves as a reference implementation of AutoMETA's multi-agent coordination framework.

3.7 Threats to Validity

The quality of extracted data may depend on the completeness of the original study reports, which we mitigate through mandatory page anchoring and exclusion rules. Variability among LLMs could influence extraction coverage, for which we report rerun stability. Although evaluated in a cardiology setting, the round-based coordination framework is domain-independent and can generalize to other evidence-synthesis contexts such as safety reporting, policy evaluation, or benchmark aggregation.

4 RESULTS

We evaluate AutoMETA against a single-agent LLM pipeline and a human-authored meta-analysis reference across eight cardiology meta-analyses (2004). All systems are assessed along three axes: (E1) protocol faithfulness, (E2) extraction coverage and consistency, and (E3) statistical reliability. Unless stated otherwise, results are aggregated across all study sets; effect-size deviations from the human reference are expressed as relative errors (both mean and median), and heterogeneity deviations as mean absolute differences with 95% CIs. For each meta-analysis, we averaged the pooled estimates across 10 independent runs; the reported means and medians are computed over these 8 per-paper averages.

[1] All codes are available at `https://github.com/AutoMETA/rgh112`.

4.1 Quantitative Evaluation

Table 1 summarizes the deviations between LLM-based systems and the human-authored baselines. We report both mean and median relative errors because the distribution is heavily skewed: a small number of outlier studies—typically caused by LLM unit-conversion mistakes—inflate the mean while the median better reflects typical-case performance.

AutoMETA (Multi, Full) achieves the best median effect-size accuracy among all configurations ($\Delta_{\text{rel}}(\%)^{\text{med}}$=6.4%), indicating that on most papers the system closely approximates human pooled estimates. The gap to the mean ($\Delta_{\text{rel}}(\%)^{\text{mean}}$=40.0%) reveals that a few papers with extraction-level outliers substantially affect the average.

Ablation results show a consistent ordering in median accuracy: **Full** (6.4%) < **NoCritique** (12.4%) < **NoProtocol** (25.1%) < **NoCrit + NoProt** (28.1%). While partially overlapping confidence intervals preclude definitive statistical separation between adjacent configurations, the directional pattern is consistent across all four conditions, suggesting that critique and protocol enforcement contribute complementary benefits. Removing critique roughly doubles the median error while also allowing more studies into the pooling stage (higher k), which stabilises heterogeneity estimates ($|\Delta Q|$=6.8, the lowest among multi-agent configurations) at the cost of uncorrected extraction errors propagating into pooled estimates. Removing the protocol quadruples the median error; disabling both yields the worst performance.

Single-agent behavior. The **Single (Full)** pipeline exhibits a bimodal profile: on favourable papers it nearly matches the human reference ($\Delta_{\text{rel}}(\%) < 4\%$), while on others it produces large deviations ($\Delta_{\text{rel}}(\%) > 60\%$), yielding an overall median of 34.0%. Notably, **Single (NoProtocol)** shows the lowest mean $\Delta_{\text{rel}}(\%)$ (21.4%), but this result should be interpreted with caution: heterogeneity diagnostics were available for only a subset of papers, precluding reliable CI estimation (†), suggesting that the pipeline completed the full workflow for only a subset of papers. A system that produces accurate effect sizes on select papers but cannot complete the full meta-analytic workflow does not constitute a reliable pipeline.

Heterogeneity deviations. Multi (Full) shows elevated $|\Delta I^2|$=30.6 compared to ablated variants (13.6–23.9), while NoCritique achieves the lowest $|\Delta Q|$ and $|\Delta \tau^2|$. This reflects an accuracy–coverage tradeoff: the full critique–protocol pipeline applies stricter inclusion filtering, reducing k to 2–3 studies in some papers, where small perturbations shift I^2 by tens of percentage points. Without critique, more studies survive inclusion, stabilising heterogeneity at the cost of effect-size fidelity. This dissociation means the ablation ordering that holds for effect-size accuracy does *not* extend to heterogeneity metrics, where the relationship is non-monotonic. Resolving this tension—e.g., through confidence-weighted inclusion rather than binary accept/reject—represents a key direction for future work.

4.2 Protocol Faithfulness and Error Correction

AutoMETA substantially reduces rule violations related to unit inconsistency and continuity correction. Audits of dialogue traces indicate that the majority of detected errors originate from inter-agent critique rounds, confirming that critique acts as a functional

Table 1: Deviations from the human-authored baselines. Entries report relative effect-size errors (mean with 95% CI; median with IQR) and mean absolute errors for heterogeneity diagnostics with 95% CIs. Ablations remove key components of coordination: *No-Critique* disables peer review, *No-Protocol* removes explicit procedural constraints, and their combination illustrates the breakdown of coordinated synthesis when both mechanisms are absent.

System	$\Delta_{\text{rel}}(\%)^{\text{mean}}$	$\Delta_{\text{rel}}(\%)^{\text{med}}$	$\lvert\Delta Q\rvert$	$\lvert\Delta\tau^2\rvert$	$\lvert\Delta I^2\rvert$
Multi (Full)	40.0 [5.8, 88.9]	**6.4** [3.9, 45.0]	13.9 [8.9, 18.2]	0.076 [.034, .118]	30.6 [6.2, 67.2]
Multi (NoCritique)	51.6 [8.6, 126.8]	12.4 [6.2, 32.4]	6.8 [2.4, 11.2]	0.024 [.014, .040]	23.9 [10.9, 34.3]
Multi (NoProtocol)	64.4 [9.2, 152.3]	25.1 [9.9, 79.6]	15.1 [10.4, 19.5]	0.065 [.018, .116]	13.6 [5.8, 21.2]
Multi (NoCrit + NoProt)	50.2 [14.0, 98.3]	28.1 [14.0, 64.1]	14.4 [11.7, 16.9]	0.057 [.021, .093]	18.7 [9.0, 36.2]
Single (Full)	38.5 [10.3, 71.1]	34.0 [23.6, 48.8]	9.8 [1.4, 22.3]	0.021 [.016, .027]	23.5 [0.0, 36.3]
Single (NoProtocol)	21.4 [5.6, 37.2]	22.6 [7.1, 37.0]	10.0†	0.032†	10.5†

Notes: Effect-size deviations are $\Delta_{\text{rel}}(\%) = 100 \times |\hat{\theta}_{\text{model}} - \hat{\theta}_{\text{human}}|/|\hat{\theta}_{\text{human}}|$. Mean errors are reported with 95% CIs; median errors are reported with interquartile ranges (IQR, Q1–Q3), providing a robust summary of the central 50% of per-paper deviations. The gap between mean and median reflects the influence of outlier studies (e.g., LLM unit-conversion errors), which disproportionately inflate the mean. Because effect-size scales differ across papers (log OR, WMD, Hedges g), only relative errors—not absolute differences—are aggregated. Heterogeneity deviations ($|\Delta Q|$, $|\Delta\tau^2|$, $|\Delta I^2|$) are means with 95% CIs. †Insufficient data to compute a reliable CI.

analogue of human peer review. Agents frequently identify inconsistencies such as mismatched denominators, ambiguous thresholds, or missing zero-cell corrections, prompting targeted revisions before pooling. The **No-Protocol** variant fails to resolve such discrepancies, producing inconsistent inclusion criteria and elevated τ^2 deviations (up to $|\Delta\tau^2|$=0.116). Figure 3 visualizes a representative interaction where a peer agent flags a threshold violation ("0.7 vs. 0.5"), leading to a corrected log(OR) from 1.21 to 0.94 and validated inclusion in the pooled analysis.

However, the same critique mechanism that improves extraction quality can also lead to over-filtering. When critique rounds repeatedly flag borderline studies as insufficiently documented, agents may exclude them from the final dataset, reducing k and destabilizing heterogeneity estimates. This over-filtering effect is most pronounced in Multi (Full), where the combination of critique and protocol enforcement creates a high inclusion bar. Future work should explore adaptive thresholds that balance inclusion breadth with extraction quality.

4.3 Critique Dynamics and Process Trace

Beyond aggregate accuracy, the critique trace in Figure 3 shows how AutoMETA performs *cooperative self-correction* in real time. In each critique–revision cycle, agents exchange structured feedback, update local assumptions (e.g., thresholds, denominators), and revalidate consistency through the protocol engine.

Rather than acting as passive reviewers, agents function as distributed auditors that uncover latent contradictions in extraction logic—such as mismatched thresholds or missing continuity corrections—and trigger targeted local rewrites. This mirrors the human dual-reviewer paradigm in systematic reviews, where independent extraction followed by reconciliation safeguards epistemic soundness. In AutoMETA, these cycles are not scripted; they emerge from the dialogue policy itself and produce verifiable audit trails of how consensus is reached.

The R0–R3 trace illustrates this mechanism concretely: a single mis-specified parameter ("0.70 → 0.50") propagates through structured dialogue into a verified correction, leading to convergence in both the reported value (log(OR) = 0.94) and protocol compliance. Such trace-level transparency provides not only interpretability but also a basis for post-hoc reliability auditing—something rarely attainable in purely end-to-end LLM systems.

4.4 Error Dynamics and Breakdown

Table 2 summarises the distribution of protocol-engine flags across all model runs (9,447 validation entries, 66,988 issues). The two most frequent categories—**threshold / outcome-type inconsistencies** (24%) and **ambiguous inclusion labels** (22%)—stem from agents misclassifying outcome types (e.g. labelling a diagnostic accuracy measure as a generic continuous variable) or applying inconsistent eligibility criteria. These structural flags are raised by the protocol engine at Round 0 and subsequently addressed during critique rounds 1–2, where peer agents cross-check JSON schemas against page-anchored evidence.

A second cluster—**denominator misalignment** (19%) and **zero-cell omission** (12%)—reflects incomplete specification of contingency-table cells. When sample sizes are missing from the extraction, the protocol engine cannot verify whether zero-cell continuity corrections are needed; critique rounds surface these gaps by requesting explicit denominators and flagging uncorrected zero cells.

Residual categories—**citation / page-reference mislinking** (12%) and **unit mismatch** (11%)—are largely detected at Round 0 by automated schema checks (e.g. `study_id` does not match assigned paper, analysis unit not specified). These persist at lower rates after critique because they depend on document-level parsing rather than inter-agent deliberation.

Across configurations, critique improved extraction quality most visibly on structurally complex papers, with per-paper validation pass-rate gains of 8–15 percentage points on challenging cases. The aggregate pass rate across all papers was similar between configurations (88% with critique vs. 87% without), as critique primarily targets the tail of difficult extractions rather than uniformly improving all entries. The boundary between *resolvable disagreement*—threshold and denominator issues that critique can fix—and *epistemic opacity*—citation ambiguities rooted in the source PDFs themselves—defines the practical ceiling of the current system.

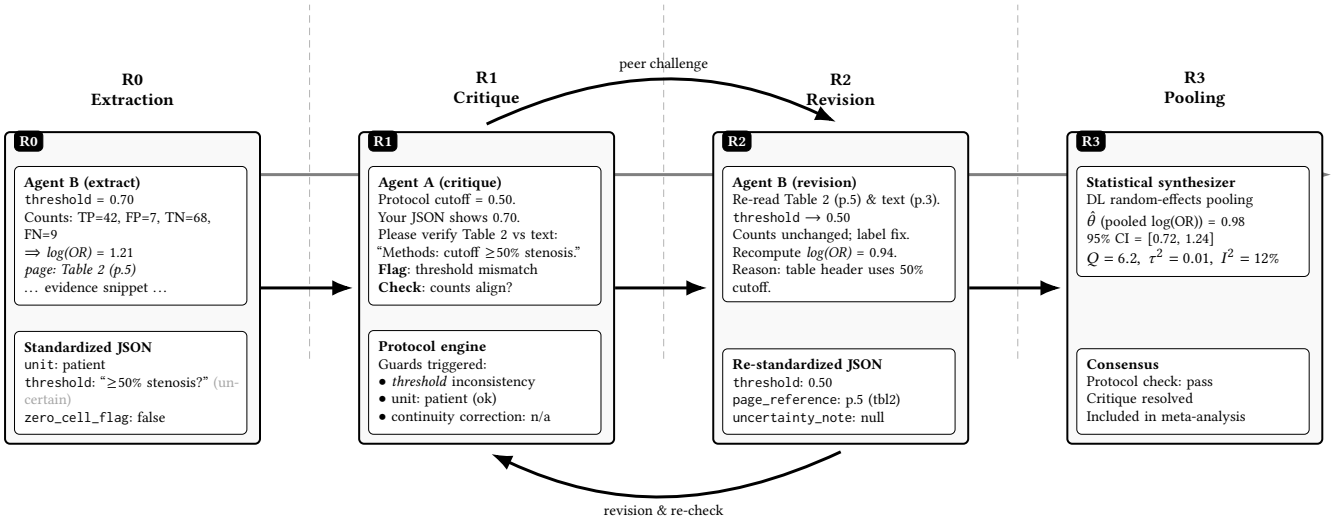

Figure 3: Representative critique–revision trace (Rounds 0–3). A study agent initially extracts outcomes with a mis-specified threshold (0.70). A peer agent flags a protocol violation (required cutoff: 0.50), prompting a source re-check and JSON restandardization. After correction, the study's *log(OR)* is updated (1.21→0.94) and passes protocol guards. The statistical synthesizer then includes the corrected entry in DerSimonian–Laird random-effects pooling and reports heterogeneity diagnostics (Q, τ^2, I^2). This exemplifies how inter-agent critique functions as a distributed quality-control layer that converts informal disagreement into auditable, protocol-faithful consensus.

Table 2: Distribution of detected errors across all model runs (9,447 validation entries, 66,988 flags). Most flags relate to outcome specification and denominator gaps, both systematically reduced by inter-agent critique.

Error Type	Frequency (%)	Detected by Round
Threshold inconsistency (e.g., outcome misclassification)	24	0–1
Ambiguous inclusion criteria or label mapping	22	1–2
Denominator misalignment in contingency tables	19	0–1
Zero-cell omission / continuity correction gap	12	1–2
Citation or page-reference mislinking	12	0
Unit mismatch (patient vs. vessel vs. segment)	11	0

5 DISCUSSION

AutoMETA demonstrates that structured multi-agent coordination can approximate human-level meta-analytic reasoning, achieving a median relative effect-size error of 6.4% across eight cardiology meta-analyses. However, the results also reveal important limitations and trade-offs that qualify the scope of this achievement.

5.1 Coordination as a Source of Reliability

The ablation analysis provides evidence that **coordination**—not model scale or prompt engineering—is a principal driver of effect-size accuracy. The consistent ordering of median $\Delta_{\text{rel}}(\%)$ across configurations (Full: 6.4% → NoCritique: 12.4% → NoProtocol: 25.1% → NoCrit+NoProt: 28.1%) suggests that critique and protocol enforcement each contribute independently, though this interpretation should be tempered by the wide confidence intervals that partially overlap between adjacent conditions. Critique catches numerical errors (unit conversions, threshold mismatches), while the protocol prevents structural inconsistencies (mixed analysis units, missing zero-cell corrections). Only their combination yields the best effect-size accuracy.

However, this complementarity is not uniformly additive. The full configuration's heterogeneity deviations are the *highest* among multi-agent variants ($|\Delta I^2|$=30.6), while removing critique paradoxically *improves* heterogeneity alignment ($|\Delta Q|$: 13.9 → 6.8; $|\Delta\tau^2|$: 0.076 → 0.024). This dissociation confirms that critique and protocol contribute complementary but partially competing benefits: together they maximise effect-size fidelity, yet the resulting study-filtering pressure destabilises heterogeneity estimation.

The reliability profiles of single-agent and multi-agent systems differ qualitatively. Specifically, the single-agent pipeline shows moderate accuracy ($\Delta_{\text{rel}}(\%)^{\text{med}}$=34.0%) but frequently omits heterogeneity diagnostics, whereas multi-agent coordination ensures complete outputs at the cost of the accuracy–coverage trade-off discussed in Section 5.2.

5.2 The Accuracy–Coverage Trade-off

A central finding of this work is that stricter quality enforcement improves per-study accuracy but reduces the number of studies included in pooled analyses. Multi (Full) achieves the best effect-size accuracy precisely because its critique rounds filter out ambiguous or poorly extracted studies. However, this filtering reduces k, making heterogeneity statistics (Q, τ^2, I^2) inherently unstable. In one paper, Multi (Full) included only k=2 studies, producing I^2=86% versus the human reference of I^2=0%—a deviation driven entirely by insufficient sample size rather than systematic error.

This trade-off mirrors a well-known challenge in human-conducted meta-analysis, where stricter inclusion criteria improve internal validity but reduce statistical power. AutoMETA's critique mechanism amplifies this effect: agents that identify potential extraction issues may conservatively exclude borderline studies rather than attempt uncertain corrections. Future work should explore confidence-weighted pooling, where studies with uncertain extractions contribute to the analysis with reduced weight rather than being excluded entirely.

5.3 Transparency and Procedural Rigor

Unlike typical LLM pipelines that output opaque summaries, AutoMETA provides a *verifiable reasoning trace* linking every pooled estimate to its page-level provenance. As illustrated in Figure 3, every correction is logged with explicit JSON-level provenance and document anchors, enabling full replayability of the synthesis process. This property has direct implications for evidence-based decision systems, where epistemic accountability—the ability to justify *why* an estimate was produced—is as crucial as numerical accuracy.

5.4 Limitations

Several limitations warrant discussion. First, the corpus is restricted to eight cardiology meta-analyses from a single publication year; generalization to broader biomedical or non-medical domains remains to be validated. Second, the wide intervals in Table 1 (e.g., $\Delta_{\text{rel}}(\%)^{\text{med}}$ IQR of [3.9, 45.0] for Multi Full) reflect substantial variability across papers, indicating that performance is not yet uniformly reliable. Third, the accuracy–coverage trade-off identified here suggests that the current binary inclusion mechanism (accept/reject) may be suboptimal; graduated confidence scoring could improve both axes simultaneously. Fourth, the gap between mean and median $\Delta_{\text{rel}}(\%)$ (40.0% vs. 6.4% for Multi Full) indicates that LLM-level extraction errors—particularly unit-conversion mistakes—remain a bottleneck that protocol-level enforcement alone cannot fully address. Fifth, our evaluation uses human-authored meta-analyses as the reference standard, but these are themselves not error-free [27]; deviations from the human reference may occasionally reflect corrections of human errors rather than system failures—a circularity inherent to any benchmarking study in this domain. Finally, AutoMETA assumes that the human-authored protocol is valid and complete; incomplete or conflicting protocols could reintroduce bias in real-world applications.

5.5 Broader Implications and Future Work

Beyond meta-analysis, the principles underlying AutoMETA extend naturally to other forms of collective reasoning and model verification. Any domain requiring protocol-faithful synthesis from multiple textual sources—including risk assessment, policy evaluation, and AI benchmark aggregation—can benefit from similar agentic coordination structures. Future work will explore (i) adaptive consensus mechanisms that weight agents by past accuracy, (ii) confidence-weighted inclusion to resolve the accuracy–coverage trade-off, (iii) alternative critique topologies beyond the current receiving-feedback design, (iv) domain transfer experiments to non-medical corpora, and (v) stronger extraction-level safeguards against unit-conversion and scale-mismatch errors. Regarding (iii), the current system revises extractions based on feedback *received* from peers; we also conducted preliminary experiments with a variant in which agents *give* feedback on others' work and reflexively improve their own extractions—a pathway supported by peer assessment research showing that giving feedback yields greater learning gains than receiving it [4, 24]. Full results are available in our repository.[1] More fundamentally, AutoMETA demonstrates that epistemic reliability in autonomous reasoning systems depends on distinguishing *resolvable disagreement* from *inherent evidence opacity*—a principle that may generalize beyond scientific synthesis to any multi-agent inference domain.

6 CONCLUSION

We presented **AutoMETA**, a multi-agent LLM system for autonomous meta-analysis. By framing evidence synthesis as a protocol-grounded coordination task, AutoMETA enables study-centered agents to extract, critique, and reconcile evidence collaboratively.

Quantitatively, AutoMETA achieved a median relative effect-size error of 6.4% across eight cardiology meta-analyses, the lowest among all configurations tested. The consistent ablation ordering—Full (6.4%) < NoCritique (12.4%) < NoProtocol (25.1%) < NoCrit+NoProt (28.1%)—suggests that critique and protocol enforcement each contribute independently to reliability, with the directional trend holding despite wide per-paper confidence intervals. Heterogeneity alignment presents a more nuanced picture: the full system's stricter study filtering improves effect-size accuracy at the cost of elevated $Q/\tau^2/I^2$ deviations, an accuracy–coverage trade-off that ablation analysis makes explicit.

These findings suggest that the reliability of autonomous evidence-synthesis systems arises from the interaction between procedural grounding and cooperative verification, rather than from model scale alone. AutoMETA's critique–revision loop reframes meta-analysis as an iterative, transparent reasoning process among autonomous agents, producing auditable provenance that links every pooled estimate to its source evidence. Beyond meta-analysis, the same principles—protocol adherence, peer critique, and traceable consensus—may extend to other collective reasoning domains. Ultimately, resolving the accuracy–coverage trade-off through confidence-weighted inclusion represents a key next step toward fully autonomous, human-comparable evidence synthesis.

ACKNOWLEDGMENTS

This research was supported by the Basic Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education under Grant 2710004669; and by the Institute of Convergence Science (ICONS), Yonsei University, and (in part) by the Yonsei University Office of Research Affairs through the Yonsei T.R.U.S.T. (Yonsei Transdisciplinary Research Union for a Sustainable Tomorrow) program (Project Y) under Grant project no.: 2025-22-0461; and by the Yonsei Frontier Lab (YFL) Program for Distinguished Overseas Faculty of Yonsei University.

REFERENCES

[1] Jawad Ibn Ahad, Rafeed Mohammad Sultan, Abraham Kaikobad, Fuad Rahman, Mohammad Ruhul Amin, Nabeel Mohammed, and Shafin Rahman. 2024. Empowering meta-analysis: Leveraging large language models for scientific synthesis. In *2024 IEEE International Conference on Big Data (BigData)*. IEEE, 1615–1624.

[2] Michael Borenstein, Larry V Hedges, Julian PT Higgins, and Hannah R Rothstein. 2021. *Introduction to meta-analysis*. John wiley & sons.

[3] Jared M Campbell, Miloslav Klugar, Sandrine Ding, Dennis P Carmody, Sasja J Hakonsen, Yuri T Jadotte, Sarahlouise White, and Zachary Munn. 2015. Diagnostic test accuracy: methods for systematic review and meta-analysis. *JBI Evidence Implementation* 13, 3 (2015), 154–162.

[4] Kwangsu Cho and Charles MacArthur. 2011. Learning by reviewing. *Journal of educational psychology* 103, 1 (2011), 73.

[5] Jacqueline Davis, Kerrie Mengersen, Sarah Bennett, and Lorraine Mazerolle. 2014. Viewing systematic reviews and meta-analysis in social research through different lenses. *SpringerPlus* 3, 1 (2014), 511.

[6] Damien de Mijolla, Wen Yang, Philippa Duckett, Christopher Frye, and Mark Worrall. 2024. Language hooks: a modular framework for augmenting LLM reasoning that decouples tool usage from the model and its prompt. *arXiv preprint arXiv:2412.05967* (2024).

[7] Rebecca DerSimonian and Nan Laird. 1986. Meta-analysis in clinical trials. *Controlled Clinical Trials* 7, 3 (1986), 177–188.

[8] Rebecca DerSimonian and Nan Laird. 2015. Meta-analysis in clinical trials revisited. *Contemporary clinical trials* 45 (2015), 139–145.

[9] Matthias Egger, George Davey Smith, and Andrew N Phillips. 1997. Meta-analysis: principles and procedures. *Bmj* 315, 7121 (1997), 1533–1537.

[10] Meta Fundamental AI Research Diplomacy Team (FAIR)†, Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, et al. 2022. Human-level play in the game of Diplomacy by combining language models with strategic reasoning. *Science* 378, 6624 (2022), 1067–1074.

[11] Alireza Ghafarollahi and Markus J Buehler. 2024. ProtAgents: protein discovery via large language model multi-agent collaborations combining physics and machine learning. *Digital Discovery* 3, 7 (2024), 1389–1409.

[12] Alireza Ghafarollahi and Markus J Buehler. 2025. Sparks: Multi-agent artificial intelligence model discovers protein design principles. *arXiv preprint arXiv:2504.19017* (2025).

[13] Zhibin Gou, Zhihong Shao, Yeyun Gong, yelong shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2024. CRITIC: Large Language Models Can Self-Correct with Tool-Interactive Critiquing. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=Sx038qxjek

[14] T Greco, Alberto Zangrillo, G Biondi-Zoccai, and Giovanni Landoni. 2013. Meta-analysis: pitfalls and hints. *Heart, lung and vessels* 5, 4 (2013), 219.

[15] Anna-Bettina Haidich. 2010. Meta-analysis in medical research. *Hippokratia* 14, Suppl 1 (2010), 29.

[16] Larry V Hedges. 1992. Meta-analysis. *Journal of Educational Statistics* 17, 4 (1992), 279–296.

[17] Julian PT Higgins, Simon G Thompson, Jonathan J Deeks, and Douglas G Altman. 2003. Measuring inconsistency in meta-analyses. *BMJ* 327, 7414 (2003), 557–560.

[18] Cheng-Yu Hsieh, Si-An Chen, Chun-Liang Li, Yasuhisa Fujii, Alexander Ratner, Chen-Yu Lee, Ranjay Krishna, and Tomas Pfister. 2023. Tool documentation enables zero-shot tool-usage with large language models. *arXiv preprint arXiv:2308.00675* (2023).

[19] Christine P Lee, David Porfirio, Xinyu Jessica Wang, Kevin Chenkai Zhao, and Bilge Mutlu. 2025. Veriplan: Integrating formal verification and llms into end-user planning. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–19.

[20] Juneyoung Lee, Kyung Won Kim, Sang Hyun Choi, Jimi Huh, and Seong Ho Park. 2015. Systematic review and meta-analysis of studies evaluating diagnostic test accuracy: a practical review for clinical researchers-part II. Statistical methods of meta-analysis. *Korean journal of radiology* 16, 6 (2015), 1188–1196.

[21] Eric Lehman, Jay DeYoung, Regina Barzilay, and Byron C Wallace. 2019. Inferring Which Medical Treatments Work from Reports of Clinical Trials. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*. 3705–3717.

[22] Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in Neural Information Processing Systems* 36 (2023), 51991–52008.

[23] Rubén López-Nicolás, Daniel Lakens, Jose A López-López, Maria Rubio-Aparicio, Alejandro Sandoval-Lentisco, Carmen López-Ibáñez, Desirée Blázquez-Rincón, and Julio Sánchez-Meca. 2024. Reproducibility of published meta-analyses on clinical-psychological interventions. *Advances in Methods and Practices in Psychological Science* 7, 1 (2024), 25152459231202929.

[24] Kristi Lundstrom and Wendy Baker. 2009. To give is better than to receive: The benefits of peer review to the reviewer's own writing. *Journal of second language writing* 18, 1 (2009), 30–43.

[25] Ronald Luo, Ziya Sastimoglu, Abu Ilius Faisal, and M Jamal Deen. 2024. Evaluating the efficacy of large language models for systematic review and meta-analysis screening. *medrxiv* (2024), 2024–06.

[26] Kangyong Ma. 2025. AI agents in chemical research: GVIM–an intelligent research assistant system. *Digital Discovery* 4, 2 (2025), 355–375.

[27] Esther Maassen, Marcel ALM Van Assen, Michèle B Nuijten, Anton Olsson-Collentine, and Jelte M Wicherts. 2020. Reproducibility of individual effect sizes in meta-analyses in psychology. *PloS one* 15, 5 (2020), e0233107.

[28] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. SELF-REFINE: iterative refinement with self-feedback. In *Proceedings of the 37th International Conference on Neural Information Processing Systems* (New Orleans, LA, USA) *(NIPS '23)*. Curran Associates Inc., Red Hook, NY, USA, Article 2019, 61 pages.

[29] Iain Marshall, Joël Kuiper, byron wallace, Derek, Sebastián Gálvez, Edward Banner, Frank, and Arash Joorabchi. 2022. *ijmarshall/robotreviewer: RobotReviewer v0.7*. https://doi.org/10.5281/zenodo.6855718

[30] Faith W Mutinda, Shuntaro Yada, Shoko Wakamiya, and Eiji Aramaki. 2022. AUTOMETA: automatic meta-analysis system employing natural language processing. In *MEDINFO 2021: One World, One Health–Global Partnership for Digital Innovation*. IOS Press, 612–616.

[31] Md Rizwan Parvez. 2025. Chain of evidences and evidence to generate: Prompting for context grounded and retrieval augmented reasoning. In *Proceedings of the 4th International Workshop on Knowledge-Augmented Methods for Natural Language Processing*. 230–245.

[32] Tim Reason, Emma Benbow, Julia Langham, Andy Gimblett, Sven L Klijn, and Bill Malcolm. 2024. Artificial intelligence to automate network meta-analyses: four case studies to evaluate the potential application of large language models. *PharmacoEconomics-Open* 8, 2 (2024), 205–220.

[33] Johannes B Reitsma, Andreas S Glas, Anne WS Rutjes, Rob Scholten, Patrick MM Bossuyt, and Aeilko H Zwinderman. 2005. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *Journal of Clinical Epidemiology* 58, 10 (2005), 982–990.

[34] Paulo Salem, Robert Sim, Christopher Olsen, Prerit Saxena, Rafael Barcelos, and Yi Ding. 2025. TinyTroupe: An LLM-powered Multiagent Persona Simulation Toolkit. *arXiv preprint arXiv:2507.09788* (2025).

[35] Kristen L Scotti, Sarah Young, Melanie A Gainey, and Haoyong Lan. 2025. Artificial Intelligence and Automation in Evidence Synthesis: An Investigation of Methods Employed in Cochrane, Campbell Collaboration, and Environmental Evidence Reviews. *Cochrane Evidence Synthesis and Methods* 3, 5 (2025), e70046.

[36] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: language agents with verbal reinforcement learning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems* (New Orleans, LA, USA) *(NIPS '23)*. Curran Associates Inc., Red Hook, NY, USA, Article 377, 19 pages.

[37] Teo Susnjak, Peter Hwang, Napoleon Reyes, Andre LC Barczak, Timothy McIntosh, and Surangika Ranathunga. 2025. Automating research synthesis with domain-specific large language model fine-tuning. *ACM Transactions on Knowledge Discovery from Data* 19, 3 (2025), 1–39.

[38] João Pedro Fernandes Torres, Catherine Mulligan, Joaquim Jorge, and Catarina Moreira. 2024. Promptheus: A human-centered pipeline to streamline slrs with llms. *arXiv preprint arXiv:2410.15978* (2024).

[39] Thomas A Trikalinos, Georgia Salanti, Elias Zintzaras, and John PA Ioannidis. 2008. Meta-analysis methods. *Advances in genetics* 60 (2008), 311–334.

[40] Rens van de Schoot, Jeroen de Bruin, Rianne Schram, Parisa Zahedi, Jannie de Boer, Floris Weijdema, Bianca Kramer, Marijn Huijts, Maarten Hoogerwerf, Godfried Ferdinands, Arjen Harkema, Jasper Willemsen, Ying Ma, Oscar Florez-Vargas, El Hadji Bah, Eric Ruijter, Michèle B. Nuijten, Erik W. Augustijn, Lucia Dominguez-Alvarez, Robbie C. M. van Aert, Marcel A. L. M. van Assen, Klaas Sijtsma, and Silvia D. Olabarriaga. 2021. An open source machine learning framework for efficient and transparent systematic reviews. *Nature Machine*

Intelligence 3 (2021), 125–133. https://doi.org/10.1038/s42256-020-00287-7

[41] Shuai Wang, Harrisen Scells, Shengyao Zhuang, Martin Potthast, Bevan Koopman, and Guido Zuccon. 2024. Zero-shot generative large language models for systematic review screening automation. In *European Conference on Information Retrieval*. Springer, 403–420.

[42] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171* (2022).

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

[44] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. 2024. Autogen: Enabling next-gen LLM applications via multi-agent conversations. In *First Conference on Language Modeling*.

[45] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

[46] Hye Sun Yun, David Pogrebitskiy, Iain J Marshall, and Byron C Wallace. 2024. Automatically extracting numerical results from randomized controlled trials with large language models. *arXiv preprint arXiv:2405.01686* (2024).