Puzzles to Paradigms: A Survey of Human–Agent Collaboration in Scientific Innovation

Anonymous ACL submission

Abstract

Agentic AI systems are increasingly embedded in scientific workflows, including literature synthesis, hypothesis generation, tool use, writing, and peer review. However, existing surveys and evaluations often emphasize task performance while underexamining how humans and agents coordinate under different epistemic conditions of scientific work. This survey introduces **CSCW×Kuhn**, a coordination-aware framework that connects computer-supported cooperative work dimensions of articulation work, common ground, and awareness with three scientific contexts: puzzle-solving in normal science, anomaly evaluation, and paradigm-shift exploration. Using this framework, we synthesize recent human–agent systems and benchmarks, showing how coordination demands and failure risks vary across contexts. We identify five recurring failures: procedural drift, grounding substitution, evaluative flattening, proxy optimization, and epistemic black-boxing, and argue for evaluations that move beyond task accuracy toward grounding, provenance, disagreement preservation, and long-horizon scientific reliability.

1 Introduction

Agentic AI systems are increasingly entering scientific research workflows. Recent systems can retrieve and synthesize literature, generate hypotheses, operate scientific tools, write and revise manuscripts, simulate peer review, coordinate multiple agents, and execute data-analysis or experimental procedures (Wang et al., 2023; Chen et al., 2025a; Wei et al., 2025; Zheng et al., 2025; Wang et al., 2024a; Guo et al., 2024). These developments are changing not only what scientific tasks can be automated, but also how scientific work is organized, monitored, and evaluated. Scientific innovation is rarely a single-step prediction problem. It requires researchers to frame problems, compare evidence, maintain provenance, interpret uncertainty, revise assumptions, and decide when a result is credible or worth pursuing (Kuhn, 2012). As agentic systems become involved in these workflows, the critical question is therefore not only whether agents can successfully complete scientific tasks, but whether humans and agents can coordinate reliably under various epistemic conditions.

This question is important because scientific workflows vary in uncertainty, evidence standards, and evaluation criteria. Some tasks resemble puzzle-solving within established methods, such as screening papers, executing analysis code, or following a protocol. Others involve evaluation of anomalies, where humans and agents must judge uncertain findings, contested claims, novelty, or reviewer disagreement. Still, others support paradigm-shift exploration, where the goal is to search for alternative concepts, distant analogies, or new research directions (Kuhn, 2012). These contexts impose different coordination demands. One system that is adequate for bounded execution may fail when evidence is contested; another that generates plausible hypotheses may still be unreliable if provenance, assumptions, and validation paths remain opaque.

Computer-supported cooperative work (CSCW) offers a useful theoretical foundation for analyzing these challenges. CSCW treats collaboration as a sociotechnical achievement that depends on articulation work, coordination mechanisms, shared understanding, and awareness of ongoing activity (Schmidt and Bannon, 1992; Schmidt and Simone, 1996; Clark and Brennan, 1991; Dourish and Bellotti, 1992; Gutwin and Greenberg, 2002). These concerns are central to scientific workflows, where researchers must align goals, evidence, methods, assumptions, intermediate artifacts, and responsibility across multiple stages.

When agents enter such workflows, they do not merely automate isolated tasks; they become part of distributed coordination arrangements. A CSCW

perspective thus helps explain why task-level performance is insufficient: a system may produce a seemingly correct output while failing to preserve workflow state, evidential grounding, disagreement, provenance, or human accountability.

Existing surveys have made substantial progress in mapping AI systems for science, agentic architectures, and human–AI collaboration. *AI4Science surveys* organize systems along research pipelines, including literature review, hypothesis generation, experimentation, analysis, and writing (Chen et al., 2025a; Zheng et al., 2025; Tie et al., 2026). *Agentic-system surveys* analyze planning, tool use, memory, retrieval, and multi-agent coordination (Wang et al., 2024a; Guo et al., 2024; Tran et al., 2025). *Human–agent collaboration surveys* examine role allocation, agency distribution, interaction patterns, and control with human involvement (Gomez et al., 2024; Vats et al., 2025; Zhang et al., 2025b; Zou et al., 2026). Still, these perspectives do not fully explain how the epistemic context of scientific work changes the coordination demands placed on human–agent systems. As a result, current evaluations often emphasize bounded task success, output correctness, or benchmark performance (Lou et al., 2025; Chen et al., 2025d; Majumder et al., 2025; Wijk et al., 2025), while under-measuring workflow integrity, evidential grounding, disagreement preservation, provenance, and accountability (Xiong et al., 2025; Jin et al., 2024; Russo Latona et al., 2025).

In this survey, we address this gap by introducing **CSCW×Kuhn**, a coordination-aware framework for analyzing human–agent collaboration in scientific innovation. We use the term "human–agent collaboration" to refer to systems in which agents contribute to planning, judgment, execution, or revision in a scientific workflow and where human decisions remain relevant for final scientific output. Drawing on CSCW, we characterize coordination demand through articulation work, common ground, and awareness. Drawing on Kuhn's account of scientific change, we distinguish three epistemic modes: puzzle-solving in normal science, anomaly evaluation, and paradigm-shift exploration. Altogether, these perspectives allow us to analyze scientific agents by how they support coordination under various forms of epistemic uncertainty.

Figure 1 summarizes the framework. Our central claim is that coordination failures emerge when the structure of human–agent collaboration is misaligned with the epistemic demands of the scientific task. For instance, *procedural drift* is salient in bounded execution workflows, *grounding substitution* is critical in hypothesis generation and literature-based reasoning, and *evaluative flattening* becomes consequential when agents participate in peer review or novelty assessment. This framing explains why task-level success can still mask failures in scientific reliability.

To summarize, our survey makes three contributions:

- **A coordination-aware taxonomy.** We introduce **CSCW×Kuhn**, a reusable taxonomy that connects CSCW coordination dimensions with epistemic modes of scientific work.
- **A synthesis of recent systems.** We organize representative human–agent scientific systems across puzzle-solving in normal science, anomaly evaluation, and paradigm-shift exploration, demonstrating how collaboration patterns and system risks vary by epistemic context.
- **An evaluation agenda.** We identify five recurring coordination failures: procedural drift, grounding substitution, evaluative flattening, proxy optimization, and epistemic black-boxing, and translate them into evaluation targets beyond task accuracy.

We use CSCW×Kuhn to synthesize recent systems and benchmarks over three scientific contexts. Section 2 defines the framework and formalizes coordination demand. Section 3 applies the framework to representative human–agent systems across puzzle-solving, anomaly evaluation, and paradigm-shift exploration. Section 4 analyzes current benchmarks and evaluation methods, showing where task-centric evaluation under-measures coordination quality. Section 5 discusses implications for system design, evaluation, and governance, and Section 6 concludes.

2 A Coordination Framework for Human–Agent Scientific Collaboration

Human–agent collaboration in scientific innovation is not only about assigning tasks to agents or evaluating final outputs. Scientific work often unfolds through multi-step workflows, including problem framing, literature review, hypothesis generation, experimentation, analysis, writing, review, and revision. In these workflows, humans and agents must

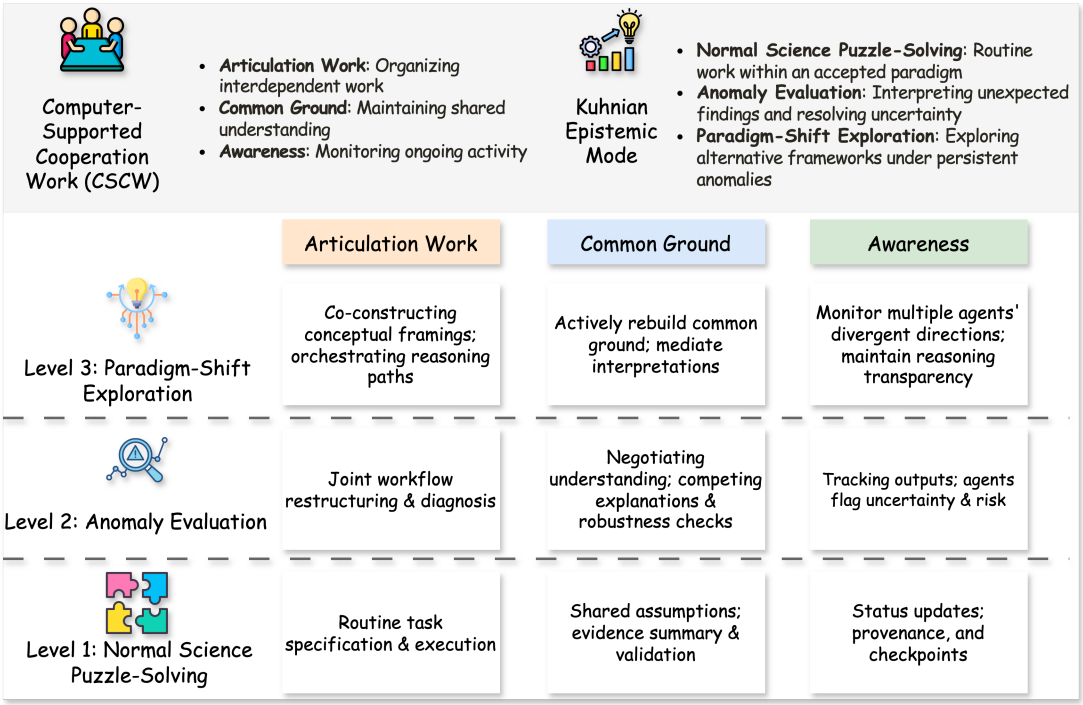

Figure 1: CSCW×Kuhn framework for human–agent scientific collaboration. It maps CSCW coordination dimensions to epistemic modes of scientific innovation, showing that higher epistemic uncertainty requires stronger coordination, grounding, transparency, and oversight.

coordinate goals, assumptions, evidence, intermediate states, and responsibilities. The required coordination varies across scientific contexts. Some tasks involve stable goals and accepted criteria; others involve uncertain claims, contested evidence, or exploratory reasoning.

We present CSCW×Kuhn as a framework for analyzing these differences. The components of CSCW ground the framework in prior work on cooperative work and coordination. In CSCW, collaboration is not treated as simply dividing a task between participants; it also requires organizing interdependent activities, repairing workflow breakdowns, maintaining common understandings, and making ongoing work visible (Schmidt and Bannon, 1992; Schmidt and Simone, 1996; Clark and Brennan, 1991; Dourish and Bellotti, 1992; Gutwin and Greenberg, 2002). This perspective is useful for human–agent scientific collaboration because agents may retrieve evidence, decompose tasks, execute tools, critique claims, or revise outputs, while humans oversee, interpret scientific meaning, and decide if the resulting work is credible. Kuhn's account of scientific change helps distinguish routine puzzle-solving, anomaly evaluation, and paradigm-shift exploration (Kuhn, 2012). Combining CSCW and Kuhn therefore allows us to analyze how coordination demands and evaluation targets change across human–agent scientific workflows.

2.1 Coordination Dimensions

We characterize coordination demand with three dimensions informed by CSCW:

- **Articulation work** (α_t): The work required to organize, assign, sequence, integrate, and repair interdependent tasks. This follows CSCW accounts of articulation work as the often-invisible organizing work that makes cooperative activity possible (Strauss, 1985; Schmidt and Bannon, 1992).

- **Common ground** (γ_t): The shared understanding of goals, assumptions, constraints, evidence, methods, and task state. This dimension draws on grounding theory, which emphasizes that collaboration requires participants to establish and maintain sufficient mutual understanding for the activity at hand (Clark and Brennan, 1991).

- **Awareness** (ω_t): The visibility of actions, intermediate outputs, decisions, errors, and consequences. This dimension follows CSCW work on awareness in shared workspaces, where participants need to understand what others are doing,

what has changed, and how their own actions should adapt (Dourish and Bellotti, 1992; Gutwin and Greenberg, 2002).

At workflow step t, we define coordination demand as:

$$\phi_t = (\alpha_t, \gamma_t, \omega_t), \tag{1}$$

where α_t, γ_t, and ω_t correspond to articulation work, common ground, and awareness. This vector is used as a codable representation of what a human–agent system must support rather than as a numerical metric.

2.2 Scientific Contexts

Coordination demands depend on the scientific context. We use Kuhn's theory (Kuhn, 2012) as an analytical heuristic and distinguish three contexts:

- **Puzzle-solving in normal science**: Work conducted within accepted goals, methods, and evaluation criteria.
- **Anomaly evaluation**: Work that assesses unexpected findings, uncertain claims, or contested contributions.
- **Paradigm-shift exploration**: Exploratory work that searches for alternative concepts, cross-domain connections, or new research directions.

Let K denote the set of scientific contexts:

$$K = \left\{ \begin{array}{l} \text{Puzzle-Solving in Normal Science,} \\ \text{Anomaly Evaluation,} \\ \text{Paradigm-Shift Exploration} \end{array} \right\}, \tag{2}$$

where $k_t \in K$ denotes the dominant context at step t. These categories are not strict boundaries; a scientific workflow may move from routine execution to anomaly evaluation or exploratory reasoning.

2.3 Framework Formulation

We model human–agent scientific collaboration as an iterative workflow that transforms an initial research intent into evidence-backed outputs. Let S denote the set of workflow stages, such as problem framing, literature review, hypothesis generation, experiment design, execution, analysis, writing, review, and revision. The coordination demand at step t depends on both the scientific context and the workflow stage:

$$\phi_t = f(k_t, s_t), \tag{3}$$

where $k_t \in K$ and $s_t \in S$.

This formulation captures the main idea of CSCW×Kuhn—different scientific contexts require different forms of human–agent coordination. In normal-science puzzle-solving, coordination mainly concerns reliable execution and workflow tracking. In anomaly evaluation, it shifts toward evidential comparison, disagreement preservation, and evaluative alignment. In paradigm-shift exploration, coordination becomes more demanding because goals, assumptions, and validation criteria may be unstable.

This framework further clarifies why task-level evaluation is insufficient. A system may complete a benchmark task while failing to preserve workflow integrity, evidential grounding, disagreement, provenance, or human accountability. Therefore, we use CSCW×Kuhn as a coding framework that links scientific context, coordination demand, failure mode, and evaluation target.

We operationalize CSCW×Kuhn as a qualitative coding scheme. For each system or workflow segment, we code for the dominant epistemic context, the level of support for articulation work, common ground, and awareness, and the corresponding failure risks and evaluation targets. Appendix B provides the full rubric and decision rules.

3 Collaboration Patterns Across Scientific Contexts

Corpus and Coding. To apply CSCW×Kuhn, we systematically coded recent human–agent systems and benchmarks along four dimensions: epistemic mode, system cluster, coordination demand, and possible failure risk. The corpus was constructed through a focused search of recent work on scientific agents, AI for science, human–agent collaboration, tool-augmented workflows, peer review, and multi-agent scientific discovery. We use this coding to organize representative systems across puzzle-solving in normal science, anomaly evaluation, and paradigm-shift exploration. Full details on search sources, inclusion criteria, screening, and coding decisions are provided in Appendix C.

Table 1 summarizes how CSCW×Kuhn applies to representative human–agent systems in scientific innovation. The table reports the dominant coordination pattern for each scientific context. Full system-level coding—including coordination-demand cells, representative systems, and failure risks—is provided in Appendix D (Table 2).

Scientific Context	Representative System Clusters	Coordination Focus	Main Failure Risks
Puzzle-solving in normal science	Literature/evidence workflows; design and optimization; protocol execution; shared planning	task decomposition, protocol execution, workflow tracking; domain constraints, data formats, methods, success criteria; intermediate outputs, checkpoints, errors, deviations	Procedural drift; grounding substitution; hidden execution errors; weak long-horizon oversight.
Anomaly evaluation	Hypothesis and novelty evaluation; LLM-as-reviewer; rebuttal and meta-review; institutional review	review criteria, novelty checks, rebuttal and decision workflows; paper–literature–benchmark–venue alignment; uncertainty, disagreement, unsupported claims, confidence signals	Evaluative flattening; shallow fluency; majority-opinion bias; weak adjudication of contested claims.
Paradigm-shift exploration	Cross-domain ideation; falsification and feedback loops; recursive AI scientist workflows; verification/search; infrastructure	evolving goals, search strategies, validation plans; problem–concept–domain–evidence alignment; idea provenance, retrieved evidence, rankings, experimental outcomes	Grounding substitution; proxy optimization; epistemic black-boxing; unsupported or non-testable ideas.

Table 1: Summary of coordination patterns across scientific contexts. Full system-level coding with icons is provided in Table 2.

3.1 Puzzle-Solving in Normal Science

Stable task frames. Puzzle-solving in normal science refers to work conducted within accepted goals, methods, and evaluation criteria (Kuhn, 2012). In this context, the main coordination challenge is reliable execution. Articulation work concerns task decomposition, tool use, protocol following, and workflow integration. Common ground depends on shared assumptions about domain constraints, data formats, methods, and success criteria. Awareness requires visible intermediate outputs, status updates, errors, and checkpoints.

Observed collaboration pattern. Systems in this context mainly support bounded scientific workflows, including literature retrieval, evidence screening, molecular design, protocol execution, and shared planning (Wang et al., 2025b; Ye et al., 2025; Kon et al., 2025; Feng et al., 2026). Agents typically assist by decomposing tasks, expanding search, generating candidates, executing procedures, or maintaining workflow artifacts. Human involvement remains important for checking constraints, validating intermediate results, and correcting deviations.

Evaluation gap. Current evaluations often emphasize task completion, output correctness, or benchmark performance. These measures are useful but incomplete. They may show that an agent completes a predefined task, while missing whether the system preserved the intended workflow, handled intermediate states correctly, and supported human oversight across long-horizon processes.

3.2 Anomaly Evaluation

Contested scientific judgment. Anomaly evaluation refers to the assessment of uncertain findings, contested claims, or contributions whose significance is not yet settled. This includes novelty assessment, peer review, rebuttal analysis, critique generation, and meta-review. In this context, the coordination challenge shifts from reliable execution to evaluative alignment. Here, humans and agents must compare evidence, surface disagreement, and make judgment criteria explicit.

Observed collaboration pattern. Systems in this context often convert scientific judgment into structured workflows (Xiong et al., 2025; Lin et al., 2025; Zhu et al., 2025; Chang et al., 2025; Wu et al., 2026). They retrieve comparison evidence, decompose review criteria, identify unsupported claims, expose reviewer disagreement, or support rebuttal reasoning. Agents are therefore most useful when they help structure judgment rather than replace human evaluation.

Coordination demands. Articulation work involves decomposing scientific merit into criteria such as novelty, soundness, relevance, actionability, and evidential support. Common ground requires alignment among the paper, prior literature, benchmarks, venue expectations, reviewer concerns, and rebuttal evidence. Awareness necessitates making uncertainty, disagreement, unsupported claims,

confidence signals, and revision state visible before humans make higher-stakes decisions.

Evaluation gap. Existing evaluations often rely on prediction accuracy, agreement with human labels, review-quality rankings, groundedness, specificity, or actionability. These metrics do not fully capture whether a system helps humans preserve minority views, distinguish substantive novelty from superficial difference, or adjudicate evidential conflict. This limitation is consequential because AI-assisted review systems may reshape reviewer confidence, rebuttal behavior, and decision outcomes (Russo Latona et al., 2025; Liang et al., 2024; Jin et al., 2024; Zhao et al., 2026).

3.3 Paradigm-Shift Exploration

Unstable goals and criteria. Paradigm-shift exploration refers to upstream work that expands the conceptual search space, connects distant bodies of knowledge, and tests whether speculative directions deserve further investigation (Kuhn, 2012). Unlike normal-science puzzle-solving, this context does not assume that goals, comparison spaces, or validation criteria are already stable.

Observed collaboration pattern. Systems in this context support cross-domain ideation, falsification loops, recursive research workflows, verification search, and scientific infrastructure (Wang et al., 2024b; Baek et al., 2025; Radensky et al., 2026; Gottweis et al., 2026; Ghareeb et al., 2026; Zhang et al., 2025a; Lu et al., 2026; Romera-Paredes et al., 2024; Shojaee et al., 2025). These systems are most defensible when speculative generation is paired with evidence retrieval, critique, executable checks, provenance tracking, or expert revision.

Coordination demands. Paradigm-shift exploration imposes the strongest coordination demands among the tasks analyzed. Articulation work must stabilize evolving goals, search strategies, and validation plans. Common ground is fragile, as agents may introduce unfamiliar concepts, distant analogies, or weakly related literature. Awareness is essential because researchers must inspect the provenance of the idea, retrieved evidence, rankings, experimental results, and verification signals before deciding whether a direction is scientifically meaningful.

Evaluation gap. Current evaluation remains limited for paradigm-shift exploration. Novelty, plausibility, or benchmark success——these criteria do not establish whether a system supports durable conceptual reframing. The question is not necessarily if an agent can generate surprising ideas, but if those ideas remain grounded, testable, interpretable, and useful after scrutiny from experts and follow-up validation.

Recurring failure modes. Over the three scientific contexts, we identify patterns of five coordination failures. *Procedural drift* occurs when collaboration deviates from the intended workflow, protocol, constraint set, or task state. *Grounding substitution* occurs when plausibility, fluency, or apparent novelty replaces evidential support or domain validity. *Evaluative flattening* occurs when contested scientific judgment is reduced to simplified scores, rankings, or majority opinions. *Proxy optimization* occurs when systems optimize benchmark metrics while losing alignment with scientific goals. *Epistemic black-box* occurs when provenance, assumptions, rationale, intermediate states, tool use, or responsibility become difficult to inspect or contest.

4 Benchmarks and Evaluation Methods

Current benchmarks provide useful measures of task performance but only partially evaluate human–agent collaboration in scientific innovation. Summarized in Table 5, most benchmarks focus on bounded outputs, such as retrieval quality, answer correctness, hypothesis plausibility, review quality, code execution, or end-to-end task completion. These metrics improve comparability but often under-measure whether humans and agents preserve workflow integrity, maintain evidential grounding, surface disagreement, and support accountable scientific judgment.

4.1 Limits of Task-Centric Evaluation

Workflow integrity. Many benchmarks decompose scientific work into standalone tasks with clear input–output formats. This design supports reproducible evaluation, but it can obscure whether an agent follows the intended workflow across intermediate steps. For example, AAAR-1.0 evaluates singular scientific tasks with explicit input–output expectations (Lou et al., 2025). Such benchmarks are suitable for measuring local task success but are less able to detect *procedural drift*, where an agent deviates from protocols, constraints, tool-use requirements, or task state. Considering this, evalu-

ation should include trace audits, intermediate-state checks, tool-use inspection, and procedure-fidelity measures in addition to final accuracy.

Evidential grounding. Output-level verification does not necessarily establish scientific grounding. A hypothesis, review, or explanation may appear plausible while remaining weakly supported, hallucinated, or inconsistent with domain knowledge (Xiong et al., 2025). This reflects the risk of *grounding substitution*, where fluency or apparent novelty replaces evidence. Evaluation should test whether claims are supported by relevant sources, whether evidence is traceable, and whether generated conclusions remain consistent with retrieved literature, data, or expert knowledge.

Scientific judgment. Some benchmarks isolate criteria that are easier to score than scientific merit as a whole. SchNovel, for instance, advances the evaluation of scholarly novelty, but novelty remains only one dimension of scientific judgment (Lin et al., 2025). In anomaly evaluation, this creates the risk of *evaluative flattening*, where contested judgments are reduced to a single score or ranking. Evaluation should preserve tradeoffs among novelty, validity, feasibility, significance, evidence quality, and disagreement rather than treating scientific value as a one-dimensional target.

Exploratory validity. Paradigm-shift exploration is especially difficult to evaluate because goals, assumptions, and validation criteria are often unstable. Existing ideation benchmarks still struggle to assess feasibility, diversity, self-evaluation, and the relation between general model capability and scientific idea quality (Si et al., 2025; Ruan et al., 2026). Novelty or plausibility alone cannot show whether an idea is testable, interpretable, or scientifically useful. Evaluation should therefore include assumption tracing, provenance checks, uncertainty reporting, expert critique, and long-horizon follow-up validation.

4.2 Toward Coordination-Aware Evaluation

Evaluation targets. A coordination-aware evaluation should complement task metrics with four additional targets. First, it should assess *process fidelity*: whether agents preserve protocols, constraints, intermediate states, and intended workflows. Second, it should assess *claim–evidence alignment*: whether generated claims are supported by relevant, traceable, and domain-appropriate evidence. Third, it should assess *judgment preservation*: whether systems maintain multiple evaluation criteria and make disagreement visible. Fourth, it should assess *epistemic transparency*: whether users can inspect assumptions, provenance, uncertainty, tool use, and responsibility.

Human oversight. Human oversight should vary with epistemic context. In puzzle-solving settings, humans should primarily verify workflow adherence, tool use, constraint satisfaction, and intermediate outputs. In anomaly evaluation, oversight should focus on evidence comparison, reviewer disagreement, minority perspectives, and the distinction between substantive novelty and superficial difference. In paradigm-shift exploration, oversight should examine how candidate ideas were generated, what assumptions shaped them, what evidence supports them, and what follow-up would make them testable.

Implications for benchmark design. Future benchmarks should move beyond final-output scoring and evaluate collaboration as a process. This requires richer traces, interpretable intermediate artifacts, explicit evidence links, human-intervention points, and longitudinal validation. Such evaluation does not replace task accuracy; instead, it situates accuracy within the broader coordination demands of scientific work. Under CSCW×Kuhn, benchmark completion is neither necessary nor sufficient for success; the operative criterion is if the system can support reliable coordination under the epistemic conditions of the scientific workflow.

5 Discussion

Implications for evaluation. CSCW×Kuhn reframes evaluation from a question of whether an agent completes a task to whether a human–agent system supports the coordination demands of scientific work. This distinction is important because the same benchmark score may reflect different levels of scientific reliability under different epistemic contexts. In puzzle-solving settings, evaluation should not stop at output correctness; it should also test whether the system preserves procedure fidelity, constraint satisfaction, tool-use validity, and intermediate-state consistency. In anomaly evaluation, the key issue is not only whether an agent predicts a human label or generates a plausible review, but whether it helps preserve disagreement, compare alternatives, expose evidential gaps, and

distinguish substantive novelty from superficial difference. In paradigm-shift exploration, evaluation must be even more cautious—novelty and plausibility are insufficient unless the system also supports provenance, feasibility assessment, falsifiability, and follow-up validation. Thus, coordination-aware evaluation complements task metrics by asking what forms of scientific accountability remain visible to human researchers.

Implications for system design. This perspective suggests limits to strong claims about end-to-end scientific automation. Current evidence more strongly supports the design of collaborative scientific support systems than fully autonomous discovery systems. For instance, ScienceAgentBench evaluates bounded workflow tasks and frames agents as assisting scientists with data processing, visualization, and related operations rather than replacing scientific judgment (Chen et al., 2025d). Similarly, benchmarks for hypothesis generation and novelty assessment show that plausible or original outputs still require grounding, expert interpretation, and downstream validation (Si et al., 2025; Xiong et al., 2025). System design should therefore prioritize modular assistance, explicit human checkpoints, inspectable intermediate artifacts, and domain-specific validation interfaces. Agent outputs should be treated as candidate analyses, hypotheses, reviews, or drafts rather than as finalized scientific conclusions. This is especially important in long-horizon workflows, where small errors in retrieval, planning, tool use, or interpretation can accumulate and amplify into misleading scientific claims.

Implications for peer review and institutional governance. Human–agent collaboration also raises institutional questions when agents participate in evaluative settings such as peer review. Review and meta-review are not merely prediction tasks; they are institutional mechanisms for allocating credibility, attention, and publication opportunities. AgentReview shows that review outcomes are shaped by reviewer heterogeneity, latent variables, and privacy constraints, and that biased reviewers can materially affect decisions (Jin et al., 2024). Related evidence suggests that AI-assisted reviewing may influence review scores and acceptance outcomes, particularly for borderline submissions (Russo Latona et al., 2025). These findings imply that the main governance issue is under what disclosure, monitoring, audit, and accountability regimes AI should be used. Institutions need policies that specify when AI assistance is permitted, how its use should be disclosed, how review evidence should be preserved, and how responsibility is assigned when agent-generated judgments influence scientific decisions.

Open challenges. Several challenges remain. First, coordination quality is difficult to benchmark because it is processual, context-dependent, and often visible only through intermediate traces rather than final outputs. Therefore, future benchmarks should include richer logs, evidence links, intervention points, and human-in-the-loop evaluation protocols. Second, long-horizon scientific value is hard to measure. A generated hypothesis may appear novel at the time of evaluation but later prove infeasible, ungrounded, or unproductive. This calls for longitudinal studies that track whether agent-supported ideas, analyses, and reviews remain useful after expert scrutiny and follow-up work. Third, CSCW×Kuhn itself should be empirically validated as a coding scheme. Future studies should test whether researchers can reliably annotate epistemic modes, coordination demands, failure modes, and evaluation targets across real scientific workflows. Finally, governance should be integrated into system development rather than treated as a post-hoc concern. As scientific agents increasingly enter research, review, and publication workflows, progress will increasingly depend on forms and design of collaboration that preserves evidence and distributes responsibility.

6 Conclusion

This survey frames human–agent collaboration in scientific innovation as an epistemically conditioned coordination process. We introduce CSCW×Kuhn to connect scientific context, coordination demand, failure mode, and evaluation target across puzzle-solving, anomaly evaluation, and paradigm-shift exploration. Our synthesis shows that current evaluations often under-measure workflow integrity, evidential grounding, disagreement preservation, transparency, and human oversight. We find that—along with task performance—scientific agents should be assessed by their ability to support accountable coordination under different epistemic conditions.

Limitations

This survey has several limitations. First, the literature on agentic scientific systems is expanding quickly, so our corpus cannot exhaustively cover all emerging systems, benchmarks, and deployment practices. Second, our coding relies on information reported in published papers; when system designs or evaluation procedures are underspecified, coordination demands and failure risks are interpreted from available evidence. Third, CSCW×Kuhn codes the dominant epistemic condition of each system or workflow segment, while real scientific workflows may transition across puzzle-solving, anomaly evaluation, and paradigm-shift exploration. Future work can extend CSCW×Kuhn through larger corpora, empirical annotation studies, and deployment-based validation.

Ethics Statement

This paper is a survey of human–agent collaboration in scientific innovation and does not introduce new human-subject experiments, collect personal data, or release new datasets. Our analysis is based on publicly available research papers and benchmarks. We acknowledge that human–agent scientific systems may raise ethical concerns related to accountability, transparency, overreliance on automated judgments, and the potential misuse of agentic systems in scientific evaluation. Accordingly, we emphasize that such systems should be designed and evaluated with human oversight, provenance tracking, uncertainty disclosure, and clear responsibility assignment.

Information About Use of AI Assistants

AI assistants were used only for language polishing, grammar correction, and improving the clarity and readability of the manuscript. They were not used to generate original scientific claims, conduct literature screening, make coding decisions, perform analysis, or produce conclusions without author verification. All content, references, interpretations, and final writing decisions were reviewed and approved by the authors, who take full responsibility for the manuscript.

References

Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, et al. 2026. Synthesizing scientific literature with retrieval-augmented language models. *Nature*, 650:857–863.

Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. 2025. ResearchAgent: Iterative research idea generation over scientific literature with large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6709–6738, Albuquerque, New Mexico. Association for Computational Linguistics.

Jun Shern Chan, Neil Chowdhury, Oliver Jaffe, James Aung, Dane Sherburn, Evan Mays, Giulio Starace, Kevin Liu, Leon Maksin, et al. 2025. MLE-bench: Evaluating machine learning agents on machine learning engineering. In *Proceedings of the International Conference on Learning Representations*.

Yuan Chang, Ziyue Li, Hengyuan Zhang, Yuanbo Kong, Yanru Wu, Hayden Kwok-Hay So, Zhijiang Guo, Liya Zhu, and Ngai Wong. 2025. TreeReview: A dynamic tree of questions framework for deep and efficient LLM-based scientific peer review. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15651–15682, Suzhou, China. Association for Computational Linguistics.

Qiguang Chen, Mingda Yang, Libo Qin, Jinhao Liu, Zheng Yan, Jiannan Guan, Dengyun Peng, Yiyan Ji, Hanjing Li, et al. 2025a. AI4Research: A survey of artificial intelligence for scientific research. *arXiv preprint arXiv:2507.01903v2*.

Tingting Chen, Srinivas Anumasa, Beibei Lin, Vedant Shah, Anirudh Goyal, and Dianbo Liu. 2025b. AutoBench: An automated benchmark for scientific discovery in LLMs. *arXiv preprint arXiv:2502.15224*.

Wei Chen, Han Ding, Meng Yuan, Zhao Zhang, Deqing Wang, and Fuzhen Zhuang. 2025c. Bridging social psychology and LLM reasoning: Conflict-aware meta-review generation via cognitive alignment. *arXiv preprint arXiv:2503.13879*.

Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, et al. 2025d. ScienceAgentBench: Toward rigorous assessment of language agents for data-driven scientific discovery. In *Proceedings of the International Conference on Learning Representations*.

Herbert H. Clark and Susan E. Brennan. 1991. Grounding in communication. In *Perspectives on Socially Shared Cognition*, pages 127–149. American Psychological Association.

Le Cong, David Smerkous, Xiaotong Wang, Di Yin, Zaixi Zhang, Ruofan Jin, Yinkai Wang, Michal Gerasimiuk, Ravi K. Dinesh, et al. 2025. LabOS: The AI-XR co-scientist that sees and works with humans. *arXiv preprint arXiv:2510.14861v2*.

Paul Dourish and Victoria Bellotti. 1992. Awareness and coordination in shared workspaces. In *Proceedings of the 1992 ACM conference on Computer-supported cooperative work*, pages 107–114.

Jiangshu Du, Yibo Wang, Wenting Zhao, Zhongfen Deng, Shuaiqi Liu, Renze Lou, Henry Zou, Pranav Narayanan Venkit, Nan Zhang, et al. 2024. Llms assist nlp researchers: Critique paper (meta-) reviewing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5081–5099.

Haoyang Fang, Boran Han, Nick Erickson, Xiyuan Zhang, Su Zhou, Anirudh Dagar, Jiani Zhang, Ali Caner Turkmen, Cuixiong Hu, et al. 2025. MLZero: A multi-agent system for end-to-end machine learning automation. In *Advances in Neural Information Processing Systems*.

K. J. Kevin Feng, Kevin Pu, Matt Latzke, Tal August, Pao Siangliulue, Jonathan Bragg, Daniel S. Weld, Amy X. Zhang, and Joseph Chee Chang. 2026. Cocoa: Co-planning and co-execution with AI agents. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery.

Shanghua Gao, Richard Zhu, Pengwei Sui, Zhenglun Kong, Sufian Aldogom, Yepeng Huang, Ayush Noori, Reza Shamji, Krishna Parvataneni, et al. 2025. Democratizing AI scientists using ToolUniverse. *arXiv preprint arXiv:2509.23426v2*.

Ali Essam Ghareeb, Benjamin Chang, Ludovico Mitchener, Angela Yiu, Caralyn J. Szostkiewicz, Dmytro Shved, Gavin J. Gyimesi, Jon M. Laurent, Samantha M. Wright, et al. 2026. A multi-agent system for automating scientific discovery. *Nature*.

Catalina Gomez, Sue Min Cho, Shichang Ke, Chien-Ming Huang, and Mathias Unberath. 2024. Human-AI collaboration is not very collaborative yet: A taxonomy of interaction patterns in AI-assisted decision making from a systematic review. *Frontiers in Computer Science*, 6.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Petar Sirkovic, Artiom Myaskovsky, Grzegorz Glowaty, Felix Weissenberger, Alessio Orlandi, et al. 2026. Accelerating scientific discovery with Co-Scientist. *Nature*.

Palash Goyal, Mihir Parmar, Yiwen Song, Hamid Palangi, Tomas Pfister, and Jinsung Yoon. 2026. ScholarPeer: A context-aware multi-agent framework for automated peer review. *arXiv preprint arXiv:2601.22638*.

Chenyang Gu, Jiahao Cheng, Meicong Zhang, Pujun Zheng, Jinquan Zheng, and Guoxiu He. 2026. MoRI: Learning motivation-grounded reasoning for scientific ideation in large language models. *arXiv preprint arXiv:2603.19044v3*.

Xuemei Gu and Mario Krenn. 2025. Interesting scientific idea generation using knowledge graphs and LLMs: Evaluations with 100 research group leaders. *arXiv preprint arXiv:2405.17044v3*.

Sikun Guo, Amir Hassan Shariatmadari, Guangzhi Xiong, Albert Huang, Myles Kim, Corey M. Williams, Stefan Bekiranov, and Aidong Zhang. 2025. IdeaBench: Benchmarking large language models for research idea generation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pages 5888–5899, New York, NY, USA. Association for Computing Machinery.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 8048–8057.

Carl Gutwin and Saul Greenberg. 2002. A descriptive framework of workspace awareness for real-time groupware. *Computer Supported Cooperative Work (CSCW)*, 11(3):411–446.

Pengcheng Jiang, Jiacheng Lin, Zhiyi Shi, Zifeng Wang, Luxi He, Yichen Wu, Ming Zhong, Peiyang Song, Qizheng Zhang, et al. 2026. Adaptation of agentic AI: A survey of post-training, memory, and skills. *arXiv preprint arXiv:2512.16301v3*.

Yiqiao Jin, Qinlin Zhao, Yiyang Wang, Hao Chen, Kaijie Zhu, Yijia Xiao, and Jindong Wang. 2024. AgentReview: Exploring peer review dynamics with LLM agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1208–1226, Miami, Florida, USA. Association for Computational Linguistics.

Liqiang Jing, Zhehui Huang, Xiaoyang Wang, Wenlin Yao, Wenhao Yu, Kaixin Ma, Hongming Zhang, Xinya Du, and Dong Yu. 2025. DSBench: How far are data science agents from becoming data science experts? In *Proceedings of the International Conference on Learning Representations*.

Nolan Koblischke, Hyunseok Jang, Kristen Menou, and Mohamad Ali-Dib. 2025. Gravity-Bench-v1: A benchmark on gravitational physics discovery for agents. In *Proceedings of the 42nd International Conference on Machine Learning*.

Patrick Tser Jern Kon, Jiachen Liu, Qiuyi Ding, Yiming Qiu, Zhenning Yang, Yibo Huang, Jayanth Srinivasa, Myungjin Lee, Mosharaf Chowdhury, and Ang Chen. 2025. Curie: Toward rigorous and automated scientific experimentation with AI agents. *arXiv preprint arXiv:2502.16069*.

Thomas S. Kuhn. 2012. *The Structure of Scientific Revolutions*, 4 edition. University of Chicago Press. 50th Anniversary Edition; with an introductory essay by Ian Hacking.

Ed Li, Junyu Ren, Xintian Pan, Cat Yan, Chuanhao Li, Dirk Bergemann, and Zhuoran Yang. 2025. Build your personalized research group: A multiagent framework for continual and interactive science automation. *arXiv preprint arXiv:2510.15624*.

Weixin Liang, Zachary Izzo, Yaohui Zhang, Haley Lepp, Hancheng Cao, Xuandong Zhao, Lingjiao Chen, Haotian Ye, Sheng Liu, et al. 2024. Monitoring AI-modified content at scale: A case study on the impact of chatGPT on AI conference peer reviews. In *Forty-first International Conference on Machine Learning*.

Ethan Lin, Zhiyuan Peng, and Yi Fang. 2025. Evaluating and enhancing large language models for novelty assessment in scholarly publications. In *Proceedings of the 1st Workshop on AI and Scientific Discovery: Directions and Opportunities*, pages 46–57, Albuquerque, New Mexico, USA. Association for Computational Linguistics.

Shengchao Liu, Jiongxiao Wang, Yijin Yang, Chengpeng Wang, Ling Liu, Hongyu Guo, and Chaowei Xiao. 2024a. Conversational drug editing using retrieval and domain feedback. In *Proceedings of the International Conference on Learning Representations*.

Sizhe Liu, Yizhou Lu, Siyu Chen, Xiyang Hu, Jieyu Zhao, Yingzhou Lu, and Yue Zhao. 2025. DrugAgent: Automating AI-aided drug discovery programming through LLM multi-agent collaboration. *arXiv preprint arXiv:2411.15692v2*.

Yujie Liu, Zonglin Yang, Tong Xie, Jinjie Ni, Ben Gao, Yuqiang Li, Shixiang Tang, Wanli Ouyang, Erik Cambria, and Dongzhan Zhou. 2026. ResearchBench: Benchmarking LLMs in scientific discovery via inspiration-based task decomposition. *arXiv preprint arXiv:2503.21248v3*.

Zijun Liu, Kaiming Liu, Yiqi Zhu, Xuanyu Lei, Zonghan Yang, Zhenhe Zhang, Peng Li, and Yang Liu. 2024b. AIGS: Generating science from AI-powered automated falsification. OpenReview Archive.

Renze Lou, Hanzi Xu, Sijia Wang, Jiangshu Du, Ryo Kamoi, Xiaoxin Lu, Jian Xie, Yuxuan Sun, Yusen Zhang, et al. 2025. AAAR-1.0: Assessing AI's potential to assist research. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 40361–40383. PMLR.

Chris Lu, Cong Lu, Robert Tjarko Lange, Yutaro Yamada, Shengran Hu, Jakob Foerster, David Ha, and Jeff Clune. 2026. Towards end-to-end automation of AI research. *Nature*, 651:914–919.

Yi Luo, Linghang Shi, Yihao Li, Aobo Zhuang, Yeyun Gong, Ling Liu, and Chen Lin. 2025a. From intention to implementation: Automating biomedical research via LLMs. *Science China Information Sciences*, 68(170105).

Ziming Luo, Zonglin Yang, Zexin Xu, Wei Yang, and Xinya Du. 2025b. LLM4SR: A survey on large language models for scientific research. *arXiv preprint arXiv:2501.04306*.

Bodhisattwa Prasad Majumder, Harshit Surana, Dhruv Agarwal, Bhavana Dalvi Mishra, Abhijeetsingh Meena, Aryan Prakhar, Tirth Vora, Tushar Khot, Ashish Sabharwal, and Peter Clark. 2025. DiscoveryBench: Towards data-driven discovery with large language models. In *Proceedings of the International Conference on Learning Representations*.

Ludovico Mitchener, Angela Yiu, Benjamin Chang, Mathieu Bourdenx, Tyler Nadolski, Arvis Sulovari, Eric C. Landsness, Daniel L. Barabasi, Siddharth Narayanan, et al. 2025. Kosmos: An AI scientist for autonomous discovery. *arXiv preprint arXiv:2511.02824v2*.

Deepak Nathani, Lovish Madaan, Nicholas Roberts, Nikolay Bashlykov, Ajay Menon, Vincent Moens, Mikhail Plekhanov, Amar Budhiraja, Despoina Magka, et al. 2025. MLGym: A new framework and benchmark for advancing AI research agents. In *Proceedings of the Conference on Language Modeling*.

Alexander Novikov, Ngân Vũ, Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, Sergey Shirobokov, Borislav Kozlovskii, Francisco J. R. Ruiz, et al. 2025. AlphaEvolve: A coding agent for scientific and algorithmic discovery. *arXiv preprint arXiv:2506.13131*.

Jiefu Ou, William Walden, Kate Sanders, Zhengping Jiang, Kaiser Sun, Jeffrey Cheng, William Jurayj, Miriam Wanner, Shaobo Liang, et al. 2025. CLAIMCHECK: How grounded are LLM critiques of scientific papers? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 21712–21735, Suzhou, China. Association for Computational Linguistics.

Yansheng Qiu, Haoquan Zhang, Zhaopan Xu, Ming Li, Diping Song, Zheng Wang, and Kaipeng Zhang. 2025. AI Idea Bench 2025: AI research idea generation benchmark. *arXiv preprint arXiv:2504.14191v3*.

Marissa Radensky, Simra Shahid, Raymond Fok, Pao Siangliulue, Tom Hope, and Daniel S. Weld. 2026. Human-LLM compound system for scientific ideation through facet recombination and novelty evaluation. *arXiv preprint arXiv:2409.14634v7*.

Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M. Pawan Kumar, Emilien Dupont, Francisco J. R. Ruiz, Jordan S. Ellenberg, Pengming Wang, et al. 2024. Mathematical discoveries from program search with large language models. *Nature*, 625:468–475.

Kai Ruan, Xuan Wang, Jixiang Hong, Peng Wang, Yang Liu, and Hao Sun. 2026. Evaluating LLMs' divergent thinking capabilities for scientific idea genera-

tion with minimal context. *Nature Communications*, 17(3625).

Giuseppe Russo Latona, Manoel Horta Ribeiro, Tim R. Davidson, Veniamin Veselovsky, and Robert West. 2025. The AI review lottery: Widespread AI-assisted peer reviews boost paper scores and acceptance rates. *Proceedings of the ACM on Human-Computer Interaction*, 9(7):1–28.

Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Michael Moor, Zicheng Liu, and Emad Barsoum. 2025. Agent laboratory: Using LLM agents as research assistants. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 5977–6043, Suzhou, China. Association for Computational Linguistics.

Kjeld Schmidt and Liam Bannon. 1992. Taking cscw seriously: Supporting articulation work. *Computer supported cooperative work (CSCW)*, 1(1):7–40.

Kjeld Schmidt and Carla Simone. 1996. Coordination mechanisms: Towards a conceptual foundation of cscw systems design. *Computer Supported Cooperative Work (CSCW)*, 5(2):155–200.

Chenyang Shao, Dehao Huang, Yu Li, Keyu Zhao, Weiquan Lin, Yining Zhang, Qingbin Zeng, Zhiyu Chen, Tianxing Li, et al. 2025. OmniScientist: Toward a co-evolving ecosystem of human and AI scientists. *arXiv preprint arXiv:2511.16931v2*.

Erzhuo Shao, Yifang Wang, Yifan Qian, Zhenyu Pan, Han Liu, and Dashun Wang. 2026. SciSciGPT: Advancing human–AI collaboration in the science of science. *Nature Computational Science*, 6:301–315.

Chuhan Shi, Xiaoquan Ren, Yifang Wang, Junze Li, Yushi Sun, Yawen Luo, and Rui Sheng. 2026. A survey of human-AI collaboration for scientific discovery. *TechRxiv*. Preprint, version 2.

Parshin Shojaee, Kazem Meidani, Shashank Gupta, Amir Barati Farimani, and Chandan K. Reddy. 2025. LLM-SR: Scientific equation discovery via programming with large language models. In *Proceedings of the International Conference on Learning Representations*.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. 2025. Can LLMs generate novel research ideas? a large-scale human study with 100+ NLP researchers. In *Proceedings of the International Conference on Learning Representations*.

Yiwen Song, Yale Song, Tomas Pfister, and Jinsung Yoon. 2026a. PaperOrchestra: A multi-agent framework for automated AI research paper writing. *arXiv preprint arXiv:2604.05018v1*.

Zhangde Song, Jieyu Lu, Yuanqi Du, Botao Yu, Thomas M. Pruyn, Yue Huang, Kehan Guo, Xiuzhe Luo, Yuanhao Qu, et al. 2026b. Evaluating large language models in scientific discovery. *arXiv preprint arXiv:2512.15567v2*.

Giulio Starace, Oliver Jaffe, Dane Sherburn, James Aung, Jun Shern Chan, Leon Maksin, Rachel Dias, Evan Mays, Benjamin Kinsella, et al. 2025. PaperBench: Evaluating AI's ability to replicate AI research. In *Proceedings of the 42nd International Conference on Machine Learning*.

Anselm Strauss. 1985. Work and the division of labor. *Sociological quarterly*, 26(1):1–19.

Haoyang Su, Renqi Chen, Shixiang Tang, Zhenfei Yin, Xinzhe Zheng, Jinzhe Li, Biqing Qi, Qi Wu, Hui Li, et al. 2025. Many heads are better than one: Improved scientific idea generation by a LLM-based multi-agent system. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28201–28240, Vienna, Austria. Association for Computational Linguistics.

Jiabin Tang, Lianghao Xia, Zhonghang Li, and Chao Huang. 2025. AI-Researcher: Autonomous scientific innovation. In *Advances in Neural Information Processing Systems*.

Guiyao Tie, Pan Zhou, and Lichao Sun. 2026. A survey of AI scientists. *arXiv preprint arXiv:2510.23045v5*.

Khanh-Tung Tran, Dung Dao, Minh-Duong Nguyen, Quoc-Viet Pham, Barry O'Sullivan, and Hoang D. Nguyen. 2025. Multi-agent collaboration mechanisms: A survey of LLMs. *arXiv preprint arXiv:2501.06322*.

Michelle Vaccaro, Abdullah Almaatouq, and Thomas Malone. 2024. When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, 8(12):2293–2303.

Rosni Vasu, Chandrayee Basu, Bhavana Dalvi Mishra, Cristina Sarasua, Peter Clark, and Abraham Bernstein. 2025. HypER: Literature-grounded hypothesis generation and distillation with provenance. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 25413–25438, Suzhou, China. Association for Computational Linguistics.

Vanshika Vats, Marzia Binta Nizam, Minghao Liu, Ziyuan Wang, Richard Ho, Mohnish Sai Prasad, Vincent Titterton, Sai Venkat Malreddy, Riya Aggarwal, et al. 2025. A survey on human-AI collaboration with large foundation models. *arXiv preprint arXiv:2403.04931v3*.

Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, et al. 2023. Scientific discovery in the age of artificial intelligence. *Nature*, 620:47–60.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, et al. 2024a. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(186345).

Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. 2024b. SciMON: Scientific inspiration machines optimized for novelty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 279–299, Bangkok, Thailand. Association for Computational Linguistics.

Yunfei Wang, Yeqin Zhang, Yuyang Wu, Liang Lu, Phi Le Nguyen, Xiaoliang Wang, and Cam-Tu Nguyen. 2025a. MLAlgo-Bench: Can machines implement machine learning algorithms? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 14298–14329, Suzhou, China. Association for Computational Linguistics.

Zifeng Wang, Lang Cao, Qiao Jin, Joey Chan, Nicholas Wan, Behdad Afzali, Hyun-Jin Cho, Chang-In Choi, Mehdi Emamverdi, et al. 2025b. A foundation model for human-AI collaboration in medical literature mining. *Nature Communications*, 16(8361).

Jiaqi Wei, Yuejin Yang, Xiang Zhang, Yuhan Chen, Xiang Zhuang, Zhangyang Gao, Dongzhan Zhou, Guangshuai Wang, Zhiqiang Gao, et al. 2025. From AI for science to agentic science: A survey on autonomous scientific discovery. *arXiv preprint arXiv:2508.14111v2*.

Yixuan Weng, Minjun Zhu, Guangsheng Bao, Hongbo Zhang, Jindong Wang, Yue Zhang, and Linyi Yang. 2025. CycleResearcher: Improving automated research via automated review. In *The Thirteenth International Conference on Learning Representations*.

Yixuan Weng, Minjun Zhu, Qiujie Xie, QiYao Sun, Zhen Lin, Sifan Liu, and Yue Zhang. 2026. DeepScientist: Advancing frontier-pushing scientific findings progressively. In *Proceedings of the International Conference on Learning Representations*.

Hjalmar Wijk, Tao Roa Lin, Joel Becker, Sami Jawhar, Neev Parikh, Thomas Broadley, Lawrence Chan, Michael Chen, Joshua M. Clymer, et al. 2025. RE-bench: Evaluating frontier AI R&D capabilities of language model agents against human experts. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 66772–66832. PMLR.

Sihong Wu, Yiling Ma, Yilun Zhao, Tiansheng Hu, Owen Jiang, Manasi Patwardhan, and Arman Cohan. 2026. RbtAct: Rebuttal as supervision for actionable review feedback generation. *arXiv preprint arXiv:2603.09723v2*.

Wenqing Wu, Chengzhi Zhang, and Yi Zhao. 2025. Automated novelty evaluation of academic paper: A collaborative approach integrating human and large language model knowledge. *Journal of the Association for Information Science and Technology*, 76(11):1452–1469.

Yanzheng Xiang, Hanqi Yan, Shuyin Ouyang, Lin Gui, and Yulan He. 2025. SciReplicate-Bench: Benchmarking LLMs in agent-driven algorithmic reproduction from research papers. In *Proceedings of the Conference on Language Modeling*.

Guangzhi Xiong, Eric Xie, Corey Williams, Myles Kim, Amir Hassan Shariatmadari, Sikun Guo, Stefan Bekiranov, and Aidong Zhang. 2025. Toward reliable scientific hypothesis generation: Evaluating truthfulness and hallucination in large language models. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 7849–7857.

Weikai Xu, Chengrui Huang, Shen Gao, and Shuo Shang. 2025. LLM-based agents for tool learning: A survey. *Data Science and Engineering*, 10:533–563.

Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. 2025. The AI scientist-v2: Workshop-level automated scientific discovery via agentic tree search. *arXiv preprint arXiv:2504.08066*.

Zonglin Yang, Xinya Du, Junxian Li, Jie Zheng, Soujanya Poria, and Erik Cambria. 2024. Large language models for automated open-domain scientific hypotheses discovery. In *ICML 2024 Workshop on AI for Science*.

Zonglin Yang, Wanhao Liu, Ben Gao, Tong Xie, Yuqiang Li, Wanli Ouyang, Soujanya Poria, Erik Cambria, and Dongzhan Zhou. 2025. MOOSE-Chem: Large language models for rediscovering unseen chemistry scientific hypotheses. In *Proceedings of the International Conference on Learning Representations*.

Geyan Ye, Xibao Cai, Houtim Lai, Xing Wang, Junhong Huang, Longyue Wang, Wei Liu, and Xiangxiang Zeng. 2025. Drugassist: A large language model for molecule optimization. *Briefings in Bioinformatics*, 26(1):bbae693.

Zahra Zahedi and Subbarao Kambhampati. 2021. Human-AI symbiosis: A survey of current approaches. *arXiv preprint arXiv:2103.09990*.

Bo Zhang, Shiyang Feng, Xiangchao Yan, Jiakang Yuan, Runmin Ma, Yusong Hu, Zhiyin Yu, Xiaohan He, Songtao Huang, et al. 2025a. InternAgent: When agent becomes the scientist – building closed-loop system from hypothesis to verification. *arXiv preprint arXiv:2505.16938*.

Shuning Zhang, Hui Wang, and Xin Yi. 2025b. Exploring collaboration patterns and strategies in human-AI co-creation through the lens of agency: A scoping review of the top-tier HCI literature. *Proceedings of the ACM on Human-Computer Interaction*, 9(7):1–43.

Yunxiang Zhang, Muhammad Khalifa, Shitanshu Bhushan, Grant D. Murphy, Lajanugen Logeswaran, Jaekyeom Kim, Moontae Lee, Honglak Lee, and Lu Wang. 2025c. MLRC-Bench: Can language

agents solve machine learning research challenges? In *Advances in Neural Information Processing Systems*.

Bingchen Zhao, Jenny Zhang, Chenxi Whitehouse, Minqi Jiang, Michael Shvartsman, Abhishek Charnalia, Despoina Magka, Tatiana Shavrina, Derek Dunfield, et al. 2026. APRES: An agentic paper revision and evaluation system. *arXiv preprint arXiv:2603.03142*.

Tianshi Zheng, Zheye Deng, Hong Ting Tsang, Weiqi Wang, Jiaxin Bai, Zihao Wang, and Yangqiu Song. 2025. From automation to autonomy: A survey on large language models in scientific discovery. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 17733–17750, Suzhou, China. Association for Computational Linguistics.

Xianrui Zhong, Bowen Jin, Siru Ouyang, Yanzhen Shen, Qiao Jin, Yin Fang, Zhiyong Lu, and Jiawei Han. 2025. Benchmarking retrieval-augmented generation for chemistry. In *Proceedings of the Conference on Language Modeling*.

Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. 2025. DeepReview: Improving LLM-based paper review with human-like deep thinking process. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29330–29355, Vienna, Austria. Association for Computational Linguistics.

Xiang Zhuang, Keyan Ding, Tianwen Lyu, Yinuo Jiang, Xiaotong Li, Zhuoyi Xiang, Zeyuan Wang, Ming Qin, Kehua Feng, et al. 2025. Advancing biomolecular understanding and design following human instructions. *Nature Machine Intelligence*, 7:1154–1167.

Henry Peng Zou, Wei-Chieh Huang, Yaozu Wu, Jizhou Guo, Yankai Chen, Chunyu Miao, Hoang Nguyen, Yue Zhou, Weizhi Zhang, et al. 2026. LLM-based human-agent collaboration and interaction systems: A survey. *arXiv preprint arXiv:2505.00753v5*.

A Survey Positioning

This appendix provides supplementary details for the survey scope, CSCW and Kuhnian conceptual grounding, operational coding rubric, corpus construction, analytic coding, and extended tables. Appendix A.1 positions our work relative to existing surveys. Appendix A.2 details the CSCW foundations of our coordination coding. Appendix A.3 clarifies how Kuhn's account of scientific change is used as an analytical lens. Appendix B provides the operational rubric for coding epistemic context, coordination demand, failure modes, and evaluation targets. Appendix C describes corpus construction and coding procedures. Appendix D reports the full system-level coding used to support the synthesis in the main text. Appendix E provides additional tables for operational definitions, survey comparison, and benchmark mapping.

A.1 Comparison with Existing Surveys

To position CSCW×Kuhn, we compare our survey with prior work on AI for science, agentic systems, multi-agent systems, and human–agent collaboration. These survey traditions provide important foundations, but they differ in their primary analytical focus. AI4Science surveys typically organize systems by scientific tasks or research pipelines. Agentic-system surveys focus on model architecture, planning, tool use, memory, and multi-agent coordination. Human–agent collaboration surveys provide stronger accounts of human roles, interaction, and agency distribution.

Our survey addresses a different gap. Rather than organizing systems only by task, architecture, or application domain, we analyze human–agent scientific collaboration as an epistemically conditioned coordination process. This enables us to connect scientific context, coordination demand, human role, evaluation target, and failure mode in a unified coding framework. Table 3 summarizes this positioning across three dimensions: scientific epistemic context, human-role framing, and explicit failure-mode analysis.

A.2 CSCW Foundations for Coordination Coding

CSCW provides the conceptual foundation for the coordination dimensions used in CSCW×Kuhn. We use CSCW as a theory of cooperative work that emphasizes the practical work required to make collaboration possible. Prior research in CSCW has shown that cooperative systems must support task execution, and also the organization of interdependent activities, the maintenance of shared understandings, and the visibility of ongoing work (Schmidt and Bannon, 1992; Schmidt and Simone, 1996; Dourish and Bellotti, 1992; Gutwin and Greenberg, 2002).

This appendix clarifies how we translate CSCW concepts into coding dimensions. **Articulation work** refers to the organizing and repair work required to coordinate interdependent tasks, including sequencing, role assignment, handoffs, integration, and recovery from breakdowns (Strauss, 1985; Schmidt and Bannon, 1992). In human–agent scientific systems, this includes agent planning, tool selection, protocol execution, workflow decompo-

sition, and human intervention points. **Common ground** refers to the extent to which humans and agents share sufficient understanding of goals, assumptions, constraints, evidence, methods, and task state for coordinated action (Clark and Brennan, 1991). In scientific workflows, this includes alignment between research questions, retrieved literature, data, claims, benchmarks, and evaluation criteria. **Awareness** refers to the visibility of actions, intermediate states, errors, decisions, and consequences during collaborative work (Dourish and Bellotti, 1992; Gutwin and Greenberg, 2002). In human–agent scientific systems, this includes logs, provenance, uncertainty signals, rationales, disagreement traces, and checkpoints that allow humans to monitor and contest agent behavior.

These three dimensions were selected because they are theoretically grounded in CSCW and operationally useful for analyzing agentic scientific workflows. The dimensions allow us to code for whether a system supports the coordination work needed for reliable scientific collaboration, rather than only if it produces a correct output. They also connect directly to the failure modes identified in the main text: procedural drift reflects failures of articulation work; grounding substitution reflects failures of common ground; evaluative flattening and epistemic black-boxing reflect failures of awareness, inspectability, and contestability.

A.3 Kuhnian Lens for Scientific Contexts

We use Kuhn's account of scientific change as an analytical heuristic for distinguishing different epistemic conditions in scientific work (Kuhn, 2012). In Kuhn's framework, scientific communities operate within paradigms that define legitimate problems, accepted methods, evidential standards, and criteria for successful explanation. Under stable paradigms, most work proceeds as normal science, where researchers solve puzzles within established assumptions. Scientific work may also involve anomalies (i.e., findings or claims that do not fit prevailing expectations). Some anomalies are resolved within the existing paradigm, whereas others motivate deeper conceptual reconsideration.

This distinction is useful for analyzing human–agent collaboration since different epistemic conditions impose different coordination demands. Routine puzzle-solving requires reliable execution, workflow tracking, and constraint management. Anomaly evaluation requires comparison among competing interpretations, careful treatment of evidence, and preservation of disagreement. Paradigm-shift exploration requires broader conceptual search, provenance over speculative ideas, and mechanisms for feasibility, falsifiability, and follow-up validation.

Accordingly, we operationalize three epistemic modes: *puzzle-solving in normal science*, *anomaly evaluation*, and *paradigm-shift exploration*. These modes are not intended as exhaustive categories of all scientific activity, nor do they imply that real workflows remain fixed within a single mode. Instead, they provide a tractable coding scheme for analyzing how coordination demands change as scientific work moves from stable execution to contested judgment and exploratory reframing.

B Operational Coding Rubric

We operationalize CSCW×Kuhn as a qualitative coding scheme that links epistemic context, coordination demand, failure mode, and evaluation target. The coding unit is a system or workflow segment described in a paper, such as literature retrieval, hypothesis generation, experimental execution, review generation, rebuttal analysis, or recursive research planning. For each unit, we first identify the dominant epistemic context, then assess coordination support along three CSCW dimensions: articulation work (α), common ground (γ), and awareness (ω). We then identify possible failure risks and map them to evaluation targets.

Epistemic context. Puzzle-solving in normal science is assigned when the task has stable goals, established methods, and predefined success criteria, such as protocol execution, benchmarked prediction, evidence extraction, or constrained optimization. Anomaly evaluation is assigned when the task involves judging uncertain, contested, novel, or potentially problematic claims, such as peer review, novelty assessment, rebuttal analysis, critique, or claim verification. Paradigm-shift exploration is assigned when the task involves open-ended search for alternative hypotheses, cross-domain analogies, conceptual recombination, or long-horizon research directions where goals and evaluation criteria are not fully stable.

Coordination demand. We code each CSCW coordination dimension on an ordinal 0–3 scale. These scores are not treated as quantitative performance measures; rather, they provide a transparent way to compare what forms of coordination support

are made available by different systems.

- **Articulation work (α).** 0 = no task decomposition or coordination structure; 1 = fixed pipeline with predefined steps; 2 = dynamic planning with explicit human checkpoints; 3 = adaptive replanning across agents and humans, including handoffs, repair, and revision.

- **Common ground (γ).** 0 = no shared state or explicit task context; 1 = prompt-level task instruction only; 2 = explicit shared memory, workspace, or task state; 3 = shared assumptions, evidence, uncertainty, constraints, and role model are visible and maintained.

- **Awareness (ω).** 0 = final output only; 1 = logs or execution traces are available; 2 = intermediate states, provenance, and decision paths are visible; 3 = humans can inspect, contest, intervene, and revise system actions or outputs.

Failure mode and evaluation target. Failure risks are identified by comparing the epistemic demands of a task with the coordination support provided by the system. **Procedural drift** is coded when a system performs multi-step work with limited task-state tracking, checkpointing, or replanning support. **Grounding substitution** is coded when outputs depend on evidence but provide weak provenance, unsupported claims, or insufficient source alignment. **Evaluative flattening** is coded when contested scientific judgment is reduced to narrow scores, rankings, or majority opinions. **Proxy optimization** is coded when a system is primarily optimized for benchmark metrics without evidence of downstream scientific validity. **Epistemic black-boxing** is coded when assumptions, intermediate reasoning, provenance, tool use, or responsibility boundaries are difficult to inspect or contest.

The corresponding evaluation targets are: workflow coordination for procedural drift, claim–evidence grounding for grounding substitution, multidimensional scientific judgment for evaluative flattening, downstream validation for proxy optimization, and inspectability or contestability for epistemic black-boxing.

C Corpus Construction and Analytic Coding

C.1 Survey Protocol

Our goal was to construct a representative corpus of agentic systems, benchmarks, and evaluation studies relevant to human–agent collaboration in scientific innovation. The corpus was designed to support conceptual synthesis and framework development rather than to serve as an exhaustive systematic review. We therefore prioritized breadth across scientific workflows, agentic system designs, and evaluation settings.

Search sources. We searched ACL Anthology, arXiv, OpenReview, ACM Digital Library, IEEE Xplore, Google Scholar, and publisher websites for recent work on scientific agents, AI for science, LLM-based scientific workflows, multi-agent systems, and human–agent collaboration. We prioritized work published between 2024 and 2026 because this period captures the rapid emergence of LLM-based agents, tool-augmented systems, and multi-agent scientific workflows. Earlier work was included when it provided important conceptual grounding or representative system designs.

Search keywords. Search queries combined terms related to scientific workflows, language agents, collaboration, and evaluation. Representative queries included: *LLM scientific discovery*, *AI scientist*, *scientific agents*, *AI for science*, *AI4Science*, *human-agent collaboration*, *human-AI collaboration*, *human-AI collaborative hypothesis generation*, *RAG scientific literature*, *LLM-as-reviewer*, *automated peer review*, *scientific ideation LLM*, *hypothesis generation LLM*, *tool-use agents science*, *tool-augmented LLM scientific workflow*, and *multi-agent scientific discovery*.

Screening process. Candidate papers were screened in three stages. First, titles and abstracts were screened for relevance to scientific research, discovery, evaluation, or scholarly workflows. Second, full papers were assessed to determine whether the system involved an agentic or LLM-based component relevant to planning, retrieval, critique, judgment, tool use, experimentation, writing, or evaluation. Third, papers were selected for synthesis if they contributed analytical coverage for at least one epistemic mode or system cluster in CSCW$\times$Kuhn. We identified 128 candidate papers. During title and abstract screen-

ing, 43 papers were excluded, leaving 85 papers for full-text assessment. After full-text assessment, 48 unique papers were included in the system synthesis, representing 52 named systems and related artifacts.

Inclusion criteria. We included papers that met at least two of the following criteria: (1) the system supports scientific research, scientific discovery, scholarly evaluation, or research-workflow automation; (2) the system uses LLMs, agents, retrieval-augmented generation, tool use, multi-agent coordination, review, critique, or ideation; (3) the system includes or implies a human role in task planning, judgment, oversight, revision, interpretation, or validation; and (4) the paper reports a concrete system, benchmark, method, or evaluation study relevant to scientific work.

Exclusion criteria. We excluded papers that focused only on general-purpose NLP tasks without a scientific workflow, domain-specific prediction without an agentic or collaborative component, purely theoretical discussions without a system or evaluation target, and applications where human involvement was absent, incidental, or not relevant to the scientific output. We also excluded papers whose relevance was limited to background AI capability rather than scientific collaboration, evaluation, or workflow coordination.

C.2 Analytic Coding

Coding dimensions. Each included work was coded along four dimensions: *epistemic mode*, *system cluster*, *coordination demand*, and *possible failure risk*. The coding was designed to support the comparative synthesis in the main text rather than to produce mutually exclusive labels for all possible system properties.

- **Epistemic mode** identifies the dominant scientific context supported by the system: puzzle-solving in normal science, anomaly evaluation, or paradigm-shift exploration. When a system could fit multiple modes, we assigned the mode corresponding to its primary workflow role and evaluation target.

- **System cluster** captures the dominant workflow function of the system, such as literature and evidence workflows, design and optimization, protocol execution, shared orchestration, novelty evaluation, peer review, rebuttal support, cross-domain ideation, falsification, recursive AI-scientist workflows, verification/search, or scientific infrastructure.

- **Coordination demand** records how the system supports the three CSCW dimensions used in our framework: articulation work (α), common ground (γ), and awareness (ω). Articulation work concerns task organization and workflow integration; common ground concerns shared assumptions, goals, evidence, and constraints; awareness concerns the visibility of actions, intermediate states, rationales, errors, and outcomes.

- **Possible failure risk** identifies the main coordination vulnerability implied by the system design, task setting, interface, or evaluation setup. These risks were used to synthesize recurring failure modes across systems.

Failure-mode synthesis. After coding individual systems, we compared failure risks across epistemic modes and system clusters. We synthesized five recurring coordination failures. **Procedural drift** refers to deviations from intended workflows, protocols, constraints, or task states. **Grounding substitution** occurs when fluency, plausibility, or apparent novelty substitutes for evidential support or domain validity. **Evaluative flattening** occurs when contested scientific judgment is reduced to simplified scores, rankings, criteria, or majority opinions. **Proxy optimization** occurs when systems optimize benchmark metrics or local evaluation signals while losing alignment with broader scientific goals. **Epistemic black-boxing** occurs when provenance, assumptions, intermediate reasoning, tool use, or responsibility become difficult for humans to inspect, contest, or revise.

Use of coding in the survey. The coding was used in two ways. First, it supported the system-level synthesis in Table 2, where representative systems are grouped by epistemic mode, system cluster, coordination demand, and failure risk. Second, it supported the benchmark analysis in Table 5, where evaluation studies are compared according to their coverage of workflow coordination, human oversight, scientific judgment, and downstream validation. This procedure allows the survey to move beyond a list of systems and instead analyze how coordination requirements change across scientific contexts.

Reliability and limitations. The current coding should be interpreted as analytic coding for conceptual synthesis, not as a finalized systematic-review annotation. Coding decisions were checked against the reported system design, workflow role, and evaluation setup of each paper. Because many systems span multiple scientific stages, epistemic modes are not treated as mutually exclusive categories. We assigned the dominant mode to support comparison, while acknowledging that real scientific workflows may move from routine puzzle-solving to anomaly evaluation or exploratory reframing. Future work can extend this codebook through independent annotators, inter-annotator agreement, larger corpora, and deployment-based validation.

D Full System-Level Coding

Table 2 reports the full system-level coding used to construct the synthesis in the main text. For each epistemic mode, we group representative systems into workflow-oriented clusters and identify their primary coordination demands and potential failure risks. Icons indicate artifact type, while the coordination-demand column reports articulation work (α), common ground (γ), and awareness (ω).

E Supplementary Tables

This section provides supplementary tables that support the main analysis. Table 2 provides the full system-level coding of representative human–agent scientific collaboration systems, grouped by epistemic mode and workflow cluster. Table 3 compares existing survey traditions with our survey. Table 4 defines the main terms used throughout the paper. Table 5 maps benchmark and evaluation studies by scientific domain, target task, metric, and coverage gap. The coding procedure behind these tables is described in Appendix B.

System Cluster	Representative Systems	Coordination Demand	Possible Failure Risks
		Puzzle-Solving in Normal Science	
Literature & evidence workflows	[model] **LEADS** (Wang et al., 2025b)	α: subtask decomposition γ: PICO, eligibility, field alignment ω: ranked candidates, rationales, extraction outputs for expert oversight	Missed studies; borderline-case errors; extraction errors
Design & optimization agents	[framework] **ChatDrug** (Liu et al., 2024a) [model] **DrugAssist** (Ye et al., 2025) [model] **InstructBioMol** (Zhuang et al., 2025)	α: generation, editing, optimization subtasks γ: property–format–constraint alignment ω: candidates, feedback, explanations, validation scores	Unsafe, uninterpretable, or infeasible candidate molecules
Execution & protocol systems	[framework, multi-agent system] **Curie** (Kon et al., 2025) [model] **CycleResearcher** (Weng et al., 2025) [multi-agent system] **BioResearcher** (Luo et al., 2025a) [multi-agent system, interface] **LabOS** (Cong et al., 2025)	α: protocol planning, code execution, documentation γ: method–data–variable, reproducibility alignment ω: step state, quality checks, deviations, errors, results	Protocol failures; variable, dataset, or deviation mishandling
Shared planning & orchestration systems	[single-agent, framework] **DrugAgent** (Liu et al., 2025) [framework, infrastructure] **AgentRxiv** (Schmidgall et al., 2025) [multi-agent system] **MLZero** (Fang et al., 2025) [framework, multi-agent system] **freephdlaber** (Li et al., 2025) [multi-agent system] **SciSciGPT** (Shao et al., 2026) [framework, multi-agent system] **PaperOrchestra** (Song et al., 2026a) [interface] **Cocoa** (Feng et al., 2026)	α: plan–subtask–role–tool orchestration, review checkpoints γ: goal–artifact–constraint–workspace alignment ω: plans, reports, memory, feedback, intervention points	Context loss; role misalignment; degraded intermediate information
		Anomaly Evaluation	
Hypothesis generation & novelty evaluation	[benchmark, method] **SchNovel / RAG-Novelty** (Lin et al., 2025) [benchmark, framework] **TruthHypo / KnowHD** (Xiong et al., 2025) [model, method] **Knowledge Interaction Model / Fusion module** (Wu et al., 2025) [model] **HypER** (Vasu et al., 2025) [model] **MoRI** (Gu et al., 2026)	α: novelty, truthfulness, grounding checks γ: review–metadata–literature alignment, knowledge graphs ω: scores, rationales, retrieved evidence, unsupported claims	Conflation of perceived novelty and plausibility with grounded substance
LLM-as-reviewer & critique systems	[framework, benchmark] **DeepReview** / [model] **DeepReviewer-14B** (Zhu et al., 2025) [model] **CycleReviewer** (Weng et al., 2025) [framework] **TreeReview** (Chang et al., 2025) [framework] **ScholarPeer** (Goyal et al., 2026)	α: novelty, soundness, baseline checks γ: venue–literature–benchmark alignment ω: evidence-cited answers, retrieved literature, question trees, verification logs, confidence signals	Shallow fluency; missed flaws or baselines; hallucinated or unusable feedback
Rebuttal & meta-review systems	[benchmark] **ReviewCritique** (Du et al., 2024) [framework] **CAF** (Chen et al., 2025c) [model, benchmark, method] **RBTACT** (Wu et al., 2026)	α: deficiency, conflict, actionable feedback checks γ: paper–review–rebuttal–decision alignment ω: deficient review segments, conflict maps, rebuttal uptake, evidence signals	Majority-opinion bias; overreliance on reviews; weak adjudication
Institutional review systems	[method] **Distributional GPT quantification** (Liang et al., 2024) [multi-agent system, framework] **AgentReview** (Jin et al., 2024) [framework] **APRES** (Zhao et al., 2026)	α: revision, review simulation, decision modeling, AI-use monitoring γ: rubric–reviewer role–rebuttal–decision alignment ω: revision comparisons, simulated review traces, decision shifts, AI-use prevalence	Black-box review or revision outputs; simplified evaluative judgments
		Paradigm-Shift Exploration	
Cross-domain ideation systems	[framework] **SCIMON** (Wang et al., 2024b) [framework] **SciMuse** (Gu and Krenn, 2025) [multi-agent system] **VIRSCI** (Su et al., 2025) [framework, multi-agent system] **ResearchAgent** (Baek et al., 2025) [interface, framework] **Scideator** (Radensky et al., 2026)	α: literature retrieval, facet extraction, cross-domain recombination, novelty checking γ: research problem–concept–facet–domain alignment ω: inspirations, provenance, novelty signals, feedback, revision paths	Weakly grounded ideas; misleading analogies; vague research directions
Falsification & feedback systems	[multi-agent system] **BABY-AIGS** (Liu et al., 2024b) [multi-agent system, framework] **InternAgent** (Zhang et al., 2025a) [multi-agent system] **Kosmos** (Mitchener et al., 2025) [multi-agent system] **Robin** (Ghareeb et al., 2026) [multi-agent system, framework] **AI co-scientist** (Gottweis et al., 2026)	α: hypothesis generation, critique, validation, data analysis, revision loops γ: objective–evidence–data/code–claim alignment ω: falsification/review agents, rankings, experiment results, citations, feedback	Amplified weak hypotheses; proxy overfitting; unsupported long-horizon claims
Recursive AI scientist workflows	[single-agent, framework] **The AI Scientist-v2** (Yamada et al., 2025) [framework, multi-agent system] **AI-Researcher** (Tang et al., 2025) [framework, multi-agent system] **DeepScientist** (Weng et al., 2026) [single-agent, framework] **The AI Scientist** (Lu et al., 2026)	α: ideation, code, experimentation, writing, result analysis, recursive refinement γ: research goal–hypothesis–result–manuscript alignment ω: experiment logs, errors, plots, reviews, manuscript drafts, feedback	Ungrounded ideas; recursive errors; result misinterpretation; hidden methodological gaps
Verification & search systems	[method, framework] **FunSearch** (Romera-Paredes et al., 2024) [framework] **LLM-SR** (Shojaee et al., 2025) [single-agent, framework] **AlphaEvolve** (Novikov et al., 2025)	α: program/equation search and evaluation γ: problem spec–evaluator–code alignment ω: scores, evaluator outputs, program databases, best candidates	Proxy-score overfitting; high-scoring but weakly interpretable outputs
Infrastructure & agent ecosystems	[infrastructure] **ToolUniverse** (Gao et al., 2025) [infrastructure, framework] **OmniScientist** (Shao et al., 2025)	α: tool, protocol, and provenance workflows γ: tool–data–knowledge-graph–norm alignment ω: traces, outputs, provenance, review signals, rankings	Hidden tool misuse; provenance gaps; weak governance; misaligned collaboration

Table 2: Full system-level coding of representative human–agent scientific collaboration systems. Systems are grouped by epistemic mode and workflow cluster. Icons indicate artifact type: model, framework, single-agent or agentic system, multi-agent system, benchmark, method, interface, and infrastructure. The coordination-demand column summarizes articulation work (α), common ground (γ), and awareness (ω).

Survey Theme	Survey Papers	Scientific Epistemic Context	Human Role Framing	Explicit Failure Modes
AI4Science	Luo et al. (2025b)	✓ Addressed	∼ Addressed for peer review/writing; no human-role taxonomy	✗ Absent
	Chen et al. (2025a)	✗ Not addressed; focuses on tasks and workflows	✓ Addressed	✗ Absent
	Wei et al. (2025)	✓ Addressed	∼ Discusses human roles in relation to agents; no human-role taxonomy	✗ Absent
	Zheng et al. (2025)	✓ Addressed	✓ Addressed	✗ Absent
	Tie et al. (2026)	✓ Addressed	∼ Discusses human roles in relation to agents; no human-role taxonomy	✗ Absent
Agentic System	Wang et al. (2024a)	∼ Not central; general application focus	✗ Not addressed in scientific epistemic contexts	✗ Absent
	Guo et al. (2024)	∼ Not central; general application focus	∼ Weakly addressed; no human-role taxonomy	✗ Absent
	Tran et al. (2025)	✗ Not addressed	∼ Weakly addressed; no human-role taxonomy	✗ Absent
	Xu et al. (2025)	✗ Not addressed	∼ Weakly addressed; no human-role taxonomy	✗ Absent
	Jiang et al. (2026)	✓ Addressed	✗ Not addressed	✗ Absent
Human–Agent Collaboration	Zahedi and Kambhampati (2021)	✗ Not addressed	✓ Addressed	✗ Absent
	Vaccaro et al. (2024)	✗ Not addressed; general application focus	∼ Weakly addressed; no human-role taxonomy	✗ Absent
	Gomez et al. (2024)	✗ Not addressed; healthcare application	✓ Broadly addressed in decision-making settings	✗ Absent
	Vats et al. (2025)	✗ Not addressed; healthcare application	✓ Addressed	∼ Not explicit
	Zhang et al. (2025b)	✗ Not addressed	✓ Addressed	✗ Absent
	Shi et al. (2026)	✓ Addressed	✓ Addressed	∼ Not explicit
	Zou et al. (2026)	✗ Not addressed	✓ Addressed	∼ Not explicit
This Survey	**Ours**	✓ **Addressed**	✓ **Addressed**	✓ **Explicit**

Table 3: Comparison of existing survey traditions. The table contrasts prior surveys and this work by whether they explicitly address scientific epistemic context, human-role framing, and failure-mode analysis.

Conceptual Block	Term	Operational Definition
Human–Agent Scientific Collaboration		
Collaboration scope	**Human–Agent Collaboration (HAC)**	A collaborative arrangement in which humans and agentic systems jointly contribute to a scientific goal. We exclude simple retrieval tools and static classifiers, and include agents that participate in planning, judgment, execution, or revision within scientific workflows.
	Agentic system	A system that integrates one or more AI agents capable of planning, reasoning, tool use, memory, workflow execution, or autonomous task decomposition.
Coordination structure	**Human–agent coordination**	The organization of interdependent work between humans and agents, including task allocation, state tracking, feedback, and intervention.
	Agent–agent coordination	The organization of interdependent work among agents in multi-agent systems, including role assignment, communication, delegation, and aggregation.
Coordination Dimensions		
Coordination demand	**Coordination demand**	The type and intensity of coordination required by a scientific workflow under a given epistemic context.
	Articulation work (α)	The work required to organize, assign, sequence, integrate, and revise interdependent tasks so that the workflow remains coherent.
	Common ground (γ)	The shared understanding of goals, assumptions, constraints, evidence, methods, task state, and responsibilities among collaborators.
	Awareness (ω)	The visibility of actions, intermediate outputs, decisions, errors, provenance, and consequences during collaboration.
Scientific Context		
Scientific activity	**Scientific innovation**	The advancement of scientific knowledge through research, discovery, validation, and refinement of methods or explanations.
	Scientific research	Systematic inquiry that formulates questions, applies methods, analyzes evidence, and produces scientific claims, models, tools, or artifacts.
	Scientific discovery	The production of new, externally verifiable scientific knowledge or explanations.
Epistemic condition	**Epistemic context**	The scientific conditions that shape collaboration, including uncertainty, assumptions, goals, evidential standards, and validation criteria.
	Anomaly	A finding, observation, or claim that does not fit prevailing assumptions and raises uncertainty about whether existing explanations remain sufficient.
Epistemic Modes		
	Epistemic mode	A form of scientific work characterized by distinct goals, uncertainty levels, evidential standards, and coordination demands.
	Puzzle-solving in normal science	Routine scientific work conducted within accepted paradigms, methods, assumptions, and evaluation criteria.
	Anomaly evaluation	The assessment of unexpected findings, uncertain claims, or contested contributions to determine their credibility, significance, and relation to existing assumptions.
	Paradigm-shift exploration	Exploratory work that searches for alternative concepts, frameworks, or research directions when existing assumptions may be insufficient.
Coordination Failure Modes		
	Failure mode	A recurring breakdown caused by misalignment between coordination structure and epistemic demands in human–agent scientific collaboration.
	Procedural drift	Deviation from the intended workflow, protocol, constraint set, or task state.
	Grounding substitution	Replacement of evidential support or domain validity with plausibility, fluency, or apparent novelty.
	Evaluative flattening	Reduction of contested scientific judgment into simplified scores, rankings, criteria, or majority opinions.
	Proxy optimization	Optimization of benchmark metrics or proxy objectives while losing alignment with the underlying scientific goal.
	Epistemic black-boxing	Limited inspectability into provenance, assumptions, rationale, intermediate states, tool use, or responsibility.

Table 4: Operational definitions of key concepts used in this survey. Terms are grouped by conceptual function to clarify how CSCW×Kuhn connects human–agent collaboration, coordination demand, epistemic context, and coordination failure.

Research Work	Science Domain	Scope		Target Task	Metric / Approach	Coverage Gap			
		Seg.	Task			C	O	J	V
Execution and Automated Workflow									
MLE-Bench [B] (Chan et al., 2025)	AI / ML	✓		Train models and submit prediction files for Kaggle competitions	Any-medal rate; medal rates; above-median rate; valid submission rate; pass@k; raw competition score			✗	
DiscoveryBench [B] (Majumder et al., 2025)	General	✓		Search and verify hypotheses from datasets and discovery goals	Hypothesis Matching Score; context F1; variable F1; relationship accuracy	✗	✗		
ScienceAgentBench [B] (Chen et al., 2025d)	General		✓	Generate self-contained Python code to solve tasks from published papers	Valid execution rate; success rate; CodeBERTScore; API cost; rubric-based human rating			✗	
Auto-Bench [B] (Chen et al., 2025b)	General		✓	Infer causal graphs through iterative oracle-guided interventions	Success rate; average iterations; adjacency-matrix match; trajectory accuracy		✗		
DSBench [B] (Jing et al., 2025)	Data Science		✓	Answer data-analysis questions and build predictive models	Task-level accuracy; competition-level accuracy; task success rate; performance gap		✗	✗	
AAAR-1.0 [B] (Lou et al., 2025)	General		✓	Validate equations, design experiments, detect paper weaknesses, identify human-review deficiencies	F1; entailment precision/recall/F1; semantic similarity; ROUGE/BERTScore; TF-IDF	✗			
RE-Bench [B] (Wijk et al., 2025)	AI / ML	✓		Solve open-ended ML research-engineering environments	Normalized score; score@k; time budget; human-expert comparison	✗			
Gravity-Bench-v1 [B] (Koblischke et al., 2025)	Physics		✓	Infer gravitational dynamics from budgeted observations in simulated binary-star systems	Task success rate; percent-error threshold; expert-reference comparison; observations used	✗			
MLAlgo-Bench [B] (Wang et al., 2025a)	AI / ML		✓	Implement ML algorithms and workflows from instructions	Pass rate; performance; time overhead; instruction-following score			✗	
PaperBench [B] (Starace et al., 2025)	AI / ML	✓		Replicate AI research papers	Replication Score; rubric pass/fail scores; JudgeEval F1; human baseline comparison		✗	✗	
MLGym-Bench [B] (Nathani et al., 2025)	AI / ML	✓		Improve baselines on open-ended AI research tasks	AUP score; performance profiles; Best Attempt@4; Best Submission@4; raw task metrics		✗	✗	
SciReplicate-Bench [B] (Xiang et al., 2025)	NLP		✓	Generate code from algorithm descriptions in recent NLP papers	Execution accuracy; CodeBLEU; reasoning graph accuracy; dependency/API recall			✗	
Scholarly Judgment and Evidential Support									
AgentReview [E] (Jin et al., 2024)	AI / ML	✓		Simulate peer-review process	Decision agreement; Jaccard index; Cohen's κ; rating mean/variance; review-text length; BERTScore/embedding similarity			✗	
TruthHypo [B] / KnowHD [E] (Xiong et al., 2025)	Biomed.		✓	Generate and verify biomedical entity-relation hypotheses	Link-level precision/recall/F1; relation accuracy; groundedness score; human/GPT-4o truthfulness preference			✗	✗
SchNovel [B] / RAG-Novelty [E] (Lin et al., 2025)	General		✓	Compare paired papers to identify the more novel work	Pairwise accuracy; field-level accuracy; year-gap accuracy; RAG-Novelty accuracy			✗	
ChemRAG-Bench [B] (Zhong et al., 2025)	Chemistry		✓	Evaluate RAG on chemistry question-answering, calculation, molecule, and reaction tasks	Accuracy; exact match; validity; molecular fingerprint similarity; BLEU/ROUGE; error accuracy				✗
MLRC-Bench [B] (Zhang et al., 2025c)	AI / ML	✓		Propose and implement novel ML methods for research competitions	Improvement to human solution; effectiveness; efficiency; code simplicity; pass@k			✗	
Wu et al. (2025) [E] (Wu et al., 2025)	AI / ML		✓	Predict method novelty from novelty sentences in reviews and LLM-generated method summaries	Accuracy; weighted F1			✗	
CLAIMCHECK [B] (Ou et al., 2025)	AI / ML		✓	Analyze and verify review weaknesses in disputed paper claims	Claim-association F1; weakness-label agreement; specificity/groundedness preference; weakness similarity		✗	✗	
ScholarQABench [B] / OpenScholar [E] (Asai et al., 2026)	General	✓		Retrieve relevant literature and synthesize citation-backed answers to scientific queries	Rubric accuracy; citation F1; correctness accuracy; ROUGE-L; LLM-judge quality scores; expert preference	✗			✗
PaperWritingBench [B] (Song et al., 2026a)	AI / ML	✓		Generate LaTeX manuscripts, synthesize targeted related work, construct citations and visuals	Citation F1/recall; literature-review/manuscript quality; simulated acceptance; human pairwise preference	✗			✗
Exploratory Ideation and Research-Direction Value									
MOOSE [B/E] (Yang et al., 2024)	Social Science	✓		Generate hypotheses from raw web corpus	Validness score; novelty score; helpfulness score; expert ratings; feedback-iteration gains				✗
Si et al. (2025) [E] (Si et al., 2025)	NLP		✓	Compare novelty in LLM-generated versus expert-written research ideas through blind expert review	Novelty score; excitement score; feasibility score; expected-effectiveness score; overall score; reviewer agreement				✗
MOOSE-Chem [B/E] (Yang et al., 2025)	Chemistry	✓		Rediscover chemistry hypotheses by inspiration retrieval, composition, and ranking	Hit ratio; Matched Score; average rank ratio; Top MS; Average MS		✗		✗
AI Idea Bench 2025 [B] (Qiu et al., 2025)	AI / ML	✓		Generate research ideas using target and inspiration papers	Idea-to-topic score; idea-to-idea score; multiple-choice success; idea-competition score; novelty score; feasibility score; feasibility-per-step score			✗	✗
IdeaBench [B] (Guo et al., 2025)	General	✓		Generate research ideas from target-paper reference abstracts	Insight Score; semantic similarity; idea-overlap rating; human-ranking alignment				✗
LiveIdeaBench [B] (Ruan et al., 2026)	General		✓	Generate scientific ideas from single-keyword prompts	Originality; feasibility; fluency; flexibility; clarity; overall score			✗	✗
ResearchBench [B] (Liu et al., 2026)	General	✓		Retrieve inspirations, compose hypotheses, and rank hypotheses	Hit ratio; normalized hypothesis-composition score; pairwise ranking accuracy		✗		✗
SDE [B/E] (Song et al., 2026b)	General	✓		Answer scenario-grounded questions; generate, test, and refine hypotheses in discovery loops	Question-level accuracy; normalized project-level subobjective scores			✗	✗

Table 5: Benchmark and evaluation studies for human–agent scientific collaboration. [B] denotes benchmark and [E] denotes evaluation study. Seg. denotes workflow-segment evaluation, and Task denotes task-specific evaluation. Coverage gaps are marked for workflow coordination (C), human oversight (O), scientific judgment (J), and downstream validation (V).