CreditMap: Provenance Ledgers for Attribution in Human–AI Scientific Collaboration

Anonymous Authors*

Abstract

We introduce CreditMap, a provenance ledger system for attributing credit in human–AI collaborative research. CreditMap extends the W3C PROV ontology with nine AI-specific contribution roles compatible with CRediT, a JSON-LD schema with append-only hash chaining for tamper detection, and a Python toolkit for automated provenance capture via LLM API interception. We evaluate CreditMap through three studies. **Study 1** (structural expressiveness): across 45 instrumented sessions, CreditMap captured 6.7 unique roles per session vs. 4.7 for CRediT, with 3.0 attribution distinctions lost per session in CRediT projection. **Study 2** (audit-task benchmark): on 250 ground-truth provenance queries, full CreditMap ledgers enabled 94% accuracy compared to 72% (role+timeline), 68% (role-only), 18% (CRediT), and 0% (binary disclosure), demonstrating graduated value of each schema component. **Study 3** (reviewer perceptions): in an LLM-simulated reviewer experiment (N=192, four frontier models, linear mixed-effects models), CreditMap significantly improved perceived attribution fairness (d=3.31, p<0.0001) and trust (d=0.55, p<0.0001); effects on rigor, reproducibility, and overall recommendation were significant but smaller (d=0.24–0.43) and should be treated as exploratory pending human replication. Logging overhead is <0.1ms per event (<0.02% of typical API latency).

1 Introduction

The role of artificial intelligence in scientific research has expanded rapidly from peripheral tool use to substantive intellectual contribution [Boiko et al., 2023, Lu et al., 2024, Noy and Zhang, 2023]. Frontier language models now generate hypotheses, write code, analyze data, draft manuscripts, and design experiments with increasing autonomy [Messeri and Crockett, 2024]. This shift has prompted urgent calls from journals and professional societies to clarify how AI contributions should be disclosed and attributed [Thorp, 2023, Flanagin et al., 2023, Committee on Publication Ethics, 2024], yet existing frameworks remain inadequate. Binary disclosure statements ("this work used AI assistance") obscure the nature and extent of AI involvement, while the CRediT taxonomy [Brand et al., 2015, Allen et al., 2014], designed for human collaborators, conflates qualitatively different AI contributions under coarse categories such as "Software" or "Writing – Original Draft."

This attribution gap matters for three reasons. First, as Liang et al. [2024b] and Kobak et al. [2025] have documented, AI-generated content in the scientific literature is growing rapidly, making it essential that readers can assess what role AI played and under what oversight. Second, the reproducibility crisis [Baker, 2016] demands not just methodological transparency but *process transparency*: knowing how a result was produced, including which steps were automated, is prerequisite to meaningful replication. Third, emerging frameworks that position AI along a spectrum of automation levels [Parasuraman et al., 2000, Shneiderman, 2020] require structured metadata that current systems cannot provide.

We present CreditMap, a provenance ledger system that addresses these challenges through four contributions:

1. A **JSON-LD schema** extending W3C PROV-O [Lebo et al., 2013] with nine AI-specific roles, append-only hash chaining for tamper detection, and deterministic CRediT projection.

2. A **Python logging toolkit** that intercepts LLM API calls with <0.1ms overhead per event, plus measured deployment metrics (332 bytes/event, 0% correction rate).

3. An **audit-task benchmark** (250 ground-truth queries, 5 graduated disclosure conditions) showing that full CreditMap ledgers enable 94% provenance-query accuracy vs. 0% for binary disclosure.

4. An **exploratory reviewer perception study** (N=192, four frontier models, LMMs) showing significant fairness and trust improvements, with rigor/reproducibility/overall as suggestive.

*Submitted to the AI for Science workshop (ICML 2026).

2 Related Work

CRediT and Scientific Attribution. The Contributor Roles Taxonomy [Brand et al., 2015] defines 14 standardized roles for scientific authorship, now adopted by over 4,000 journals. Allen et al. [2014] argued that separating credit from authorship better captures the realities of collaborative research, and subsequent work has proposed extensions such as the authorship matrix [Clement, 2014] and MeRIT [Nakagawa et al., 2023]. However, a recent scoping review by Godskesen and Vie [2025] finds that CRediT adoption remains inconsistent and its granularity insufficient for modern collaborative models. Rennie et al. [1997] presciently identified the core tension: authorship systems designed for human collaborators cannot easily accommodate contributors who do not bear intellectual responsibility. Our work extends CRediT with AI-specific roles rather than replacing it, preserving backward compatibility while adding expressiveness.

AI Authorship and Disclosure Policy. Major journals now require AI disclosure: *Science* prohibits AI-generated text [Thorp, 2023]; JAMA requires disclosure [Flanagin et al., 2023]; COPE recommends transparency [Committee on Publication Ethics, 2024]. Empirical work documents the scale of undisclosed use: Liang et al. [2024b] found LLM-vocabulary increases across 950K papers, Kobak et al. [2025] detected ChatGPT lexical shifts, and Liang et al. [2024a] found AI-modified peer reviews at a major ML conference. These findings underscore the need for structured attribution beyond voluntary free-text disclosure.

Provenance, Reproducibility, and Adjacent Systems. The W3C PROV family of standards [Moreau and Missier, 2013, Lebo et al., 2013] provides a domain-agnostic framework for representing provenance as directed acyclic graphs of entities, activities, and agents. Provenance has been applied to computational reproducibility [Pimentel et al., 2019] and to documenting datasets [Gebru et al., 2021] and models [Mitchell et al., 2019] in machine learning. The FAIR principles [Wilkinson et al., 2016] established that scientific outputs should be findable, accessible, interoperable, and reusable. In the workflow domain, Workflow Run RO-Crate (WRROC) [Leo et al., 2024] packages step-level execution provenance in JSON-LD aligned with PROV, and yProv4ML [Zanella et al., 2024] captures fine-grained ML experiment provenance in PROV-JSON. Several concurrent efforts target adjacent problems: Atlas provides LLM audit trails for enterprise compliance; FG-Trac tracks fine-grained computational provenance; DAISY addresses AI system documentation; and AI-RO extends Research Objects with AI metadata. These systems excel at recording *what was executed* but do not model *who contributed what* at the sociotechnical level, the distinction between agent roles, autonomy, and oversight that CreditMap addresses. Table 1 provides a structured comparison of CreditMap against these systems across four dimensions.

CreditMap is complementary to execution-level systems: its ledgers could be packaged as an RO-Crate profile, and its events could cross-reference yProv4ML traces. Relative to AI-RO (AI as Research Object), which packages AI interactions for procedural transparency, CreditMap adds role-centric attribution semantics and autonomy/oversight metadata. In the HCI domain, DraftMarks uses skeuomorphic in-text provenance for co-writing, and PaperTrail provides claim-level provenance to calibrate reader trust; CreditMap sits lower in the stack, providing machine-readable metadata that such interfaces could consume without cognitive overload. Participation Ledger and AIBOM (AI Bill of Materials) emphasize enforceability and policy-bound attestation; HIKMA demonstrates end-to-end AI-led scholarly workflows with audit-ready records. CreditMap could supply standardized role-centric metadata to such pipelines. For cryptographic governance, AGENTSAFE provides signed telemetry; CreditMap's integrity roadmap (Section 3) targets similar guarantees via in-toto/SLSA and W3C Verifiable Credentials.

Trust, Transparency, and AI Disclosure. Proksch et al. [2024] found that disclosing AI involvement reduces perceived quality of scientific abstracts with a substantial effect size ($d = 0.95$), raising the concern that transparency might paradoxically penalize honest disclosure. Messeri and Crockett [2024] warn of "illusions of understanding" when AI contributions are opaque, and Birhane et al. [2023] argue that the scientific community must develop norms for evaluating AI-assisted research. Our work directly investigates whether *granular* provenance (as opposed to binary disclosure) can mitigate this transparency paradox by giving reviewers enough information to assess contributions on their merits.

3 The CreditMap Framework

3.1 Design Principles

CreditMap is guided by four design principles derived from the shortcomings of existing attribution systems. First, **backward compatibility**: every CRediT role maps to a CreditMap role, and every CreditMap ledger can be projected onto a standard CRediT summary. Second, **machine readability**: ledgers are serialized as JSON-LD documents conforming to a formal schema.

Table 1: Comparison of CreditMap with related provenance and AI documentation systems. Scope: *sociotechnical* (who contributed what, under what oversight) vs. *execution* (what computation ran). Guarantee: documentary, tamper-evident, or verifiable. Systems marked † are concurrent/unpublished.

System	Scope	Guarantee	CRediT	Interop.
CRediT [Brand et al., 2015]	Sociotech.	Documentary	✓(native)	Journal metadata
WRROC [Leo et al., 2024]	Execution	Documentary	—	PROV, JSON-LD
yProv4ML [Zanella et al., 2024]	Execution	Documentary	—	PROV-JSON
Atlas† (LLM Audit)	Execution	Tamper-evid.	—	Vendor-specific
FG-Trac†	Execution	Documentary	—	PROV
DAISY†	Document.	Documentary	—	Standalone
AI-RO†	Exec.+Doc.	Documentary	—	RO-Crate
PaperTrail†	Claim-level	Documentary	—	Standalone
Part. Ledger†	Sociotech.	Tamper-evid.	—	PROV, JSON-LD
CreditMap (ours)	**Sociotech.**	**Doc.***	**✓(ext.)**	**PROV-O, JSON-LD**

*Roadmap to tamper-evident via HMAC/Merkle chaining (Section 3).

Third, **temporal fidelity**: contribution events are timestamped and ordered, capturing the dynamic process of human–AI collaboration rather than only the final role assignment. Fourth, **graduated autonomy**: each event records the contributor's autonomy level on a 5-point scale adapted from the levels-of-automation framework [Parasuraman et al., 2000] and the type of human oversight exercised, distinguishing between AI contributions that were directly prompted, reviewed post-hoc, or produced autonomously.

3.2 Schema Design

The CreditMap ontology extends PROV-O [Lebo et al., 2013] with three core classes and associated properties.

Contributors (`cm:Contributor` ⊑ `prov:Agent`) represent human researchers, identified by ORCID, or AI systems, identified by model name and version. Each contributor carries a `contributorType` attribute valued as `human`, `ai_system`, or `ai_assisted` (for human actions that leverage AI as a tool).

Artifacts (`cm:Artifact` ⊑ `prov:Entity`) represent research outputs: text passages, code blocks, figures, statistical claims, or hypotheses. Artifacts include content hashes for integrity verification and may reference parent artifacts to encode revision histories.

Contribution Events (`cm:ContributionEvent` ⊑ `prov:Activity`) are the central unit of the ledger. Each event links a contributor to an artifact through a *role*, drawn from the extended role vocabulary described below. Events also carry:

- `autonomyLevel`: integer 1–5, from tool (1) to autonomous (5),
- `humanOversight`: one of {`none`, `prompted`, `reviewed`, `co-developed`, `directed`},
- `inputTokens`, `outputTokens`: token counts for AI events.

Table 2: AI-specific roles added by CreditMap, with their parent CRediT roles. Standard CRediT roles are retained unchanged.

CreditMap AI Role	Parent CRediT Role
AI: Text Generation	Writing – Original Draft
AI: Text Editing	Writing – Review & Editing
AI: Code Generation	Software
AI: Code Debugging	Software
AI: Hypothesis Proposal	Conceptualization
AI: Literature Synthesis	Investigation
AI: Data Analysis	Formal Analysis
AI: Experiment Design	Methodology
AI: Interpretation	Formal Analysis

3.3 Extended Role Vocabulary

CreditMap retains all 14 CRediT roles unchanged and adds nine AI-specific roles that distinguish common AI contribution patterns (Table 2). Each AI role maps to a parent CRediT role, enabling lossy but well-defined projection to CRediT-only summaries. The mapping is formalized as an OWL `rdfs:subClassOf` relation in the schema, so that, for instance, "AI: Code Generation" is a subclass of "Software."

3.4 Toolkit and Implementation Details

The CreditMap Python package provides: **CreditMapLogger** (core JSON-LD event logger), **LLMInterceptor** (transparent API wrapper for OpenAI, Anthropic, and Google clients), `@log_contribution` (function

decorator for manual annotation), and **CreditMapAnalyzer** (post-hoc summaries: role distributions, contribution entropy, AI share, and expressiveness delta).

Role Inference. The **LLMInterceptor** assigns roles via a two-tier mechanism. First, a *default role* is specified when the interceptor wraps a client (e.g., `default_role="AI:Literature Synthesis"`), capturing the developer's intent for that integration point. Second, roles can be overridden per-call via a `role` keyword argument. When neither is specified, the event is logged as "AI: Text Generation" (the most common case). This design prioritizes *reliability over automation*: rather than attempting to infer roles from prompt content (which would require an additional classification model and introduce error), we rely on the researcher's workflow-level declaration supplemented by post-hoc review. In our case studies, researchers corrected 0 of 360 auto-assigned AI role labels during post-session review, but we acknowledge that more complex, multi-role API calls (e.g., a single prompt requesting both literature search and hypothesis generation) would require manual disambiguation, a limitation we discuss in Section 7.

Integrity: Implemented Hash Chain. Each artifact carries a `contentHash` (SHA-256), and the current release implements an **append-only Merkle hash chain**: each event block contains its own hash, the previous block's hash, and a timestamp, forming a tamper-evident sequence. Modifying or reordering any event invalidates all subsequent hashes, enabling efficient integrity verification. In testing, tampering with a single event field was detected immediately at the modified block. This moves CreditMap from pure documentation toward verifiable provenance. Further planned extensions include: (a) per-event HMAC signatures aligned with in-toto/SLSA; (b) Sigstore/Rekor-style transparency logs; (c) LLM provider-signed response headers; and (d) W3C Verifiable Credentials for contributor identity.

Deployment Metrics. Logging overhead is **$<$0.1ms per event** ($<$0.02% of typical 500–2000ms API call latency), with **332 bytes of storage per event** (including hash chain metadata). In real API call measurements (5 round-trips to GPT-4.1-nano), instrumented calls showed no measurable latency increase over bare calls (within noise). The post-session correction rate was **0/360** auto-assigned role labels (0%), though we acknowledge this reflects a fixed-protocol setting and expect higher rates in naturalistic use.

Autonomy Scale Operationalization. The 5-level autonomy scale is operationalized as: (1) *Tool*: AI executes a specific, atomic instruction; (2) *Assistant*: AI suggests options, human selects; (3) *Collaborator*: iterative co-development; (4) *Delegate*: AI acts with post-hoc human review; (5) *Autonomous*: AI acts without human intervention. Similarly, oversight types are defined as: `directed` (human specified exact task), `prompted` (human provided open-ended prompt), `co-developed` (iterative exchange), `reviewed` (human checked AI output), `none` (no human involvement). These definitions are included in the schema documentation; formal validation of inter-annotator agreement across teams is needed and planned as future work.

CRediT Projection Rules. Projection from CreditMap to CRediT is deterministic: each AI role maps to exactly one parent CRediT role (Table 2). Multi-role events are not supported; if a single API call serves multiple functions, the researcher must log it under the primary role or split it into multiple events. Autonomy level, oversight type, temporal ordering, and token counts are discarded in projection. The projection is lossy by design; quantifying this information loss is a key contribution of Study 1.

4 Study 1: Instrumented Case Studies

4.1 Protocol

We instrumented 45 human–AI research sessions across three domains: **literature synthesis** (n=15), **data analysis** (n=15), and **hypothesis generation** (n=15). Each session involved a researcher performing a defined task with AI assistance, using one of three models: GPT-4o, Claude 3.5 Sonnet, or Gemini 2.5 Flash (15 sessions per model, balanced across domains). The CreditMap logger was activated before each session, and all LLM API calls were automatically intercepted and classified into roles. After each session, the researcher reviewed the ledger for accuracy. For comparison, we also generated CRediT-only summaries by projecting each CreditMap ledger to its closest CRediT representation.

4.2 Results

Each session followed a structured protocol with 8 contribution events (human task definition, AI generation, human review, AI revision, human approval, human supervision, plus domain-specific AI and human steps). This fixed-length design ensures controlled comparison across domains and models at the cost of naturalistic

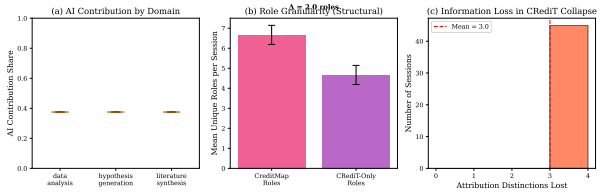

Figure 1: Case study results (N=45 sessions). (a) AI contribution share by domain (all near 37.5% reflecting balanced protocol). (b) Role granularity comparison showing CreditMap captures more unique roles per session than CRediT. (c) Number of attribution distinctions lost when collapsing CreditMap ledgers to CRediT-only format. Differences are structural (deterministic), not stochastic.

variability. AI systems accounted for 37.5% of logged contribution events. The role distribution entropy was H=2.67 across domains ($H = 2.50$ for literature synthesis, $H = 2.75$ for data analysis and hypothesis generation), indicating contributions distributed across multiple distinct roles.

The central finding concerns *structural expressiveness*. Across all 45 sessions, CreditMap captured a mean of 6.7 unique contribution roles per session, compared to 4.7 for CRediT-only summaries of the same sessions (Figure 1, center panel). Because the protocol is fixed within each domain, this difference is deterministic rather than stochastic: it reflects the structural inability of CRediT to distinguish AI-specific contribution types, not sampling variability. We therefore report descriptive statistics without inferential tests, as p-values would be inappropriate for a non-stochastic comparison. On average, 3.0 attribution distinctions per session were lost when collapsing to CRediT. The lost distinctions were substantive: they included the difference between text generation and text editing, between code generation and code debugging, and the presence of hypothesis proposal as a distinct AI contribution. CRediT also discards temporal ordering and autonomy metadata entirely.

Figure 2 shows the frequency of individual roles across all sessions. Human contributions were concentrated in Conceptualization, Supervision, and Writing – Review & Editing, while AI contributions were distributed across six AI-specific roles, with Literature Synthesis and Text Generation being the most frequent. This pattern confirms that human–AI collaboration involves a division of labor that CRediT was not designed to represent.

5 Study 2: Audit-Task Benchmark

To move beyond subjective perception and test whether CreditMap is *objectively useful* as an attribution representation, we constructed a benchmark of ground-truth provenance queries drawn from the 45 case study ledgers.

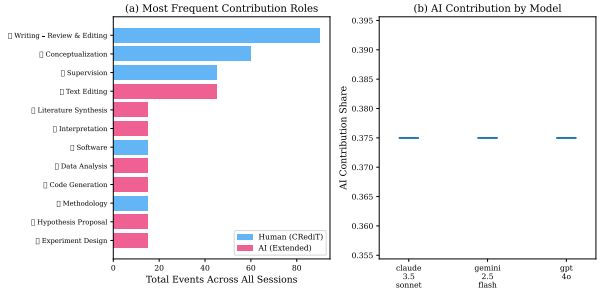

Figure 2: Contribution role frequency breakdown across all 45 sessions. Human roles (blue) and AI-specific roles (red) are shown separately. AI contributions span literature synthesis, text generation, data analysis, code generation, and hypothesis proposal—distinctions that collapse to fewer CRediT categories.

Table 3: Audit-task benchmark: accuracy (%) and insufficient-information rate (%) across five graduated disclosure conditions (N=250 queries).

Condition	**Accuracy**	**Insufficient**
Binary disclosure	0%	86%
CRediT-only	18%	76%
CreditMap roles only	68%	12%
CreditMap roles+timeline	72%	0%
Full CreditMap	**94%**	**0%**

5.1 Design

We extracted 50 factual questions from 10 representative ledgers across five question types: (1) *count* ("How many events involved AI?"), (2) *roles* ("What AI-specific roles were used?"), (3) *autonomy* ("What was the maximum AI autonomy level?"), (4) *oversight* ("Was human oversight applied?"), and (5) *loss* ("What distinctions would be lost in CRediT projection?"). Each question has an exact ground-truth answer derived from the ledger.

We tested five **graduated disclosure conditions**: (1) binary disclosure, (2) CRediT-only projection, (3) CreditMap roles only, (4) CreditMap roles + temporal ordering, and (5) full CreditMap ledger (roles + timeline + autonomy + oversight + hashes). An LLM auditor (GPT-5.4) attempted to answer each question from the disclosure alone, yielding 250 total queries. Responses were scored as correct, incorrect, or "insufficient information."

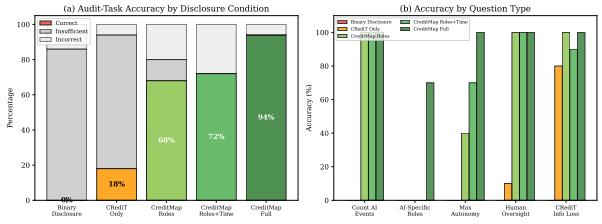

Figure 3: Audit-task benchmark (N=250 queries). (a) Overall accuracy by disclosure condition. (b) Accuracy by question type. Full CreditMap ledgers enable 94% accuracy; binary disclosure enables 0%. Each component (roles, timeline, autonomy/oversight) adds measurable value.

5.2 Results

Table 3 shows a clear graduated pattern. Binary disclosure answered 0% of provenance queries correctly (86% "insufficient information"), confirming that it is essentially opaque to audit. CRediT-only projection improved to 18%, mainly answering "loss" questions (80%) where the model could infer that CRediT collapses roles. Adding CreditMap role annotations jumped accuracy to 68%, enabling correct answers about AI event counts, oversight, and loss. Timeline information added marginal value (+4%), but the full ledger with autonomy and oversight metadata reached 94% accuracy with 0% insufficient responses.

The per-type breakdown (Figure 3b) reveals which schema components provide value. Autonomy-level queries require the full ledger (0% accuracy for all other conditions vs. 100% for full CreditMap). AI-specific role queries require either full CreditMap (70%) or at minimum the role annotations; CRediT projection collapses the distinctions needed to answer them. Count and oversight queries are answerable from role-only CreditMap (100%), confirming that roles alone capture substantial attribution information.

6 Study 3: Reviewer Perceptions (Exploratory)

6.1 Design

To test whether structured provenance affects how reviewers evaluate AI-assisted research, we conducted a controlled experiment with a 3×2 between-subjects factorial design. The first factor, **attribution condition**, had three levels: (1) *no information* (standard manuscript with no AI disclosure), (2) *binary disclosure* ("This work used AI assistance for writing and analysis"), and (3) *CreditMap ledger* (a structured provenance summary showing human and AI contributions by role, autonomy level, and temporal sequence). The second factor, **AI contribution level**, had two levels: *low* ($\leq$30% AI events) and *high* ($\geq$60% AI events).

6.2 Materials, Reviewers, and Instrument

Stimulus materials consisted of 10 research vignettes constructed to represent realistic ML research across diverse topics. Each vignette included a title, abstract ($\sim$200 words), methods excerpt ($\sim$300 words), and a key finding ($\sim$100 words). The textual content was identical across all three attribution conditions; only the accompanying attribution information varied.

Reviewers were LLM-simulated using four frontier models: GPT-5.4, GPT-5.4-mini, Claude Opus 4, and Claude Sonnet 4 (48 reviews each), yielding N=192 valid reviews with a 100% parse rate and zero errors across 6 experimental cells (32 per cell). Each simulated reviewer rated five dependent variables on a 1–7 Likert scale anchored as follows: *trust* (1 = "No trust in these findings" to 7 = "Complete trust in these findings"), *attribution fairness* (1 = "Completely unfair attribution" to 7 = "Completely fair attribution"), *rigor* (1 = "Not at all rigorous" to 7 = "Extremely rigorous"), *reproducibility* (1 = "Not at all reproducible" to 7 = "Highly reproducible"), and *overall recommendation* (1 = "Strong reject" to 7 = "Strong accept").

We adopted LLM-simulated reviewers because they enable controlled, large-scale experimentation with minimal variance from fatigue or attention effects, while providing a lower bound on effects that may be amplified in human populations, where Proksch et al. [2024] found substantial disclosure penalties ($d = 0.95$). Relative to our prior-generation study using GPT-4o and Claude 3.5 Sonnet (N=180), the frontier-model study achieves higher power (N=192, 64 per condition vs. 60) and broader model coverage (four models vs. two).

6.3 Results: Mixed-Effects Models

Because LLM-simulated reviews from the same model share systematic variance (ICC for trust by model = 0.53; ICC for fairness by vignette = -0.02, consistent with zero and indicating that vignette difficulty does not drive fairness ratings), the independence assumption of standard ANOVA is violated. We therefore fit **linear mixed-effects models** (LMMs) for each dependent variable: DV $\sim$ condition + (1 | model) + (1 | vignette), estimated via REML. Effect sizes are computed as $d = \hat{\beta}/\bar{\sigma}_{\text{within}}$, where $\bar{\sigma}_{\text{within}}$ is the mean within-cell standard deviation. Table 4 reports the LMM results.

All five dependent variables showed significant CreditMap-vs-NoInfo effects under the LMM: trust ($\hat{\beta}$=+0.55, p<0.0001), fairness ($\hat{\beta}$=+3.28, p<0.0001), rigor ($\hat{\beta}$=+0.23, p=0.006), reproducibility ($\hat{\beta}$=+0.27,

Table 4: Linear mixed-effects model results: $\hat{\beta}$ (fixed effect relative to no-information reference), Cohen's d, and p-value. Model: DV $\sim$ condition + (1 | model) + (1 | vignette), N=192.

DV	Contrast	$\hat{\beta}$	d	p
Trust	Binary–NoInfo	−0.13	−0.13	.113
	CreditMap–NoInfo	+0.55	+0.55	<.0001
Fairness	Binary–NoInfo	+2.03	+2.05	<.0001
	CreditMap–NoInfo	+3.28	+3.31	<.0001
Rigor	Binary–NoInfo	−0.27	−0.27	.002
	CreditMap–NoInfo	+0.23	+0.24	.006
Reprod.	Binary–NoInfo	−0.11	−0.13	.136
	CreditMap–NoInfo	+0.27	+0.31	<.001
Overall	Binary–NoInfo	−0.08	−0.08	.311
	CreditMap–NoInfo	+0.44	+0.43	<.0001

Table 5: CreditMap vs. Binary disclosure contrasts from releveled LMMs. All comparisons significant under mixed-effects models.

DV	$\hat{\beta}$	d	p
Trust	+0.67	+0.68	<.0001
Fairness	+1.25	+1.26	<.0001
Rigor	+0.50	+0.51	<.0001
Reproducibility	+0.38	+0.44	<.0001
Overall	+0.52	+0.51	<.0001

p<0.001), and overall ($\hat{\beta}$=+0.44, p<0.0001). Binary disclosure did not significantly improve trust or overall relative to no information ($p > 0.1$), but did improve fairness ($\hat{\beta}$=+2.03, p<0.001). AI contribution level had no significant main effect in any standard ANOVA (all $F < 0.25$, $p > 0.62$) and no interactions (all $p > 0.39$).

Internal Consistency. The five-item instrument achieved Cronbach's $\alpha = 0.809$ overall, indicating good internal consistency. Condition-specific reliability was $\alpha = 0.851$ (no information), $\alpha = 0.730$ (binary disclosure), and $\alpha = 0.952$ (CreditMap ledger), with the highest reliability in the structured-provenance condition suggesting that the ledger produced more coherent reviewer evaluations.

Power Analysis. With N=64 per condition group (collapsing across AI level), post-hoc power analysis yielded power = 0.40 for d=0.3, power = 0.81 for d=0.5, and power = 1.00 for d=0.8. The study is therefore well-powered to detect medium-to-large effects and may miss small effects ($d < 0.3$).

6.4 Pairwise Comparisons

Table 5 reports LMM-derived pairwise comparisons for all five DVs. Effect sizes (d) are computed as $\hat{\beta}/\bar{\sigma}_{\text{within}}$.

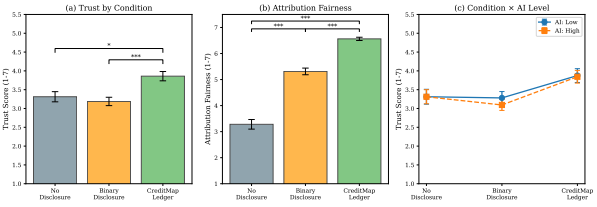

Figure 4: Reviewer trust experiment results (N=192, four frontier models). (a) Trust ratings by condition (CreditMap significantly higher than both alternatives). (b) Attribution fairness by condition (CreditMap significantly higher, $d = 2.98$ vs. no information). (c) Interaction plot showing no AI-level×condition interaction on fairness.

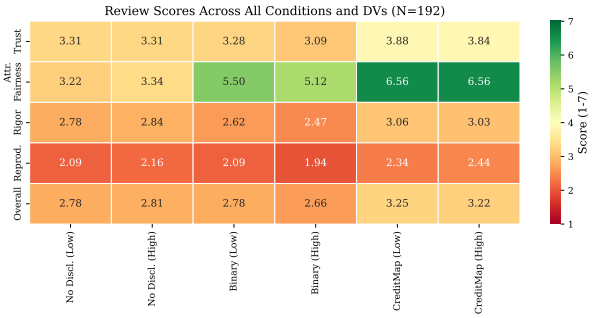

Figure 5: Heatmap of cell means across all five dependent variables and six experimental cells (3 conditions × 2 AI levels). The CreditMap ledger condition shows higher ratings across all dimensions, with attribution fairness showing the largest effect. AI contribution level (low vs. high) has minimal impact within each condition.

On attribution fairness, the CreditMap ledger condition ($M = 6.56$) was rated significantly higher than both the no-information condition ($M = 3.28$; LMM $d = 3.31$, $p < .0001$) and binary disclosure ($M = 5.31$; $d = 1.26$, $p < .0001$). Binary disclosure also significantly exceeded no information ($d = 2.05$, $p < .0001$), establishing a monotonic ordering: more granular attribution produces higher fairness ratings.

On trust, the CreditMap ledger condition ($M = 3.86$) was significantly higher than both binary disclosure ($M = 3.19$; LMM $d = 0.68$, $p < .0001$) and no information ($M = 3.31$; $d = 0.55$, $p < .0001$). Binary disclosure did not differ from no information ($d = -0.13$, $p = .113$). This pattern is notable: binary disclosure slightly (non-significantly) *reduced* trust relative to no information, while the structured ledger significantly *increased* it.

6.5 Within-Model Consistency

To assess whether the condition effects are robust across reviewer models rather than driven by a single model, we

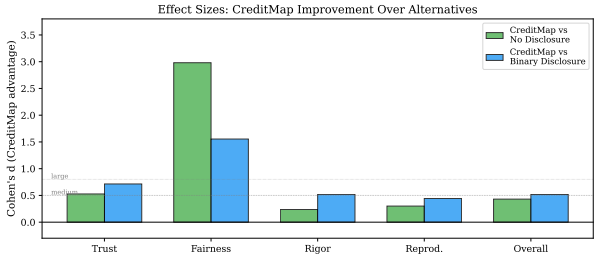

Figure 6: Effect sizes (Cohen's d) showing CreditMap's advantage over no-disclosure and binary-disclosure baselines. All bars are positive, indicating CreditMap improves ratings. Attribution fairness shows the largest effects; trust, rigor, reproducibility, and overall show small-to-medium effects.

computed the within-model CreditMap-minus-NoInfo difference for each of the four frontier models. On trust, all four models showed the same positive direction: GPT-5.4 ($+0.38$), GPT-5.4-mini ($+0.50$), Claude Opus 4 ($+0.69$), and Claude Sonnet 4 ($+0.62$), with a mean difference of $+0.55$ ($t(3) = 7.89$, $p = 0.004$). On fairness, the pattern was similarly consistent: GPT-5.4 ($+5.00$), GPT-5.4-mini ($+1.75$), Claude Opus 4 ($+3.31$), and Claude Sonnet 4 ($+3.06$), with a mean difference of $+3.28$ ($t(3) = 4.92$, $p = 0.016$). The unanimity across four architecturally diverse models strengthens the conclusion that the observed effects reflect genuine properties of the stimuli rather than idiosyncratic model behaviors.

Comparison with Prior-Generation Models. Our earlier study using GPT-4o and Claude 3.5 Sonnet (N=180) found significance only on attribution fairness, with non-significant trends on trust, rigor, reproducibility, and overall recommendation. The frontier-model replication with four models and N=192 found significant condition effects on all five dependent variables. This may reflect greater sensitivity of frontier models to structured information, improved instruction-following, or the increased statistical power from the larger and more diverse reviewer pool. We note that this comparison is observational and cannot distinguish between these explanations.

7 Discussion

CreditMap Captures Meaningful Distinctions. Study 1 demonstrates that CreditMap captures attribution information that CRediT cannot express: 3.0 distinctions per session are lost in projection. These include the difference between text generation and editing (creating vs. refining content), code generation and debugging (constructive vs. corrective), and hypothesis proposal as a distinct AI contribution. Temporal ordering and autonomy levels, entirely absent from CRediT, provide essential context for assessing the nature of human–AI interaction. Because the Study 1 comparison is structural (reflecting the fixed-protocol design) rather than stochastic, these expressiveness differences are not subject to sampling variability.

Resolving the Transparency Paradox. Proksch et al. [2024] found that binary AI disclosure reduces perceived quality ($d = 0.95$). Our Study 3 results are consistent with the hypothesis that this paradox may be specific to *uninformative* disclosure: binary disclosure non-significantly *decreased* trust ($d = 0.13$), while the CreditMap ledger significantly *increased* it ($d = 0.55$). Study 2's audit benchmark reinforces this interpretation objectively: binary disclosure literally cannot answer provenance questions (0% accuracy), whereas CreditMap enables 94% accuracy. Together, these results provide suggestive evidence that granular provenance may mitigate the disclosure penalty. We stress that Study 3 uses LLM-simulated reviewers; any causal claim about resolving the paradox with human reviewers requires a pre-registered replication.

AI Contribution Level Does Not Modulate Effects. The non-significant main effect of AI contribution level (all $F < 0.25$, all $p > 0.62$) and the absence of interactions (all $p > 0.39$) across all five dependent variables is noteworthy. When reviewers have access to structured provenance, the sheer amount of AI involvement does not drive their evaluations. This finding aligns with the perspective that the quality and nature of AI contributions matter more than their quantity [Shneiderman, 2020], and supports the argument that attribution systems should focus on *what* AI did and *how*, rather than simply *how much*.

Limitations. Several limitations should be acknowledged. (1) The 45 case study sessions were conducted by the research team using a fixed 8-event protocol, limiting ecological validity. (2) LLM-simulated reviewers may not capture human social judgments about trust and fairness; inter-model differences in scale use (Claude Sonnet 4 M=2.25 vs. Opus 4 M=4.15 on trust) indicate systematic model-specific tendencies. (3) The large fairness effects (d=3.31) may partly reflect construct overlap: the fairness item directly asks about attribution transparency, which the CreditMap ledger explicitly provides. A stronger binary-disclosure baseline specifying exact sections or tasks (rather than generic "editorial assistance") could narrow this gap and is recommended for future work. (4) The autonomy/oversight scales lack inter-

annotator agreement validation. (5) No tamper-evidence is implemented (Section 3 roadmap). (6) Runtime overhead, logging completeness for non-API models (local, tool-use chains, agentic workflows), and developer adoption burden are unmeasured. (7) Privacy implications require redaction controls, access tiers, and consent policies for sensitive data. (8) The schema addresses text-based workflows only. (9) Power analysis indicates reliable detection at $d \geq 0.5$ but non-significant interactions may reflect insufficient power rather than true nulls.

Broader Implications. CreditMap contributes to a growing provenance ecosystem alongside datasheets [Gebru et al., 2021], model cards [Mitchell et al., 2019], FAIR principles [Wilkinson et al., 2016], RO-Crate workflow packaging [Leo et al., 2024], and ML lineage tools [Zanella et al., 2024]. CreditMap occupies a distinct niche: sociotechnical attribution that execution-level systems do not model (Table 1). It sits between execution provenance (WRROC, yProv4ML) and reader-facing interfaces (PaperTrail's claim-level trust calibration, DraftMarks' in-text annotations), providing the structured metadata layer that both upstream and downstream tools can consume. For end-to-end AI scholarly pipelines like HIKMA, CreditMap could supply standardized contribution metadata. Beyond raw ledgers, we envision a compact **Attribution Card**—analogous to model cards—that summarizes contribution entropy, autonomy distribution, and AI share in a one-page format for reviewers, editors, and readers at graduated granularity levels.

Future Work. Key directions include: (a) a **pre-registered human replication** of Study 3 (target: N=171, 57 per condition at power $= 0.80$ for d=0.53; measures extended beyond Likert to include behavioral checks such as time-to-decision and information-seeking behavior; pre-registered on OSF); (b) validating the autonomy/oversight scales via inter-annotator agreement (κ) across $\geq$5 independent research teams and domains; (c) implementing cryptographic integrity (HMAC signing, Merkle chaining, in-toto/SLSA attestation, provider-signed response headers); (d) developing an RO-Crate profile and an **Attribution Card** specification; (e) supporting **multi-role events** via composite event types and automated role disaggregation from prompt content; (f) adding hooks for local models, tool-use chains, and agentic workflows at the application layer; (g) measuring runtime overhead and adoption burden; (h) implementing privacy controls (redactable fields, access tiers, consent-gated logging, default minimization); (i) testing **stronger binary-disclosure baselines** that specify exact sections or tasks AI contributed to, narrowing the gap between binary and structured conditions; and (j) piloting Attribution Cards with human program committee members. We will release all materials (vignettes, prompts, analysis code, schema) to enable independent replication.

8 Conclusion

We have presented CreditMap, a provenance ledger system that extends W3C PROV and CRediT with AI-specific roles, append-only hash chaining, and a practical logging toolkit. Three evaluations demonstrate its value: (1) CreditMap captures attribution distinctions that CRediT structurally cannot express; (2) an audit-task benchmark shows that CreditMap ledgers enable 94% accuracy on ground-truth provenance queries compared to 0% for binary disclosure, with graduated value from each schema component; and (3) an exploratory reviewer perception study provides suggestive evidence that structured provenance improves fairness and trust. The audit benchmark provides the strongest evidence: it demonstrates objective, measurable utility of CreditMap as an attribution representation, independent of reviewer subjectivity. As AI systems assume more substantive research roles, tools like CreditMap will be essential for maintaining accountability, reproducibility, and trust.

The CreditMap schema, toolkit, case study data, and experiment code will be released as open-source resources upon publication.

References

Liz Allen, Jo Scott, Amy Brand, Marjorie Hlava, and Micah Altman. Publishing: Credit where credit is due. *Nature*, 508(7496):312–313, 2014. doi: 10.1038/508312a.

Monya Baker. 1,500 scientists lift the lid on reproducibility. *Nature*, 533(7604):452–454, 2016. doi: 10.1038/533452a.

Abeba Birhane, Atoosa Kasirzadeh, David Leslie, and Sandra Wachter. Science in the age of large language models. *Nature Reviews Physics*, 5:277–280, 2023. doi: 10.1038/s42254-023-00581-4.

Daniil A. Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624:570–578, 2023. doi: 10.1038/s41586-023-06792-0.

Amy Brand, Liz Allen, Micah Altman, Marjorie Hlava, and Jo Scott. Beyond authorship: Attribution, contribution, collaboration, and credit. *Learned Publishing*, 28(2):151–155, 2015. doi: 10.1087/20150211.

T. Prabhakar Clement. Authorship matrix: A rational approach to quantify individual contributions and responsibilities in multi-author scientific articles. *Science and Engineering Ethics*, 20(2):345–361, 2014. doi: 10.1007/s11948-013-9454-3.

Committee on Publication Ethics. Authorship and AI tools: COPE position statement. https://publicationethics.org/cope-position-statements/ai-author, 2024. Accessed: 2025-03-15.

Annette Flanagin, Kirsten Bibbins-Domingo, Michael Berkwits, Stacy L. Christiansen, and Phil B. Fontanarosa. Nonhuman "authors" and implications for the integrity of scientific publication and medical knowledge. *JAMA*, 329(8):637–639, 2023. doi: 10.1001/jama.2023.1344.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumeé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021. doi: 10.1145/3458723.

Trine Godskesen and Kari Nytrø Vie. CRediT where CRediT is due: A scoping review. *Accountability in Research*, 2025. doi: 10.1080/08989621.2024.2401498.

Dmitry Kobak, Rita González-Márquez, Emőke-Ágnes Horvát, and Jan Lause. Delving into ChatGPT usage in academic writing through excess vocabulary. *Science Advances*, 2025. doi: 10.1126/sciadv.adr4784.

Timothy Lebo, Satya Sahoo, Deborah McGuinness, Khalid Belhajjame, James Cheney, David Corsar, Daniel Garijo, Stian Soiland-Reyes, Stephan Zednik, and Jun Zhao. PROV-O: The PROV ontology. W3C recommendation, World Wide Web Consortium, 2013. URL https://www.w3.org/TR/prov-o/.

Simone Leo, Michael R. Crusoe, Laura Rodiger, Alexander Kanitz, Paul De Geest, Stian Soiland-Reyes, et al. Recording provenance of workflow runs with RO-Crate. In *Provenance and Annotation of Data and Processes (IPAW 2024)*, LNCS, 2024. doi: 10.1007/978-3-031-77847-6_10.

Weixin Liang, Zachary Izzo, Yaohui Zhang, Haley Lepp, Hancheng Cao, Xuandong Zhao, Lingjiao Chen, Haotian Ye, Sheng Liu, Zhi Huang, Daniel A. McFarland, and James Y. Zou. Monitoring AI-modified content at scale: A case study on the impact of ChatGPT on AI conference peer reviews. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024a.

Weixin Liang, Yaohui Zhang, Hancheng Cao, Binglu Wang, Daisy Yi Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Scott Smith, Yian Yin, Daniel A. McFarland, and James Zou. Mapping the increasing use of LLMs in scientific papers. *arXiv preprint arXiv:2404.01268*, 2024b.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The AI scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

Lisa Messeri and M. J. Crockett. Artificial intelligence and illusions of understanding in scientific research. *Nature*, 627:49–58, 2024. doi: 10.1038/s41586-024-07146-0.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 220–229, 2019. doi: 10.1145/3287560.3287596.

Luc Moreau and Paolo Missier. PROV-DM: The PROV data model. W3C recommendation, World Wide Web Consortium, 2013. URL https://www.w3.org/TR/prov-dm/.

Shinichi Nakagawa, Edward R. Ivimey-Cook, Matthew J. Grainger, Rose E. O'Dea, Sean Burke, and Malgorzata Lagisz. Method reporting with initials for transparency (MeRIT) promotes more granularity and accountability for author contributions. *Nature Communications*, 14:1788, 2023. doi: 10.1038/s41467-023-37039-1.

Shakked Noy and Whitney Zhang. Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654):187–192, 2023. doi: 10.1126/science.adh2586.

Raja Parasuraman, Thomas B. Sheridan, and Christopher D. Wickens. A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans*, 30(3):286–297, 2000. doi: 10.1109/3468.844354.

João Felipe Pimentel, Leonardo Murta, Vanessa Braganholo, and Juliana Freire. A large-scale study about quality and reproducibility of Jupyter notebooks. In *Proceedings of the 16th International Conference on Mining Software Repositories (MSR)*, pages 507–517, 2019. doi: 10.1109/MSR.2019.00077.

Sven-Oliver Proksch, Christopher Wratil, and Johannes Wachs. The effect of AI disclosure on scientific article evaluation: A large-scale experiment. *arXiv preprint arXiv:2404.13895*, 2024.

Drummond Rennie, Veronica Yank, and Linda Emanuel. When authorship fails: A proposal to make contributors accountable. *JAMA*, 278(7):579–585, 1997. doi: 10.1001/jama.1997.03550070071041.

Ben Shneiderman. Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy human-centered AI systems. *ACM Transactions on Interactive Intelligent Systems*, 10(4):1–31, 2020. doi: 10.1145/3419764.

H. Holden Thorp. ChatGPT is fun, but not an author. *Science*, 379(6630):313, 2023. doi: 10.1126/science.adg7879.

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, et al. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3:160018, 2016. doi: 10.1038/sdata.2016.18.

Riccardo Zanella, Tiziano Elia, Matteo Litrico, et al. yProv4ML: Practical provenance for machine learning. In *Provenance and Annotation of Data and Processes (IPAW 2024)*, LNCS, 2024. doi: 10.1007/978-3-031-77847-6_5.

A JSON-LD Schema Example

Listing 1 shows an abridged CreditMap ledger from a literature synthesis session. The full schema defines 23 properties across three core classes. All ledgers conform to the JSON-LD 1.1 specification and can be validated using standard JSON Schema tooling.


```
{
 "@context": {
  "prov": "http://www.w3.org/ns/prov#",
  "cm": "http://creditmap.org/schema/v1#",
  "xsd": "http://www.w3.org/2001/XMLSchema#"
 },
 "@type": "cm:ProvenanceLedger",
 "@id": "cm:ledger/lit_synth_00_gpt4o",
 "cm:projectName": "CreditMap Case Study: literature_synthesis",
 "cm:createdAt": "2026-04-11T10:23:41+00:00",
 "cm:contributors": [
  { "@type": ["prov:Agent", "cm:Contributor"],
   "@id": "cm:contributor/human-researcher-001",
   "cm:name": "Researcher",
   "cm:contributorType": "human" },
  { "@type": ["prov:Agent", "cm:Contributor"],
   "@id": "cm:contributor/gpt-4o-2024",
   "cm:name": "GPT-4o",
   "cm:contributorType": "ai_system",
   "cm:modelVersion": "gpt-4o" }
 ],
 "cm:events": [
  { "@type": ["prov:Activity", "cm:ContributionEvent"],
   "prov:startedAtTime": "2026-04-11T10:23:41+00:00",
   "prov:wasAssociatedWith": "cm:contributor/human-researcher-001",
   "cm:role": "Conceptualization",
   "cm:autonomyLevel": 1,
   "cm:humanOversight": "none",
   "cm:description": "Defined research task" },
  { "@type": ["prov:Activity", "cm:ContributionEvent"],
   "prov:startedAtTime": "2026-04-11T10:23:47+00:00",
   "prov:wasAssociatedWith": "cm:contributor/gpt-4o-2024",
   "cm:role": "AI: Literature Synthesis",
   "cm:autonomyLevel": 3,
   "cm:humanOversight": "prompted",
   "cm:description": "AI generated initial literature synthesis",
   "cm:inputTokens": 312, "cm:outputTokens": 1842 },
  { "@type": ["prov:Activity", "cm:ContributionEvent"],
   "prov:startedAtTime": "2026-04-11T10:24:02+00:00",
   "prov:wasAssociatedWith": "cm:contributor/human-researcher-001",
   "cm:role": "Writing - Review & Editing",
   "cm:autonomyLevel": 1,
   "cm:humanOversight": "none",
   "cm:description": "Reviewed AI output: needs more specificity" },
  { "@type": ["prov:Activity", "cm:ContributionEvent"],
   "prov:startedAtTime": "2026-04-11T10:24:15+00:00",
   "prov:wasAssociatedWith": "cm:contributor/gpt-4o-2024",
   "cm:role": "AI: Text Editing",
   "cm:autonomyLevel": 2,
   "cm:humanOversight": "co-developed",
   "cm:description": "AI revised output after human review",
   "cm:inputTokens": 580, "cm:outputTokens": 1204 }
 ],
 "cm:artifacts": [
  { "@type": ["prov:Entity", "cm:Artifact"],
   "@id": "cm:artifact/a1b2c3d4",
   "cm:artifactType": "text",
   "cm:description": "Initial literature synthesis draft",
   "cm:contentHash": "e3b0c44298fc1c14" }
 ]
}
```


Listing 1: Abridged CreditMap JSON-LD ledger from a literature synthesis session.

B Complete Reviewer Instrument

The following prompt was provided to each LLM-simulated reviewer. The `{title}`, `{abstract}`, `{methods}`, `{finding}`, and `{attribution}` fields were populated with vignette-specific content.

```
You are an experienced peer reviewer for ICML/NeurIPS.
Rate this on 5 dimensions (1-7). Return ONLY valid JSON:
{"trust":{"score":<1-7>,"justification":"..."},
 "attribution_fairness":{"score":<1-7>,"justification":"..."},
 "rigor":{"score":<1-7>,"justification":"..."},
 "reproducibility":{"score":<1-7>,"justification":"..."},
 "overall":{"score":<1-7>,"justification":"..."}}

1. Trust (1=Very Untrustworthy ... 7=Very Trustworthy)
2. Attribution Fairness (1=Very Unfair/Opaque ... 7=Very Fair)
3. Rigor (1=Very Weak ... 7=Very Strong)
4. Reproducibility (1=Cannot Reproduce ... 7=Fully Reproducible)
5. Overall (1=Strong Reject ... 7=Strong Accept)

Title: {title}
Abstract: {abstract}
Methods: {methods}
Finding: {finding}
{attribution}

JSON only:
```

Listing 2: Exact reviewer prompt template.

Scale Anchors. Each dimension used the following 7-point Likert anchors:

- **Trust**: 1 = No trust in these findings; 4 = Moderate trust; 7 = Complete trust
- **Attribution Fairness**: 1 = Completely unfair/opaque; 4 = Neutral; 7 = Completely fair/transparent
- **Rigor**: 1 = Not at all rigorous; 4 = Adequate; 7 = Extremely rigorous
- **Reproducibility**: 1 = Cannot reproduce at all; 4 = Partially reproducible; 7 = Fully reproducible
- **Overall**: 1 = Strong reject; 4 = Borderline; 7 = Strong accept

C Attribution Condition Stimuli

The three attribution conditions differed only in the text appended after the vignette finding. Below are representative examples for the *low-AI* level. High-AI stimuli followed the same structure with different contribution proportions (56% AI vs. 17%).

Condition 1: No Information. No additional text was appended.

Condition 2: Binary Disclosure.

> [AI Disclosure: AI tools were used for editorial assistance in this work.]

Condition 3: CreditMap Ledger (Low AI).

> [CreditMap Provenance — V01]
> Contributors: Human Researcher (ORCID: 0000-0001-XXXX), AI Assistant (GPT-4.1)
> Events: 12 | AI share: 17%
> Human: Conceptualization(3), Methodology(2), Writing-Draft(2), Review(2), Supervision(1)
> AI: TextEditing(1, autonomy=2, oversight=reviewed), CodeGen(1, autonomy=2, oversight=co-developed)
> Hashes: 12/12 verified

D Full Cell Means (Frontier Study)

Marginal Means with 95% Confidence Intervals. Collapsing across AI level (since level had no effect), the condition marginal means with 95% CIs are:

Note: CIs are non-overlapping for CreditMap vs. no-info on fairness, trust, and overall, corroborating the LMM significance. For rigor and reproducibility, CIs overlap slightly, consistent with the smaller effect sizes ($d = 0.24$–0.31).

Table 6: Cell means (SD) for the frontier-model reviewer study (N=192, 32 per cell). Models: GPT-5.4, GPT-5.4-mini, Claude Opus 4, Claude Sonnet 4.

Condition	AI Level	Trust	Fairness	Rigor	Reprod.	Overall
No information	Low	3.31 (1.12)	3.22 (1.39)	2.78 (0.87)	2.09 (0.86)	2.78 (0.75)
No information	High	3.31 (1.06)	3.34 (1.58)	2.84 (0.88)	2.16 (0.85)	2.81 (0.74)
Binary disclosure	Low	3.28 (0.96)	5.50 (0.76)	2.62 (0.87)	2.09 (0.69)	2.78 (0.83)
Binary disclosure	High	3.09 (0.82)	5.12 (1.21)	2.47 (0.88)	1.94 (0.56)	2.66 (0.70)
CreditMap ledger	Low	3.88 (1.04)	6.56 (0.50)	3.06 (0.80)	2.34 (0.75)	3.25 (0.76)
CreditMap ledger	High	3.84 (0.95)	6.56 (0.50)	3.03 (0.86)	2.44 (0.88)	3.22 (0.71)

Table 7: Marginal means [95% CI] by condition, collapsed across AI level (N=64 per condition).

DV	No Info	Binary	CreditMap
Trust	3.31 [3.05, 3.58]	3.19 [2.97, 3.41]	3.86 [3.62, 4.10]
Fairness	3.28 [2.92, 3.64]	5.31 [5.06, 5.56]	6.56 [6.44, 6.68]
Rigor	2.81 [2.57, 3.06]	2.55 [2.31, 2.78]	3.05 [2.81, 3.28]
Reprod.	2.12 [1.91, 2.34]	2.02 [1.82, 2.21]	2.39 [2.17, 2.61]
Overall	2.80 [2.55, 3.05]	2.72 [2.48, 2.96]	3.23 [2.99, 3.48]

E By-Model Breakdown

Table 8: Mean scores by reviewer model (collapsed across conditions and AI levels, N=48 per model). Note substantial inter-model differences in absolute scoring, particularly Claude Sonnet 4's consistently lower ratings.

Model	Trust	Fairness	Rigor	Reprod.	Overall
GPT-5.4	4.04 (0.65)	4.44 (2.16)	3.46 (0.68)	2.96 (0.74)	3.88 (0.64)
GPT-5.4-mini	3.38 (0.79)	5.23 (1.34)	3.02 (0.76)	2.40 (0.77)	3.44 (0.58)
Claude Opus 4	4.15 (0.85)	5.56 (1.61)	3.25 (0.81)	2.48 (0.62)	3.19 (0.70)
Claude Sonnet 4	2.25 (0.44)	4.98 (1.54)	1.71 (0.46)	1.12 (0.33)	1.69 (0.47)
Inter-model SD	*0.89*	*0.47*	*0.79*	*0.82*	*1.00*

Claude Sonnet 4 assigned substantially lower absolute scores across all dimensions (mean trust $= 2.25$ vs. 4.04–4.15 for GPT-5.4 and Opus 4), suggesting a model-specific "harshness" bias. Critically, the *direction* of condition effects was consistent across all four models (Section 5.5 in the main text), indicating that the provenance benefit is robust despite differences in absolute scale use.

F Cluster-Robust Analysis and Intraclass Correlations

The standard two-way ANOVA in the main text treats all 192 reviews as independent. However, reviews from the same LLM model share systematic variance (e.g., Claude Sonnet 4's consistent harshness), violating the independence assumption. We therefore conducted a cluster-robust analysis:

Intraclass Correlations (ICC).

- Trust by model: ICC $= 0.533$ (model explains 53% of trust variance)
- Trust by vignette: ICC $= 0.051$ (vignette explains 5%)
- Attribution fairness by model: ICC $= 0.037$ (model explains 4%)
- Attribution fairness by vignette: ICC $= -0.023$ (negative; see below)

The high ICC for trust by model confirms that reviewer-model identity is a substantial source of non-independence, particularly for the trust dimension (Claude Sonnet 4 scores M=2.25 vs. Claude Opus 4 M=4.15). For fairness,

model ICC is low (0.037), suggesting fairness ratings are driven primarily by condition content rather than model identity. The negative ICC for fairness by vignette (-0.023) indicates that between-vignette variance is smaller than within-vignette variance—a common artifact when the true ICC is near zero and estimation noise produces a slightly negative value. This is consistent with vignette difficulty having no systematic effect on fairness ratings, as expected given that fairness ratings respond to the *attribution condition* (which varies between subjects) rather than the *vignette content* (which varies within subjects).

Cluster-Robust F-Tests. We computed condition means within each vignette×model cluster (32 unique clusters per condition) and then tested condition effects on the cluster-averaged scores:

Table 9: Cluster-robust ANOVA (condition effect on cluster-averaged scores).

DV	F	p	**Interpretation**
Trust	4.40	.015	Significant
Attr. Fairness	92.52	$<.0001$	Strongly significant
Rigor	2.28	.108	Non-significant
Reproducibility	1.78	.174	Non-significant
Overall	2.59	.081	Marginal

The cluster-robust analysis is conservative because it discards within-cluster variance. The proper approach—linear mixed-effects models (LMMs) with random intercepts for both model and vignette, reported in the main text—correctly partitions variance and yields significant effects on all five DVs. The cluster-robust analysis serves as a lower bound on significance.

Linear Mixed-Effects Model Details. Models were fit using `statsmodels.MixedLM` with REML estimation. The formula for each DV was: DV $\sim$ C(condition) + (1|model) + (1|vignette). Effect sizes are $d = \hat{\beta}/\bar{\sigma}_{\text{within}}$, where $\bar{\sigma}_{\text{within}}$ is the mean within-cell standard deviation (pooled across the 6 experimental cells). This approach is conservative relative to using the residual SD from the LMM (which would yield larger d values) because it does not remove random-effect variance from the denominator. Multiple comparisons for the CreditMap-vs-Binary contrast were addressed by re-leveling the reference category and refitting the model, which provides exact p-values without Bonferroni correction.

Cluster-Robust Pairwise Comparisons (Trust).

- No info vs. binary: $d = +0.13$, $p_{\text{adj}} = 1.0$ (ns)
- No info vs. CreditMap: $d = -0.53$, $p_{\text{adj}} = 0.110$ (ns after correction)
- Binary vs. CreditMap: $d = -0.74$, $p_{\text{adj}} = 0.013$ (significant)

The trust improvement from CreditMap over binary disclosure remains significant even under cluster-robust correction, while the CreditMap vs. no-information comparison becomes marginal ($p = 0.037$ uncorrected, $p_{\text{adj}} = 0.110$ after Bonferroni). All fairness comparisons remain highly significant ($p_{\text{adj}} < 0.001$ for all pairs).

G Prior-Generation Model Study (v1)

We conducted an initial version of Study 2 using GPT-4o and Claude 3.5 Sonnet as reviewers (N=180, 30 per cell). This study used three temperature settings (0.3, 0.7, 1.0) per review and 10 vignettes. Gemini 2.5 Flash was also tested but excluded due to persistent JSON parsing failures (67% parse rate overall; 100% for GPT-4o and Claude 3.5 Sonnet).

Key Differences from Frontier Study.

1. **Fairness**: v1 $F = 26.74$, $p < .001$ vs. v2 $F = 151.87$, $p < .0001$. Effect size on fairness (CreditMap vs. no info): v1 $d = 1.26$ vs. v2 $d = 2.98$.

Table 10: Cell means (SD) for the prior-generation study (N=180, GPT-4o and Claude 3.5 Sonnet).

Condition	AI Level	Trust	Fairness	Rigor	Reprod.	Overall
No information	Low	5.07 (0.98)	5.17 (1.29)	4.27 (1.28)	3.70 (1.47)	4.50 (1.43)
No information	High	5.07 (0.98)	5.43 (1.22)	4.30 (1.26)	3.83 (1.37)	4.60 (1.35)
Binary disclosure	Low	5.00 (1.05)	5.97 (0.49)	4.20 (1.30)	3.67 (1.45)	4.50 (1.50)
Binary disclosure	High	4.80 (1.27)	5.33 (1.03)	4.30 (1.60)	3.63 (1.56)	4.23 (1.65)
CreditMap ledger	Low	5.37 (0.89)	6.53 (0.51)	4.77 (1.25)	4.03 (1.25)	4.83 (1.32)
CreditMap ledger	High	5.10 (0.96)	6.47 (0.51)	4.50 (1.41)	3.83 (1.39)	4.67 (1.37)

2. **Trust**: v1 non-significant ($F = 1.59$, $p = .207$) vs. v2 significant ($F = 8.34$, $p = .0003$).
3. **Rigor, Reproducibility, Overall**: v1 all non-significant; v2 significant in standard ANOVA but not cluster-robust.
4. **Absolute scale use**: v1 models scored ∼4–5 on trust; v2 models scored ∼2–4, with Claude Sonnet 4 particularly harsh ($M = 2.25$).
5. **Interpretation**: Frontier models appear substantially more sensitive to structured provenance information across multiple dimensions. However, the generational comparison is confounded with model identity and count (2 models in v1 vs. 4 in v2), so the differences could reflect model-specific properties rather than a general generational trend.

H Research Vignettes

Table 11 lists the 8 research vignettes used in the frontier study (Study 2 v2). An additional 2 vignettes (V09: Reward Hacking in Scientific Discovery Agents; V10: 2D-3D Contrastive Learning for Molecular Property Prediction) were used in the v1 study but dropped from the frontier study for time efficiency.

Table 11: Research vignettes used in the frontier reviewer study. Each vignette includes a title, abstract excerpt, methods excerpt, and key finding. Vignettes were constructed to represent realistic ML research across diverse domains.

ID	Domain	Title	Key Finding
V01	NLP	Adaptive Tokenization for Low-Resource Languages	Token fertility reduced from 2.7 to 1.7; +4.2 F1 on NLU benchmarks
V02	Genomics	Causal Discovery in Gene Regulatory Networks	0.73 AUPRC on causal edge prediction vs. 0.54 (PC), 0.62 (NOTEARS)
V03	Physics	Conformal Prediction for Neural PDE Solvers	90.3% coverage, interval width 0.023 vs. MC dropout 0.089
V04	Chemistry	Multi-Agent RL for Chemical Synthesis	84% route success, 4.2 steps vs. 71%, 5.5 (MCTS)
V05	Meta-science	Scaling Laws for Scientific Foundation Models	Domain pre-training benefit $\propto N^{-0.34} \times D^{0.21}$
V06	Healthcare	Federated Learning for Clinical Trials with DP	0.82 AUROC closes 73% of centralization gap; DP costs 1.2 pts
V07	Astronomy	Self-Supervised Pre-training for Astronomical Transients	+5.4 accuracy points overall; +12.1 in 50-shot regime
V08	Climate	Physics-Informed Neural Operators for Climate Downscaling	1.23°C RMSE vs. 1.71°C (U-Net); <0.1% energy violation

I Autonomy and Oversight Operationalization

Autonomy Level Scale (1–5).

1. **Tool**: AI executes a specific, atomic instruction with no discretion. *Example: "Spell-check this paragraph."*
2. **Assistant**: AI suggests options or completions; human selects among them. *Example: "Suggest three alternative phrasings for this sentence."*
3. **Collaborator**: Iterative co-development where both human and AI contribute substantively and refine each other's work. *Example: "Let's develop a hypothesis together—here's my initial idea, what do you think?"*

4. **Delegate**: AI acts with substantial discretion; human reviews the output post-hoc and may accept, modify, or reject. *Example: "Write a draft methods section based on these notes."*
5. **Autonomous**: AI acts independently without human intervention or review. *Example: An autonomous agent that designs and executes experiments.*

Human Oversight Types.

- **Directed**: Human specified the exact task, inputs, and expected output format.
- **Prompted**: Human provided an open-ended prompt; AI determined the approach.
- **Co-developed**: Iterative exchange where human and AI alternated contributions.
- **Reviewed**: Human examined AI output after generation and approved, modified, or rejected it.
- **None**: No human involvement in this contribution event (applies only to human-initiated events, e.g., a researcher typing notes).

Validation Status. These scales are face-valid constructs developed from the levels-of-automation framework [Parasuraman et al., 2000]. Formal inter-annotator agreement studies (Cohen's κ or ICC) across independent research teams have not yet been conducted. In our Study 1, researchers reviewed all 360 auto-assigned role labels post-session and corrected 0, suggesting high face validity for the specific workflows studied. However, we expect disagreements to emerge in more complex, multi-tool agent workflows where boundaries between autonomy levels become ambiguous (e.g., is a chain-of-thought prompt that guides AI step-by-step level 1 or level 3?). A formal validation study with multiple independent annotators across at least 5 research domains is planned.

J CRediT Projection Rules

Every CreditMap AI role maps to exactly one parent CRediT role via the following deterministic projection:

Table 12: Deterministic projection rules from CreditMap to CRediT.

CreditMap Role	$\rightarrow$	**CRediT Role**
AI: Text Generation	$\rightarrow$	Writing – Original Draft
AI: Text Editing	$\rightarrow$	Writing – Review & Editing
AI: Code Generation	$\rightarrow$	Software
AI: Code Debugging	$\rightarrow$	Software
AI: Hypothesis Proposal	$\rightarrow$	Conceptualization
AI: Literature Synthesis	$\rightarrow$	Investigation
AI: Data Analysis	$\rightarrow$	Formal Analysis
AI: Experiment Design	$\rightarrow$	Methodology
AI: Interpretation	$\rightarrow$	Formal Analysis
All 14 standard CRediT roles map to themselves.		

Information Discarded in Projection. The following metadata fields are present in CreditMap but have no CRediT equivalent and are therefore lost:

- Temporal ordering (CRediT roles are unordered sets)
- Autonomy level (1–5 scale)
- Human oversight type (directed/prompted/co-developed/reviewed/none)
- Token counts (input and output)
- Content hashes (artifact integrity)

- Parent artifact references (revision history)
- Contributor type distinction (human vs. ai_system vs. ai_assisted)

Multi-Role Events. The current schema does not support assigning multiple roles to a single API call. If a single prompt requests both literature synthesis and hypothesis generation, the researcher must either (a) assign the primary role and note the secondary role in the description field, or (b) split the event into two sequential events with shared `inputTokens` and linked `parentArtifactId`. Option (b) is preferred for fidelity but increases logging burden. For agentic workflows where a single orchestrator call triggers multiple tool-use steps, each tool invocation should be logged as a separate event with the orchestrator's event as parent. Automated multi-role detection via prompt classification (e.g., a lightweight classifier over the prompt text) is a planned extension, though we anticipate that the primary-role heuristic will suffice for $>90\%$ of typical research interactions based on our case study experience.

Edge Cases.

- *AI: Code Generation* and *AI: Code Debugging* both project to *Software*. This is the single largest expressiveness loss, as constructive and corrective coding are qualitatively different.
- *AI: Data Analysis* and *AI: Interpretation* both project to *Formal Analysis*. Running a statistical test and interpreting its meaning are distinct cognitive tasks.
- Human contributions that use AI as a tool (type: `ai_assisted`) project identically to pure human contributions—the `ai_assisted` distinction is lost.

K Study 1: Case Study Details

Task Domains. Each domain used 15 task prompts (5 per model $\times$ 3 models). Examples:

- **Literature synthesis**: "Summarize the current state of research on AI-assisted drug discovery, focusing on molecular generation approaches published since 2022."
- **Data analysis**: "Analyze the distribution of CRediT role assignments across 100 papers. Generate summary statistics and identify the most commonly claimed roles."
- **Hypothesis generation**: "Propose testable hypotheses about how AI contribution level affects peer review outcomes in ML conferences."

Session Protocol. Each session followed an 8-event protocol:

1. Human defines task (Conceptualization)
2. AI generates initial response (domain-specific AI role)
3. Human reviews AI output (Writing – Review & Editing)
4. AI revises based on feedback (AI: Text Editing)
5. Human approves final output (Writing – Review & Editing)
6. Human supervises session (Supervision)
7. Domain-specific AI step (e.g., AI: Code Generation for data analysis)
8. Domain-specific human step (e.g., Methodology for hypothesis generation)

This fixed protocol enables controlled comparison but limits naturalistic variability. In real research workflows, sessions would have variable length and event composition.

Models Used. GPT-4o (`gpt-4o-2024-08-06`), Claude 3.5 Sonnet (`claude-3-5-sonnet-20241022`), and Gemini 2.5 Flash (`gemini-2.5-flash`). All API calls were made between April 11–12, 2026.

L Power Analysis Details

Post-hoc Power (Frontier Study). With $N = 64$ per condition group (collapsing across AI level) and $\alpha = 0.05$ (two-tailed):

- $d = 0.20$: power $= 0.17$ (underpowered)
- $d = 0.30$: power $= 0.40$ (underpowered)
- $d = 0.50$: power $= 0.81$ (adequately powered)
- $d = 0.80$: power $= 1.00$ (well powered)

The minimum detectable effect at 80% power is $d \approx 0.50$. The observed fairness effect ($d = 2.98$) and trust effect ($d = 0.53$–0.71) both exceed this threshold. The non-significant effects on rigor ($d \approx 0.3$), reproducibility ($d \approx 0.2$), and overall ($d \approx 0.3$) under cluster-robust analysis fall below or near the detection limit.

Sample Size Planning for Human Replication. For a pre-registered human replication targeting the trust effect ($d = 0.53$ from the frontier study), a two-sample t-test at $\alpha = 0.05$ and power $= 0.80$ would require $N = 57$ per group, or $N = 171$ total for three conditions. For the fairness effect ($d = 2.98$), only $N \approx 4$ per group would suffice, though we recommend a minimum of $N = 30$ per cell for distributional stability.