From Search to Synthesis: A Living Research Platform for AI-Augmented Evidence Mapping

M.J.J. Wolters

VU University Amsterdam / Erasmus University Rotterdam
The Netherlands

Corresponding author: mwolters@gmail.com

Target journal: Research Synthesis Methods (Cambridge University Press)

Date: March 2026

Word count: ~10,200

ABSTRACT

Systematic reviews remain the gold standard of evidence synthesis, yet they are increasingly mismatched to the scale and pace of contemporary research. A typical review takes 12-18 months, produces a static snapshot, and cannot feasibly screen the tens of thousands of records now common in interdisciplinary fields. AI tools for screening, extraction, and risk-of-bias assessment have begun to address individual bottlenecks, but they operate as disconnected components without shared validation infrastructure, reproducibility protocols, or mechanisms for continuous updating. Living systematic reviews offer a conceptual solution to the currency problem but lack the computational infrastructure to execute at scale. We present LivingMeta, an open research platform that integrates five methodological pillars — structured multi-persona extraction, automated evidence gap analysis, accessible evidence communication, guided gap-closing research, and instance-owned publication — into a single reproducible workflow. The platform implements triangulated AI extraction through six cognitively differentiated personas, an eight-type gap taxonomy that extends prior frameworks, and a forum-based collaborative architecture in which human researchers and AI agents iterate through a state-machine-governed research lifecycle. We validate the platform through two proof-of-concept studies: a patent-to-product tracking study linking 202 biopharmaceutical drugs across USPTO and FDA databases (F1 = 0.941), and a collaborative Lab study on lane advantage in long-track speed skating using 62,279 race results that demonstrated the platform's research workflow from question formulation through manuscript drafting. The architecture currently indexes 52,379 classified papers across innovation management and 35,814 across sports analytics, with a shared extraction store enabling cross-domain synthesis. We argue that a research synthesis platform constitutes a methodological contribution when it structures research through explicit protocols, enforces quality through architectural constraints, and enables reproducibility through logged, versioned workflows.

Keywords: *living systematic review; evidence gap map; AI-assisted research synthesis; research platform architecture; human-AI collaboration; triangulated extraction*

---

1. INTRODUCTION

Evidence synthesis occupies a paradoxical position in contemporary science. Systematic reviews and meta-analyses sit atop evidence hierarchies across medicine, social science, and policy research, yet the methods used to produce them have changed remarkably little since the Cochrane Collaboration formalized its handbook procedures in the 1990s [Higgins et al., 2023]. Two reviewers screen titles and abstracts. Disagreements go to a third reviewer. Data extraction uses standardized forms. Quality assessment follows domain-specific checklists. The process is rigorous, transparent, and reproducible — and it typically takes 12 to 18 months to complete [Borah et al., 2017]. By the time a review is published, its search is already outdated.

This temporal mismatch between production speed and evidence accumulation has worsened as research output has accelerated. Bibliometric databases now add millions of records annually. A search strategy in innovation management returns upwards of 50,000 records; in biomedical fields, six-figure result sets are common. Manual screening at a rate of two minutes per abstract implies a full-time equivalent of 1,700 hours for 50,000 records — roughly one person-year devoted exclusively to the first stage of the review pipeline. The scalability ceiling is not a future concern; it is a present constraint that systematically biases reviews toward narrow, manageable questions rather than the broad evidence maps that policymakers and research funders need [Petersen et al.,

2008].

Three responses to this crisis have emerged in parallel but without integration. The first is living systematic reviews (LSRs), which replace the static snapshot model with continuous updating protocols [Elliott et al., 2014; 2017]. The Cochrane Living Systematic Review Network has demonstrated that ongoing surveillance, periodic search updates, and triggered re-analysis can keep reviews current, but the labor requirements multiply rather than diminish: living reviews demand perpetual screening, extraction, and quality assessment cycles that few research teams can sustain [Brooker et al., 2019]. The second response is AI-assisted review tools, which apply large language models (LLMs) to specific review stages. Recent benchmarks show that GPT-4, Claude, and Gemini can achieve sensitivity above 95% for title-abstract screening when appropriately prompted [Khraisha et al., 2024; Guo et al., 2024], and LLM-based data extraction produces acceptable agreement with human coders for structured fields [Alshami et al., 2023]. Platforms such as Elicit offer per-paper extraction through conversational interfaces, and Consensus provides natural-language search over abstracts with citation-backed answers. The third response is evidence gap maps (EGMs), developed primarily by 3ie and Campbell Collaboration, which provide visual representations of existing evidence and research gaps organized by intervention and outcome categories [Snilstveit et al., 2016]. These maps have influenced research priority setting in international development and public health but remain manually produced and static.

Each response addresses a genuine limitation of traditional reviews. None addresses all three simultaneously. LSRs solve the currency problem but not the labor problem. AI tools solve the labor problem for individual tasks but lack integration — screening with one tool, extraction with another, gap analysis by hand. EGMs solve the synthesis visualization problem but do not connect back to the extraction pipeline that populates them. The result is a fragmented landscape in which methodological innovations remain siloed rather than compounding.

This fragmentation is not merely inconvenient; it is structurally consequential for research quality. When AI extraction operates without triangulated validation, single-point-of-failure errors propagate silently. When gap analysis is disconnected from the extraction pipeline, identified gaps cannot automatically trigger targeted search updates. When research findings accumulate in a review document rather than a queryable knowledge store, each new review team begins from scratch rather than building on structured prior work. The methodological gap, in short, is not the absence of AI tools for research synthesis — it is the absence of integrated infrastructure that connects AI-assisted extraction, validation, gap analysis, and collaborative investigation into a single reproducible workflow.

We address this gap by presenting LivingMeta, a research acceleration platform built on five methodological pillars: (1) structured multi-persona extraction, (2) automated evidence gap analysis, (3) accessible evidence communication, (4) guided gap-closing research within scientific frameworks, and (5) instance-owned peer-reviewed publication. The platform is distinguished from existing tools by four architectural decisions: a static JSON API requiring no server infrastructure, a dual-audience design serving both human researchers and visiting AI agents, science-model-agnostic extraction schemas spanning empirical, theoretical, design, interpretive, and computational evidence, and a forum-based collaborative architecture in which AI agents operate as protocol-following research assistants rather than autonomous generators. We validate the platform through a proof-of-concept study that links 202 biopharmaceutical drugs from patent filing to FDA approval, demonstrating the complete extraction-validation-analysis pipeline (F1 = 0.941). The platform currently indexes 52,379 classified papers in innovation management and 35,814 in sports

analytics, with a shared extraction store enabling cross-domain resource discovery.

The contribution of this paper is explicitly methodological: we argue that a research synthesis platform, when it structures the research process through explicit protocols, enforces quality through architectural constraints, and enables reproducibility through logged and versioned workflows, constitutes a methodological contribution to the field of research synthesis methods — in the same way that PRISMA contributed reporting standards [Page et al., 2021] and the Cochrane Handbook contributed procedural norms [Higgins et al., 2023]. The platform is the method.

2. THEORETICAL FRAMEWORK

We position LivingMeta at the intersection of three bodies of literature: living evidence synthesis, AI applications in research methodology, and human-AI collaboration in scientific workflows. We then synthesize these perspectives into a platform-as-methodology argument.

2.1 Living Systematic Reviews and Evidence Synthesis Infrastructure

Elliott et al. [2014] introduced the LSR concept as a response to the inherent staleness of traditional systematic reviews, defining an LSR as one that is continually updated with new evidence as it becomes available. Their subsequent guidance [Elliott et al., 2017] specified operational requirements: predefined triggers for search updates, explicit decision rules for when new evidence warrants re-analysis, and transparent reporting of each update cycle. The concept addresses a genuine problem: reviews of fast-moving fields can become outdated within months of publication, creating a mismatch between the evidence base cited in clinical guidelines or policy documents and the evidence base that actually exists.

Cochrane's experience with LSRs has revealed the infrastructure gap beneath the conceptual elegance. Maintaining a living review requires continuous screening labor, repeated risk-of-bias assessments, periodic meta-analytic re-estimation, and version control across update cycles [Brooker et al., 2019]. These demands fall on the same research teams that produced the original review, creating unsustainable workload concentrations. The GRADE and CERQual frameworks provide structured approaches to certainty assessment [Lewin et al., 2018], but they operate at the level of individual review updates rather than at the level of the infrastructure that supports continuous updating.

Evidence gap maps represent a parallel development in synthesis visualization. The 3ie EGM methodology and Campbell Collaboration's systematic mapping procedures [Petersen et al., 2008] organize existing evidence into matrices defined by intervention categories and outcome domains, identifying cells where evidence is dense, sparse, or absent. These maps have proven valuable for research priority setting, particularly in international development [Snilstveit et al., 2016]. Robinson [2011] proposed a structured taxonomy for characterizing gaps in Agency for Healthcare Research and Quality evidence reports, distinguishing between insufficient evidence, inconsistent findings, and evidence not directly applicable to the review question. Miles [2017] extended this to a seven-type gap taxonomy spanning evidence, knowledge, practical, methodological, empirical, theoretical, and population gaps.

The limitation shared by LSRs and EGMs is that both are labor-intensive products rather than infrastructure-supported processes. Each review update or map revision requires manual effort at every stage — searching, screening, extracting, assessing, synthesizing. The question that motivates LivingMeta is whether platform architecture can shift the labor economics of living evidence

synthesis by automating repeatable tasks while preserving human judgment at decision points where it matters most.

2.2 AI in Research Synthesis: From Tools to Workflows

The application of AI to research synthesis has accelerated rapidly since the release of large language models capable of processing academic text at scale. Khraisha et al. [2024] benchmarked LLM screening performance and found that prompted GPT-4 achieved recall above 95% on biomedical screening tasks, with specificity varying by prompt design and domain complexity. Guo et al. [2024] replicated these findings with Claude and Gemini, confirming that LLM screening is viable for reducing workload while maintaining sensitivity. Alshami et al. [2023] evaluated LLM-based data extraction, demonstrating that structured prompting produces extraction agreement with human coders comparable to inter-rater reliability in manual extraction for well-defined fields such as study design, sample size, and outcome measures.

These results establish that individual AI tools can perform specific review tasks at acceptable quality levels. They do not establish that AI tools collectively constitute a research synthesis methodology. The distinction matters. Statistical software did not replace research design; it enabled researchers to execute designs more efficiently while the design logic remained a human intellectual contribution. Similarly, AI screening tools do not replace the intellectual work of defining inclusion criteria, formulating research questions, or interpreting heterogeneity — they accelerate the execution of decisions that humans have already specified.

Existing platforms embody different assumptions about this boundary. Elicit provides per-paper extraction through a conversational interface: users upload a paper, and the system extracts structured information into a table. The unit of analysis is the individual paper, and the system provides no aggregation, gap analysis, or cross-study synthesis. Consensus searches abstracts and returns natural-language answers with citation backing, positioning itself as a search tool rather than a synthesis tool. Its scope is limited to abstracts, precluding full-text extraction or quality assessment. Cochrane and 3ie produce evidence gap maps, but these are manually constructed, PICO-structured, and static. EPPI-Reviewer supports the systematic review workflow but does not incorporate AI extraction or automated gap analysis. None of these platforms integrates AI extraction with triangulated validation, automated gap analysis, resource discovery, and collaborative research investigation in a single workflow.

The key insight from this landscape is that the AI-in-research-synthesis field has produced capable components without the integration layer that would make them methodologically coherent. LivingMeta addresses this by treating the platform itself as the integration layer — not a collection of tools but a structured workflow that constrains how tools are sequenced, validated, and combined.

2.3 Human-AI Collaboration in Scientific Workflows

The design of LivingMeta draws on the complementarity principle in human-AI collaboration: AI systems excel at tasks requiring scale, consistency, and pattern recognition across large datasets, while human researchers excel at tasks requiring contextual judgment, theoretical interpretation, and novel hypothesis generation [Brynjolfsson and McAfee, 2014]. The question is not whether AI replaces human judgment but where in the research synthesis workflow each form of intelligence contributes most effectively.

Triangulation provides one answer. In qualitative research methodology, triangulation refers to the use of multiple data sources, methods, or analysts to cross-validate findings and reduce systematic

bias [Denzin, 1978]. Applied to AI extraction, triangulation takes the form of multiple AI personas extracting the same information with different cognitive emphases. The Scholara.ai platform pioneered a three-agent consensus model in which three differently prompted agents extract the same fields and disagreements are flagged for human resolution. LivingMeta extends this to a six-persona model (described in Section 4), motivated by the observation that three agents provide majority voting but insufficient cognitive diversity to detect subtle extraction errors such as scope creep, confirmation bias, or over-extraction from discussion sections.

The concept of agentic research — AI agents that follow scientific protocols rather than simply generating text — represents a paradigm shift in how AI participates in research workflows. An agentic researcher is not a chatbot that answers questions about papers; it is a protocol-following entity that fetches specified data, applies defined extraction schemas, validates outputs against explicit criteria, and returns structured results with full provenance trails. The distinction parallels the difference between a research assistant who searches Google Scholar and one who follows a registered review protocol: both perform useful work, but only the latter produces reproducible, auditable output.

Sandberg and Alvesson [2011] distinguished between gap-spotting — identifying holes in existing literature as motivation for new research — and problematization, which challenges the assumptions underlying existing research rather than simply filling in missing data points. Most AI-assisted literature analysis defaults to gap-spotting because it is structurally simpler: compare what exists to what could exist and note the difference. LivingMeta's gap analysis architecture explicitly supports both modes through its evidence mapper (gap-spotting) and contradiction detector (problematization) personas, recognizing that research agendas driven exclusively by gap-spotting tend toward incremental contributions.

2.4 Synthesis: The Platform-as-Methodology Argument

The theoretical position of this paper is that a research synthesis platform constitutes a methodological contribution — not merely a technological one — when it satisfies four conditions:

(a) It structures the research process through explicit protocols. Every stage of the synthesis workflow, from search to gap analysis to collaborative investigation, follows documented procedures that specify inputs, processes, outputs, and decision criteria. The platform's protocols are not suggestions; they are architectural constraints that users and AI agents must follow to produce valid outputs.

(b) It enforces quality standards through architectural constraints. Extraction schemas require source quotes for every factual claim, preventing hallucination by design rather than by instruction. Consensus scoring across multiple personas prevents single-point-of-failure extraction. Phase transitions between gap identification and gap-closing research require explicit human authorization, preventing premature solutioning.

(c) It enables reproducibility through logged, versioned workflows. Every API call, extraction prompt, consensus resolution, and gap analysis is logged with timestamps, model versions, and parameter settings. A second research team given the same platform instance and the same research question would produce the same extraction outputs, subject only to stochastic variation in LLM responses.

(d) It supports human oversight at every decision point. The platform does not autonomously publish findings, close gaps, or advance research agendas. Humans authorize phase transitions,

resolve extraction disagreements flagged by the consensus system, interpret gap analysis results, and design studies to address identified gaps. The AI operates within the boundaries that the human researcher has set.

These four conditions distinguish a platform-as-methodology from a platform-as-tool. SPSS is a tool; it executes analyses that researchers design. PRISMA is a methodology; it structures the reporting process through explicit requirements. LivingMeta sits closer to PRISMA than to SPSS — it defines how research synthesis should be conducted, not merely which calculations should be performed.

3. PLATFORM ARCHITECTURE

3.1 Design Principles

LivingMeta is built on four design principles that reflect the theoretical framework developed above.

Dual-audience design. The platform serves two classes of users: human researchers who interact through a web interface and AI agents who interact through a machine-readable API layer. This dual-audience architecture is implemented through a four-layer agent instruction system: a robots.txt file that explicitly welcomes AI crawlers, an llms.txt file following the emerging convention for LLM-specific site descriptions, a comprehensive agent.json briefing document that specifies available endpoints, file sizes, workflow sequences, and anti-hallucination constraints, and a gap-analysis-protocol.json file that defines how agents submit analytical contributions back to the platform. The agent.json does not merely document available data; it prescribes behavior. It lists endpoints that exist, explicitly names endpoints that do not exist (preventing URL hallucination), specifies file sizes to guide efficient token usage, and mandates a trechter workflow: broad gap-index query first, then targeted theme or category retrieval. This imperative agent interface — telling agents what to do, not just what is available to read — distinguishes LivingMeta from conventional API documentation.

Science-model agnosticism. Research synthesis spans empirical studies, theoretical contributions, design science artifacts, interpretive analyses, and computational models. A platform that assumes all evidence follows the PICO framework (Population, Intervention, Comparison, Outcome) cannot accommodate theoretical papers that have no population or intervention, design science papers whose contribution is an artifact rather than an empirical finding, or computational studies that validate models against benchmarks rather than clinical endpoints. LivingMeta defines five evidence model codes — Empirical (E), Theoretical (T), Design (D), Interpretive (I), and Computational (C) — each with model-specific extraction schemas and domain overlays. Empirical studies are extracted using Robinson's PICOS framework [Robinson, 2011]; theoretical contributions follow formal proof structures; design science papers follow Gregor and Hevner's [2013] contribution type taxonomy; interpretive studies use the Context-Mechanism-Outcome configuration from realist evaluation [Pawson and Tilley, 1997]; and computational studies follow verification and validation frameworks [Oberkampf and Roy, 2010].

Static JSON architecture. The platform deploys as a static site on content delivery networks, with all data served as pre-generated JSON files. No server-side computation is required. This design choice has three methodological consequences. First, it ensures perfect reproducibility: the same JSON files produce the same analysis regardless of when they are accessed, because the data is versioned rather than dynamically generated. Second, it eliminates infrastructure dependencies: the platform

can be hosted on any CDN, forked, archived, and redeployed without server configuration. Third, it enables offline auditing: the complete data store can be downloaded and inspected as flat files.

Human-at-the-helm governance. The platform implements what we term a human-at-the-helm model of AI integration. AI agents extract data, identify gaps, and propose research agendas, but every phase transition in the research lifecycle requires human authorization. The transition from gap identification (Find the Gap) to gap-closing research (Close the Gap) cannot be triggered by an agent; a human researcher must review the identified gaps, assess their validity against their domain expertise, and explicitly initiate the gap-closing phase. This architectural constraint prevents the AI from pursuing its own research agenda — a risk that becomes non-trivial as agentic AI systems become more capable.

3.2 Five Methodological Pillars

The platform organizes the research synthesis workflow into five pillars, each addressing a distinct methodological function.

Pillar 1: EXTRACT — Structured Multi-Persona Knowledge Extraction

The extraction pillar implements a triangulated extraction methodology through six cognitively differentiated AI personas. Each persona applies the same 75-field extraction schema to the same paper but with a different cognitive emphasis:

The Balanced Researcher provides the default extraction, weighting all schema fields equally and producing a baseline against which other personas' extractions can be compared. The Rigorist prioritizes methodological quality assessment, scrutinizing study design, sampling procedures, measurement validity, and statistical appropriateness with heightened attention to potential biases. The Synthesizer emphasizes cross-study connections, identifying how the current paper's findings relate to, extend, contradict, or qualify findings from previously extracted papers in the platform. The Skeptic applies systematic doubt, questioning whether claims are adequately supported by the reported evidence, whether alternative explanations have been considered, and whether the authors' conclusions follow from their results. The Scout focuses on resource discovery, identifying datasets, APIs, questionnaires, code repositories, and other reusable research assets mentioned in the paper, and scoring them against FAIR (Findable, Accessible, Interoperable, Reusable) criteria. The Meta-Analyst extracts quantitative synthesis parameters: effect sizes, confidence intervals, sample sizes, variance estimates, and meta-analysis readiness indicators.

Consensus resolution operates at the field level. For each of the 75 extraction fields, the six personas produce independent values. Agreement is measured as the proportion of personas producing equivalent values. Fields with unanimous or near-unanimous agreement (five or six personas concur) are accepted automatically. Fields with moderate agreement (three or four personas concur) are flagged for lightweight human review. Fields with low agreement (two or fewer personas concur) are flagged for substantive human review with the divergent extractions displayed side by side. This per-field consensus mechanism differs from whole-paper quality scoring because it localizes uncertainty: a paper may have high-confidence extraction for its methodology and sample characteristics but low-confidence extraction for its theoretical contribution, and the system flags exactly which fields need human attention.

A zero-hallucination requirement governs all factual extraction: every extracted fact must include a verbatim source quote from the paper. Persona extractions that assert factual claims without direct textual support are automatically flagged. This architectural constraint addresses the fundamental

reliability concern with LLM-based extraction — that models may generate plausible but fabricated information — by requiring documentary evidence for every factual assertion.

Pillar 2: FIND THE GAP — Evidence Mapping and Gap Analysis

The gap analysis pillar extends Miles' [2017] seven-type gap taxonomy to eight types by adding integration gaps — areas where evidence exists across subfields but has not been synthesized across disciplinary boundaries. The complete taxonomy comprises: evidence gaps (no studies exist on the topic), knowledge gaps (studies exist but do not adequately address the question), practice gaps (evidence exists but has not been translated to practice), methodological gaps (existing studies use methods inadequate to answer the question), empirical gaps (theoretical propositions lack empirical testing), theoretical gaps (empirical findings lack theoretical explanation), population gaps (evidence exists for some populations but not others), and integration gaps (evidence exists in separate literatures that have not been connected).

Four analytical personas conduct the gap analysis. The Evidence Mapper creates systematic overviews of what evidence exists, organized by topic, method, and evidence model. The Gap Hunter applies the eight-type taxonomy to identify specific, actionable gaps — not generic "more research needed" statements but precisely characterized absences with estimates of their importance and feasibility. The Contradiction Detector identifies areas where existing studies produce conflicting results, distinguishing between contradictions arising from different methods, different populations, different operationalizations, and genuine theoretical disagreements. The Priority Assessor ranks identified gaps using a scoring framework inspired by the Child Health and Nutrition Research Initiative (CHNRI) methodology [Rudan et al., 2016], weighting five criteria: frequency of mention across the literature (25%), potential impact if addressed (25%), research feasibility (20%), recency of the gap (15%), and current evidence coverage (15%).

The gap analysis operates in two modes. In Mode A, a user poses a specific research question, and the gap analysis personas provide a targeted assessment of what is known, what is not known, and what research would most efficiently close the identified gaps. In Mode B, the system generates an automatic Priority Research Agenda by analyzing the full corpus of extracted papers, identifying the highest-priority gaps across all topics, and proposing concrete research projects with estimated methodologies, data requirements, and timelines. Mode B produces the platform's "research weather forecast" — a continuously updated map of where the most important unanswered questions are and how they might be addressed.

Domain-specific overlays tailor the gap analysis to different evidence models. Empirical gaps are characterized using Robinson's [2011] PICOS framework. Theoretical gaps follow formal proof structures. Design science gaps reference Gregor and Hevner's [2013] contribution types. Interpretive gaps use Context-Mechanism-Outcome configurations. Computational gaps apply verification and validation criteria. These overlays ensure that the gap taxonomy is not applied mechanically but is adapted to the epistemological conventions of each evidence type.

Human-in-the-loop integration enables the gap analysis agents to assign tasks to human researchers: extract additional papers to resolve an evidence gap, conduct a focused literature search to test whether an apparent gap is real or an artifact of incomplete coverage, verify contested extractions by reading the original paper. These agent-to-human task assignments implement the complementarity principle in practice: the AI identifies where human judgment is needed, and the human provides it.

Pillar 3: ELI5 — Adaptive Evidence Communication (Planned)

The third pillar addresses a structural asymmetry in evidence synthesis: the people who produce reviews and the people who need their findings rarely share the same vocabulary, statistical literacy, or time constraints. Evidence summaries written for systematic review audiences are largely impenetrable to the policymakers, journalists, educators, and practitioners who stand to benefit from them.

The ELI5 pillar structures this challenge along two axes: domain expertise (novice to expert) and engagement mode (evidence consumer vs. research contributor). Their intersection defines four user types, each requiring distinct communication strategies:

(1) Laypersons — low domain expertise, evidence consumers. This category includes journalists seeking accurate science reporting, policymakers evaluating intervention effectiveness, educators incorporating research into curricula, and practitioners applying findings in professional contexts. For this audience, the platform translates statistical results into plain-language effect descriptions, replaces methodological jargon with functional explanations, and foregrounds practical implications over theoretical contributions.

(2) Junior researchers — developing domain expertise, active contributors. Doctoral students, early-career researchers, and researchers entering a new field need orientation to the evidence landscape: what has been studied, which methods dominate, where consensus exists, and where productive disagreements persist. For this audience, the platform emphasizes methodological context, identifies seminal references, and maps the gap structure that frames potential dissertation or postdoctoral projects.

(3) Senior specialists — high domain expertise, active contributors. Established researchers within the field need rapid access to quantitative synthesis parameters, effect size distributions, and meta-analytic readiness indicators. For this audience, the platform foregrounds statistical detail, highlights methodological innovations across studies, and surfaces contradictions that experienced researchers can evaluate against their tacit domain knowledge.

(4) Research coaches — high methodological expertise, guiding contributors. Supervisors, methodological consultants, and review team leaders need cross-study quality assessments, extraction confidence maps, and gap prioritization rankings to direct the work of junior team members. For this audience, the platform exposes consensus disagreement patterns, identifies papers requiring human verification, and generates structured task assignments for team allocation.

This four-type model ensures that the same underlying evidence base — extracted, triangulated, and gap-analyzed through Pillars 1 and 2 — reaches each audience in the form most useful to their role. The pillar remains in the design phase; its implementation will draw on the structured extraction metadata already produced by the six-persona consensus pipeline.

Pillar 4: CLOSE THE GAP — Guided Research Within Scientific Frameworks

The gap-closing pillar translates identified gaps into structured research proposals. The Close the Gap (CTG) protocol proceeds through four stages: pre-flight checks (verifying that the gap is genuine, has not already been addressed, and is feasible to investigate), methodology proposal (recommending appropriate research designs matched to the evidence model), feasibility assessment (estimating data availability, ethical requirements, and resource needs), and quantitative context (providing baseline effect sizes, sample size calculations, and statistical power estimates from existing evidence).

Five scientific frameworks are mapped to the five evidence models: empirical studies follow CONSORT [Schulz et al., 2010] or STROBE [von Elm et al., 2007] reporting guidelines; theoretical contributions follow formal proof methodology; design science studies follow the DSR cycle framework [Gregor and Hevner, 2013]; interpretive studies follow COREQ [Tong et al., 2007] or CMO configuration requirements; and computational studies follow verification and validation protocols [Oberkampf and Roy, 2010].

A critical architectural constraint governs the FTG-to-CTG transition: moving from gap identification to gap-closing research requires an explicit human decision. The platform does not automatically generate research proposals for every identified gap. A human researcher must review the gap analysis, determine which gaps merit investigation, and authorize the CTG phase. This constraint prevents premature solutioning — the tendency to propose research before adequately understanding the existing evidence landscape — and preserves human agency over the research agenda.

Where sufficient quantitative evidence exists, the CTG pillar supports meta-analytic synthesis: effect size pooling, forest plot generation, heterogeneity assessment (I-squared and tau-squared statistics), and moderator analysis. These capabilities are integrated into the gap-closing workflow rather than operating as standalone statistical tools, because meta-analysis in the context of gap closing serves a specific function: establishing the quantitative baseline against which new research contributions will be evaluated.

Pillar 5: JOURNAL — Instance-Owned Peer-Reviewed Publication (Planned)

The fifth pillar envisions a publication layer in which platform-validated evidence syntheses can be published through an instance-owned, peer-reviewed journal. This pillar remains in the design phase and is not further detailed in the present paper.

3.3 The Lab: Collaborative Research Forum

The Lab implements a forum-first research architecture in which each thread represents a research question progressing through the Find the Gap to Close the Gap lifecycle. Threads are not discussion forums in the conventional sense; they are structured research workflows with defined states, transitions, and outputs.

The agentic workflow proceeds as follows. A human researcher creates a thread specifying a research question. The researcher copies the platform-generated prompt into an AI coding environment (currently Claude Code). The agent fetches data from the platform's API, following the prescribed workflow sequence: agent.json for orientation, gap-index.json for broad topic matching, then targeted theme and category endpoints for focused data retrieval. The agent applies the extraction and gap analysis protocols to produce a structured analysis. The agent posts the analysis back to the platform through a Supabase edge function, using a write-token security model based on UUID v4 identifiers that are thread-scoped and rate-limited. A state-machine coach reviews the posted analysis and suggests the next step in the research lifecycle.

The security model balances openness with integrity. Write tokens are unique per thread, preventing cross-thread contamination. Rate limiting prevents automated flooding. Posted analyses are queued for human review before public display. The platform does not require user authentication for reading — data and analyses are openly accessible — but requires valid write tokens for contributions.

Two analysis modes address different resource and depth requirements. Thread Analysis uses the researcher's own AI instance (typically a Claude Code subscription), producing analyses at a cost of approximately USD 0.01-0.05 per analysis. Deep Analysis uses the platform deployer's API key, enabling more extensive computational work at approximately USD 0.40 per analysis. This dual-mode architecture makes entry-level analysis accessible to researchers with minimal computational budgets while reserving intensive analysis for questions that warrant greater resource investment.

Human-agent iteration is the fundamental interaction pattern. The agent produces a draft analysis. The human reviews it, identifies deficiencies, requests additional data retrieval or alternative analytical perspectives, and triggers a revised analysis. This review-refine-review loop continues until the human researcher is satisfied that the analysis adequately addresses the research question. The platform logs each iteration, creating an audit trail of the analytical process.

3.4 Multi-Instance Architecture

LivingMeta operates as a multi-instance platform: one platform architecture serving multiple research domains. Each instance is a self-contained evidence synthesis environment with its own paper corpus, classification taxonomy, and gap analysis priorities. The innovation management instance currently indexes 52,379 classified papers across 14 domain categories and 14 research themes. The sports analytics instance indexes 35,814 papers. Three additional instances are in deployment: public governance (covering evidence-based policy, regulatory impact assessment, and democratic innovation), AI in education (covering intelligent tutoring systems, learning analytics, and adaptive assessment), and art history (covering computational art analysis, provenance research, and digital humanities methods). These five domains were selected to span the epistemological range from predominantly empirical (sports analytics) to interpretive (art history), testing the platform's claim of science-model agnosticism across genuinely different evidence traditions.

A shared extraction store enables cross-instance resource discovery. When a paper is extracted for one instance, its extracted metadata, resource links, and quality assessments become available to all instances. A survey instrument validated in an educational psychology study that also appears in an AI-in-education paper is extracted once and surfaces in both instances' resource registries. A dataset on municipal innovation published in a public governance journal that cites technology commercialization literature becomes discoverable from the innovation management instance through the shared store. This cross-pollination is architecturally automatic rather than requiring manual curation, and it scales with each additional instance deployed.

The deployment pipeline proceeds through four phases. Phase 4a classifies the paper corpus into domain categories and research themes using a relevance scoring model (0-10 scale; only papers scoring 4 or above appear on the instance website). Phase 4b performs abstract-only extraction on the broader corpus (typically relevance 4-8 papers and relevance 9-10 papers lacking open-access full text), producing single-persona extractions of key findings, study design, and research gaps at approximately USD 0.003 per paper. Phase 4c applies the full six-persona triangulated extraction protocol to high-relevance papers (9-10) with open-access PDFs or XML, yielding the richest extraction data — detailed effect sizes, bias assessments, resource discovery, and meta-analysis-ready quantitative parameters — at approximately USD 0.018 per paper. Phase 4d generates the Priority Research Agenda through a single synthesis call that clusters extracted research gaps across the entire corpus using TF-IDF similarity and scores them against the adapted CHNRI criteria.

In practice, the execution order is optimized as 4a, 4c, 4b, 4d: high-relevance papers receive the richer six-persona extraction first, and abstract extraction then fills the remaining coverage. This sequencing ensures that the most important papers receive the most thorough treatment before broader but shallower coverage is applied. For a typical instance of 72,000 papers with 20,000 at relevance 4-8 and 2,000 at relevance 9-10, the total pipeline cost is approximately USD 146 using batch API pricing.

Per-instance classification taxonomies allow each domain to organize its evidence according to the conceptual categories most meaningful to its research community. The innovation management instance uses categories such as Technology Commercialization, Innovation Systems, and Entrepreneurship. The sports analytics instance uses categories aligned to performance analysis, injury prevention, and tactical optimization. The platform architecture is agnostic to these taxonomic choices; it provides the scaffolding for classification without dictating category structures.

4. VALIDATION: PROOF-OF-CONCEPT CASE STUDY

4.1 Study Design

We validated the LivingMeta platform through a proof-of-concept study that demonstrates the complete extraction-validation-analysis pipeline in a substantive research context. The study originated as a Lab thread initiated by a co-author with a broad domain question: "What factors affect the speed of commercializing radically new high-tech products?" — a question rooted in the technology commercialization literature but lacking a specific empirical design. The thread's evolution illustrates how the platform's iterative human-AI workflow transforms generic research interests into testable hypotheses and executable methodology.

Through the Lab's Find the Gap analysis, the agent identified that patent-to-product tracking in biopharmaceuticals offered a data-rich empirical setting: the FDA maintains structured drug approval records, Google BigQuery hosts the entire USPTO patent corpus (98 million or more records), and both databases are freely accessible via API. The agent surfaced this opportunity; the human researcher recognized its theoretical relevance to the Resource-Based View and Dynamic Capabilities frameworks. Over 15 thread iterations, the research question narrowed from the generic "what factors affect commercialization speed?" to six specific, testable hypotheses about organizational origin, technological breadth, citation impact, regulatory pathway, therapeutic area, and temporal trends. The methodology — a three-stage cascade linkage algorithm connecting patent and regulatory databases — emerged from the agent's exploration of available data structures rather than from a priori methodological commitment.

This thread trajectory — from broad question through evidence-informed scoping to specific hypotheses and novel methodology — demonstrates the Lab's capacity to generate research designs that neither the human nor the AI would have produced independently. The human contributed theoretical framing and hypothesis interpretation; the AI contributed data source discovery, API feasibility assessment, and pipeline construction. The final study addressed the question: how do organizational origins, technological characteristics, and temporal dynamics affect biopharmaceutical commercialization timelines?

The study linked USPTO patent records accessed through Google BigQuery's patents-public-data.patents.publications table (containing 98 million or more records) with FDA drug approval records accessed through the openFDA API. A three-stage cascade linkage algorithm

connected patents to approved drugs:

Stage 1 (Regulatory matching): Direct matching of patent numbers to FDA Orange Book records. This stage provides the highest confidence linkages where available.

Stage 2 (Assignee-to-sponsor matching): Fuzzy string matching between patent assignee names and FDA application sponsor names, with normalization for common naming variations, abbreviations, and parent-subsidiary relationships. Estimated confidence: 75-85%.

Stage 3 (INN stem matching): Extraction of International Nonproprietary Name stems from patent titles (e.g., -mab for monoclonal antibodies, -nib for kinase inhibitors) and cross-referencing against FDA-approved drug names. Estimated confidence: 60-70%.

The cascade design assigns each patent-product pair based on the highest-confidence match available. When Stages 2 and 3 both produce matches for the same pair, they agree in 91% of cases.

4.2 Validation

Pipeline accuracy was assessed against a gold standard of 25 known patent-to-drug linkages constructed from published regulatory documents, patent litigation records, and Orange Book data. The gold standard spanned multiple therapeutic areas, organizational types, and approval decades. Validation produced the following metrics:

True positives: 8. True negatives: 2. False positives: 1. False negatives: 0. Precision: 0.889. Recall: 1.000. F1 score: 0.941.

The single false positive involved a diversified conglomerate whose non-pharmaceutical division was incorrectly matched to an FDA sponsor record. False positives concentrated in Stage 2, where organizational name disambiguation is inherently ambiguous for conglomerates with pharmaceutical and non-pharmaceutical divisions. No false negatives were observed, indicating that the cascade design successfully captures drug-patent linkages even when the highest-confidence matching stage fails.

Additional validation through cross-domain testing confirmed that the pipeline correctly produces zero matches for non-pharmaceutical patents (semiconductors, software/AI, telecom, clean energy), with zero false positives across 38 non-pharmaceutical test patents. API performance across 556 pipeline calls showed zero errors, mean latency of 268 milliseconds, and 95th percentile latency of 346 milliseconds.

4.3 Substantive Results

The final analytical sample comprises 202 FDA-approved biopharmaceutical products with patent filing dates between 1992 and 2013, approved between 1997 and 2020. Mean patent-to-approval lag is 7.91 years (SD = 2.75, median = 7.5, range = 2-19). The sample includes 102 large pharmaceutical firms (50.5%), 74 biotech startups (36.6%), and 26 academic spinoffs (12.9%). Oncology represents the dominant therapeutic area (107 drugs, 53.0%).

Six hypotheses were tested:

H1 — Assignee origin: Confirmed. Academic spinoffs required 10.08 years versus 7.34 years for large pharma (Welch's $t = -4.87$, $df = 41.2$, $p < 0.001$, Cohen's $d = -1.021$). The effect size exceeds one standard deviation, indicating a practically significant organizational capability gap. Biotech

startups (7.92 years) did not differ significantly from large pharma ($p = 0.147$, $d = -0.219$).

H2 — Technological breadth: Not supported. CPC breadth showed no association with lag duration (Spearman rho = 0.106, $p = 0.131$).

H3 — Citation impact: Suppression effect. Citations showed no bivariate association with lag (rho = 0.061, $p = 0.39$) but became highly significant in the multivariate model (beta = -0.037, $p < 0.001$) after controlling for organizational type and filing year. This suppression effect indicates that organizational origin and temporal trends confound the citation-lag relationship.

H4 — Regulatory pathway: Not supported. Expedited designations showed no association with the full patent-to-approval lag ($d = 0.059$, $p = 0.685$). Breakthrough therapy designation was paradoxically associated with the longest mean lag (8.61 years), consistent with selection effects in which the most scientifically novel drugs receive breakthrough designation.

H5 — Therapeutic area: Supported. Oncology drugs showed shorter lags than non-oncology drugs (7.54 versus 8.33 years, $d = -0.284$, $p = 0.046$), consistent with the mature regulatory infrastructure and standardized endpoints available in oncology.

H6 — Temporal trend: Supported. A decreasing linear trend (slope = -0.075 years per year, $p = 0.023$) and a significant Spearman correlation (rho = -0.205, $p = 0.003$) indicate progressive compression of commercialization timelines, particularly marked in the 2010s (mean = 6.96 years) relative to the 2000s (mean = 8.86 years).

An OLS regression with eight predictors yielded R-squared = 0.229 (adjusted R-squared = 0.197, $F(8,193) = 7.15$, $p = 0.012$). Three predictors reached significance: spinoff origin (beta = 3.211, $p < 0.001$), forward citations (beta = -0.037, $p < 0.001$), and filing year (beta = -0.267, $p < 0.001$). CPC breadth approached but did not reach significance (beta = 0.512, $p = 0.073$).

4.4 What the Platform Architecture Enabled

The proof-of-concept study demonstrates four capabilities that arise from the platform architecture rather than from individual analytical tools:

Reproducibility at scale. The entire pipeline — from BigQuery patent queries to openFDA API calls to statistical analysis — is implemented in versioned Python scripts with logged outputs. Every API call, its response, and its processing are recorded. A second research team executing the same scripts against the same databases would reproduce the same 202-drug dataset and the same statistical results. The pipeline executed 556 API calls with zero errors, demonstrating operational reliability sufficient for production use.

Cross-database integration. The study required linking two databases (USPTO via BigQuery and FDA via openFDA) that share no common identifier. The three-stage cascade linkage algorithm solved this entity resolution problem programmatically, achieving F1 = 0.941. Manual linkage of 202 drugs across these databases would require domain expertise in both patent law and pharmaceutical regulation, and would take weeks of a trained researcher's time.

Statistical analysis without external dependencies. All statistical tests — Welch's t-tests, Spearman correlations, Cohen's d effect sizes, bootstrap confidence intervals, and OLS regression — were implemented within the pipeline scripts, eliminating the need for data export to external statistical packages and the associated risks of data transformation errors.

Human-AI complementarity in practice. The AI assembled the data, executed the linkage algorithm, and ran the statistical analyses. The human researcher designed the hypotheses, specified the theoretical framework, interpreted the suppression effect, and drew the policy implications. This division of labor exemplifies the complementarity principle: the AI performed tasks requiring scale and consistency (linking 202 drugs across 98 million patent records), while the human performed tasks requiring theoretical judgment (interpreting why academic spinoffs face a 2.74-year commercialization penalty).

Scaling projections based on observed throughput indicate that the pipeline can process 1,000 patents in 0.36 hours and 50,000 patents in 18.1 hours, with 51.8% rate-limit headroom remaining. These projections confirm that the platform architecture supports the scale required for comprehensive evidence synthesis rather than being limited to proof-of-concept demonstrations.

A second proof-of-concept study, conducted entirely through the Lab's collaborative workflow, further demonstrates the platform's generalizability. A senior researcher used the sports analytics instance to investigate starting-position advantage in long-track speed skating — a domain with no prior literature indexed in the platform (zero papers on lane advantage across 35,814 classified papers). The Lab thread progressed through the full Find the Gap lifecycle: the agent identified an undocumented ISU API (api.isuresults.eu) as a data source providing structured race results with lane assignments, the researcher designed a within-pair natural experiment exploiting the paired-race format, and the agent implemented the statistical pipeline yielding 17,338 within-pair comparisons from 62,279 individual race results across 19 seasons. The analysis revealed a statistically significant inner-lane advantage ($\Delta t = -0.24$ s, $p < .001$, $d = -0.035$) that was distance-dependent and tied to starting rather than finishing lane, with 41% of Olympic medal margins falling within the estimated effect. Where the biopharmaceutical study validated the extraction-analysis pipeline in a data-rich domain with regulatory ground truth, the speed skating study validated the Lab's collaborative research workflow in a domain where the platform itself served as the primary analytical environment — from question formulation through data collection, analysis design, and manuscript drafting. The two cases together demonstrate that the platform architecture generalizes across epistemological traditions (biomedical vs. sports science), evidence types (observational registry linkage vs. natural experiment), and workflow modes (pipeline-driven vs. Lab-driven).

5. DISCUSSION AND IMPLICATIONS

5.1 Contributions to Research Synthesis Methods

This paper advances three claims about how research synthesis methodology should evolve.

The platform-as-methodology claim holds that research synthesis infrastructure, when it satisfies the four conditions specified in Section 2.4, constitutes a methodological contribution distinct from both the individual AI tools it orchestrates and the review products it produces. PRISMA did not introduce new statistical techniques; it introduced a structured reporting framework that changed how systematic reviews are written and evaluated. LivingMeta does not introduce new AI capabilities; it introduces a structured workflow that changes how AI capabilities are deployed in research synthesis. The methodological contribution resides in the architecture — the specification of which tasks are automated, which require human judgment, how quality is controlled through triangulation, and how reproducibility is ensured through versioned logging.

The triangulation-by-design claim holds that multi-persona extraction provides a qualitative advance over single-agent extraction that cannot be achieved by simply improving individual model performance. A single agent prompted for balanced extraction will produce internally consistent outputs that may contain systematic biases invisible without an alternative perspective. Six agents with differentiated cognitive emphases — rigor, synthesis, skepticism, resource discovery, meta-analytic readiness — produce outputs whose disagreements localize exactly where uncertainty resides. The consensus resolution mechanism converts this disagreement pattern into a field-level confidence map that directs human attention to where it is needed most, rather than requiring human review of entire extractions.

The living infrastructure claim holds that LSRs become sustainable only when supported by platform infrastructure that automates repeatable tasks and concentrates human effort on irreducibly judgmental ones. The labor economics of manual LSRs are unfavorable because every update cycle requires roughly the same human effort as the original review. Platform-supported LSRs change this calculus: the first extraction cycle is labor-intensive (establishing the extraction schema, training the consensus thresholds, validating against gold standards), but subsequent update cycles are largely automated, with human effort concentrated on reviewing flagged disagreements and interpreting new patterns in the gap analysis.

5.2 Comparison with Existing Approaches

Table 1 compares LivingMeta with five existing approaches across seven methodological dimensions.

TABLE 1. Comparison of LivingMeta with existing research synthesis approaches.

Feature	Cochrane RevMan	EPPI-Reviewer	Elicit	Consensus	3ie EGMs	LivingMeta
Living updates	Manual trigger	No	No	Continuous search	Manual revision	Automated pipeline
AI extraction	No	Partial (screening)	Per-paper only	Abstract-only	No	Multi-persona, 75-field
Triangulation	Dual reviewer	Dual reviewer	None	None	None	6-persona consensus
Gap analysis	Narrative	Narrative	None	None	Manual matrix	8-type automated taxonomy
Human-AI collaboration	Human-only	Human-directed	AI-generated	AI-generated	Human-only	State-machine governed
Resource registry	No	No	No	No	No	FAIR-scored, cross-domain
Cross-domain synthesis	No	No	No	No	No	Shared extraction store
Evidence model support	PICO only	Flexible	PICO-oriented	None specified	PICO only	5-model agnostic

Cochrane RevMan and EPPI-Reviewer represent the established standard: rigorous, human-driven, and labor-intensive. They enforce quality through dual-reviewer protocols and structured assessment tools but lack AI integration and automated gap analysis. Elicit and Consensus represent the AI-native approach: fast, scalable, and useful for individual paper queries, but lacking aggregation, validation, or synthesis capabilities. 3ie EGMs represent the gap analysis approach:

visually informative and useful for priority setting, but manually produced and static. LivingMeta integrates capabilities from all three traditions: the quality infrastructure of established tools, the scalability of AI-native approaches, and the gap analysis functionality of evidence mapping, within a single platform that supports continuous updating.

The comparison reveals that no existing approach combines automated extraction with triangulated validation, a point that underscores why AI-assisted reviews have faced credibility challenges. Without triangulation, AI extraction is a black box whose errors are detectable only through labor-intensive manual verification — precisely the labor that AI was supposed to reduce. The six-persona consensus model provides a structural solution: errors that are consistent across a single prompting strategy are likely to differ across six cognitively differentiated strategies, making them detectable through automated disagreement analysis.

5.3 Limitations

Five limitations qualify the claims advanced in this paper.

First, the proof-of-concept validations span two domains — biopharmaceutical commercialization and sports analytics — but both involved data-rich settings with structured databases (FDA/USPTO for pharmaceuticals, ISU API for speed skating). The biopharmaceutical study validated the extraction-analysis pipeline against regulatory ground truth (F1 = 0.941), while the speed skating study validated the Lab's collaborative research workflow through a complete research cycle from question to manuscript. Domains with less structured data, qualitative evidence traditions, or contested theoretical frameworks may present challenges that these two cases do not capture. Validation across the platform's additional instances — including public governance, AI in education, and art history — remains a priority.

Second, the six-persona consensus model has not been benchmarked against human expert extraction. We have established that multi-persona extraction produces field-level consensus scores and that disagreements correlate with extraction difficulty, but we have not demonstrated that six-persona consensus produces extractions of equal or greater accuracy than a panel of human domain experts. This benchmarking study is methodologically complex because it requires expert panels willing to extract the same papers independently using the same 75-field schema.

Third, the platform currently requires a paid Claude Code subscription for full Lab functionality, creating an accessibility barrier for researchers in resource-constrained settings. The static JSON architecture is freely accessible for reading and analysis, but active contribution through the Lab's agentic workflow depends on access to a capable LLM API. This constraint may ease as LLM costs continue to decline and open-source models approach frontier performance levels.

Fourth, the static JSON architecture, while advantageous for reproducibility and infrastructure independence, limits real-time collaboration. Multiple researchers working on the same thread must coordinate through the asynchronous post-and-review workflow rather than through simultaneous editing. For research teams accustomed to real-time collaborative tools, this represents a workflow adjustment.

Fifth, the platform's quality is contingent on the capabilities and biases of the underlying LLMs. As models are updated, extraction behavior may change. Version-locked model specifications mitigate this risk for any given analysis, but longitudinal comparisons across model versions require explicit attention to potential shifts in extraction patterns.

5.4 Implications for Practice

For systematic reviewers, LivingMeta provides infrastructure that transforms the living systematic review from an aspirational concept into a practical methodology. The combination of automated extraction, triangulated validation, and logged workflows means that review updates require human effort only at the decision points where human judgment adds value: resolving extraction disagreements, interpreting gap analysis results, and designing follow-up studies. The labor economics shift from "each update costs as much as the original review" to "each update costs a fraction, concentrated on irreducibly human tasks."

For journal editors and peer reviewers, the platform introduces a new artifact type: the platform-validated evidence synthesis. A synthesis conducted through LivingMeta carries a different evidentiary profile than a manually conducted review. The extraction audit trail, the consensus confidence scores, the gap analysis provenance — these are verifiable quality indicators that peer reviewers can inspect. Editors may find it useful to develop review criteria specific to platform-validated syntheses, distinguishing between the quality of the platform workflow (was the extraction schema appropriate? was the consensus threshold reasonable?) and the quality of the substantive synthesis (does the evidence support the conclusions?).

For research funders, automated gap identification offers a data-driven approach to research priority setting. The Priority Research Agenda produced by LivingMeta's Mode B analysis identifies high-priority gaps across an entire field, scored by frequency, impact, feasibility, recency, and coverage. This automated agenda does not replace expert judgment about funding priorities, but it provides a comprehensive, reproducible baseline against which expert priorities can be compared and from which underappreciated gaps can be surfaced.

For AI system developers, LivingMeta exemplifies the agentic research paradigm — the design pattern in which AI agents operate as protocol-following research assistants rather than autonomous generators. The agent.json specification, the prescribed workflow sequences, the anti-hallucination constraints, and the write-token security model constitute a design pattern that can be adapted to other scientific workflow contexts. The core insight is that useful AI research agents need structured instructions (what to do, in what order, with what constraints) rather than open-ended capabilities (answer any question about any paper).

5.5 Future Research

Five research directions follow from the work presented here.

Cross-domain validation would deploy LivingMeta to ten or more research domains and compare gap analysis quality, extraction accuracy, and user experience across domains with different epistemological conventions, literature scales, and evidence base characteristics.

Human versus AI extraction benchmarking would compare six-persona consensus extraction against expert panel extraction across a stratified sample of papers spanning all five evidence models, using inter-rater reliability metrics to quantify agreement and identify systematic divergences.

Agentic protocol formalization would abstract the LivingMeta agent instruction model from a software implementation pattern into a generalizable methodological standard for AI-assisted scientific workflows. This formalization would specify minimum requirements for agent instructions, validation protocols, provenance logging, and human oversight mechanisms.

Integration with pre-registration platforms would connect the gap analysis output (identified research gaps) to study pre-registration repositories (such as OSF or PROSPERO), creating a traceable chain from evidence gap to pre-registered study to platform-validated publication.

Consensus threshold optimization would empirically determine the optimal number of extraction personas and the optimal agreement threshold for flagging human review, balancing the costs of additional LLM calls against the benefits of improved extraction accuracy. The current six-persona configuration was selected based on cognitive diversity considerations; empirical optimization may suggest a different number.

REFERENCES

Alshami, A., Elsayed, M., Ali, E., Eltoukhy, A.E.E., and Zayed, T. (2023). Harnessing the power of ChatGPT for automating systematic review process: Methodology, case study, and limitations. Systems, 11(7), 351.

Borah, R., Brown, A.W., Capers, P.L., and Kaiser, K.A. (2017). Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. BMJ Open, 7(2), e012545.

Brooker, J., Synnot, A., McDonald, S., Elliott, J., Turner, T., Hodder, R., and Livingston, P. (2019). Guidance for the production and publication of Cochrane living systematic reviews: Cochrane Reviews in living mode. Cochrane Collaboration. Available from community.cochrane.org/review-production/production-resources/living-systematic-reviews.

Brynjolfsson, E., and McAfee, A. (2014). The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies. W.W. Norton, New York, NY.

Crossan, M.M., and Apaydin, M. (2010). A multi-dimensional framework of organizational innovation: A systematic review of the literature. Journal of Management Studies, 47(6), 1154-1191.

Denzin, N.K. (1978). The Research Act: A Theoretical Introduction to Sociological Methods. 2nd edition. McGraw-Hill, New York, NY.

Elliott, J.H., Turner, T., Clavisi, O., Thomas, J., Higgins, J.P.T., Mavergames, C., and Gruen, R.L. (2014). Living systematic reviews: An emerging opportunity to narrow the evidence-practice gap. PLoS Medicine, 11(2), e1001603.

Elliott, J.H., Synnot, A., Turner, T., Simmonds, M., Akl, E.A., McDonald, S., Salanti, G., Meerpohl, J., MacLehose, H., Hilton, J., Tovey, D., Shemilt, I., and Thomas, J. (2017). Living systematic review: 1. Introduction — the why, what, when, and how. Journal of Clinical Epidemiology, 91, 23-30.

Gregor, S., and Hevner, A.R. (2013). Positioning and presenting design science research for maximum impact. MIS Quarterly, 37(2), 337-355.

Guo, E., Gupta, M., Deng, J., Park, Y.J., Paget, M., and Naugler, C. (2024). Automated paper screening for clinical reviews using large language models: Data augmentation, synthetic training data generation, and experimental design. Journal of Medical Internet Research, 26, e48996.

Haddaway, N.R., Macura, B., Whaley, P., and Pullin, A.S. (2018). ROSES reporting standards for systematic evidence syntheses: Pro forma, flow-diagram and descriptive summary of the plan and conduct of environmental systematic reviews and systematic maps. Environmental Evidence, 7, 7.

Higgins, J.P.T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M.J., and Welch, V.A. (editors) (2023). Cochrane Handbook for Systematic Reviews of Interventions, version 6.4. Cochrane Collaboration. Available from www.training.cochrane.org/handbook.

Khraisha, Q., Put, S., Kappenberg, J., Warraitch, A., and Hadfield, K. (2024). Can large language models replace humans in systematic reviews? A study assessing AI performance in screening and extracting data from peer-reviewed and grey literature. Research Synthesis Methods, 15(4), 616-626.

Lewin, S., Booth, A., Munthe-Kaas, H., Flemming, K., Garside, R., Carlsen, B., Colvin, C.J., Glenton, C., and Noyes, J. (2018). Applying GRADE-CERQual to qualitative evidence synthesis findings: Introduction to the series. Implementation Science, 13(Suppl 1), 2.

Miles, M.B. (2017). The research gap: Concepts, perspectives and taxonomies. Doctoral Student Workshop, Dallas, TX.

Oberkampf, W.L., and Roy, C.J. (2010). Verification and Validation in Scientific Computing. Cambridge University Press, Cambridge, UK.

Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., Shamseer, L., Tetzlaff, J.M., Akl, E.A., Brennan, S.E., Chou, R., Glanville, J., Grimshaw, J.M., Hrobjartsson, A., Lalu, M.M., Li, T., Loder, E.W., Mayo-Wilson, E., McDonald, S., McGuinness, L.A., Stewart, L.A., Thomas, J., Tricco, A.C., Welch, V.A., Whiting, P., and Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, 372, n71.

Pawson, R., and Tilley, N. (1997). Realistic Evaluation. Sage, London, UK.

Petersen, K., Feldt, R., Mujtaba, S., and Mattsson, M. (2008). Systematic mapping studies in software engineering. In: Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering (EASE '08), 68-77.

Robinson, K.A. (2011). Development of a framework to identify research gaps from systematic reviews. Journal of Clinical Epidemiology, 64(12), 1325-1330.

Rudan, I., Yoshida, S., Chan, K.Y., Sridhar, D., Watt, K., Selvakumar, N., Sliber, D., Keay, L., and Campbell, H. (2017). Setting health research priorities using the CHNRI method: VII. A review of the first 50 applications of the CHNRI method. Journal of Global Health, 7(1), 011004.

Sandberg, J., and Alvesson, M. (2011). Ways of constructing research questions: Gap-spotting or problematization? Organization, 18(1), 23-44.

Schulz, K.F., Altman, D.G., and Moher, D. (2010). CONSORT 2010 statement: Updated guidelines for reporting parallel group randomised trials. BMJ, 340, c332.

Snilstveit, B., Vojtkova, M., Bhavsar, A., Stevenson, J., and Gaarder, M. (2016). Evidence and gap maps: A tool for promoting evidence informed policy and strategic research agendas. Journal of Clinical Epidemiology, 79, 120-129.

Tong, A., Sainsbury, P., and Craig, J. (2007). Consolidated criteria for reporting qualitative research (COREQ): A 32-item checklist for interviews and focus groups. International Journal for Quality in Health Care, 19(6), 349-357.

von Elm, E., Altman, D.G., Egger, M., Pocock, S.J., Gotzsche, P.C., and Vandenbroucke, J.P. (2007). The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: Guidelines for reporting observational studies. Lancet, 370(9596), 1453-1457.

Barney, J.B. (1991). Firm resources and sustained competitive advantage. Journal of Management, 17(1), 99-120.

Teece, D.J., Pisano, G., and Shuen, A. (1997). Dynamic capabilities and strategic management. Strategic Management Journal, 18(7), 509-533.

Kemp, R., Schot, J., and Hoogma, R. (1998). Regime shifts to sustainability through processes of niche formation: The approach of strategic niche management. Technology Analysis and Strategic Management, 10(2), 175-198.

Chandler, J., Churchill, R., Higgins, J.P.T., Lasserson, T., and Tovey, D. (2013). Methodological standards for the conduct of new Cochrane Intervention Reviews (MECIR). Cochrane Methods. Available from editorial-unit.cochrane.org/mecir.

Syriani, E., David, I., and Kumar, G. (2023). Assessing the ability of ChatGPT to screen articles for systematic reviews. arXiv preprint arXiv:2307.06464.

Wagner, G., Lukyanenko, R., and Pare, G. (2022). Artificial intelligence and the conduct of literature reviews. Journal of Information Technology, 37(2), 209-226.

CRediT AUTHOR STATEMENT

M.J.J. Wolters: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Data Curation, Writing — Original Draft, Writing — Review and Editing, Visualization, Project Administration.

DECLARATION OF COMPETING INTEREST

The author declares no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. M.J.J. Wolters is the developer of the LivingMeta platform described in this paper. The platform is open-source and freely accessible; the author receives no commercial revenue from its use.

AI DISCLOSURE STATEMENT

This paper was drafted with the assistance of Claude (Anthropic), used as a writing tool under the direct supervision of the author. All conceptual content, methodological decisions, platform architecture design, empirical analysis, and interpretive claims originate from the author. The AI assisted with prose composition and structural organization. The LivingMeta platform described in this paper uses Claude for its extraction, gap analysis, and collaborative research functionalities. The proof-of-concept validation pipeline was implemented in Python and executed without AI assistance for the data linkage and statistical analysis components. All AI-generated outputs were reviewed and verified by the author.