VIBE RESEARCH DRIVES SCIENTIFIC INQUIRY SHALLOWER UNLESS HUMAN JUDGMENT IS SCAFFOLDED

Anonymous authors
Paper under double-blind review

ABSTRACT

As AI systems grow more capable, researchers increasingly collaborate with them in scientific inquiry, a practice we term *vibe research*. This position paper argues that **vibe research drives scientific inquiry shallower unless human judgment is scaffolded**. We identify the root cause as Knowledge-Judgment (K-J) decoupling: at the research frontier, where no established ground truth exists, human judgment is the primary means of evaluating AI-generated knowledge, yet it is fallible, and its errors may go undetected. We propose the Asymmetric Coupling Framework (ACF), which models human–AI research collaboration as two nested loops. An exploratory study provides initial support for the framework's core implication: human judgment, not AI usage volume, determines the quality of AI-assisted research. To address K-J decoupling, we propose *decision readiness scaffolding*, through which AI helps researchers build the evaluative capacity needed at specific decision points. Scaffolding substantially reduces the expertise deficit: all 13 participants working outside their expertise produce ideas that meet or exceed the quality baseline they set within it. These results point to a broader principle: as AI takes a larger role in science, reliable research depends less on more capable AI than on deliberately cultivating the human judgment it cannot replace.

1 INTRODUCTION

AI systems increasingly search literature, generate hypotheses, design experiments, and analyze data in science (Boiko et al., 2023; Bran et al., 2024; Lu et al., 2026; Huang et al., 2025; Mitchener et al., 2025). Yet their outputs can only be validated through the researcher's judgment, and at the research frontier no established ground truth exists as an independent check (Messeri & Crockett, 2024). Consider a computational biologist whose AI, applying a dimensionality-reduction method outside her training to single-cell RNA sequencing data, proposes three novel cell-subtype hypotheses: she cannot tell genuine discoveries from methodological artifacts. Accepted uncritically, errors in such outputs may pass peer review, enter the published record, and distort the knowledge base.

We term this emerging practice *vibe research*, after vibe coding (Sapkota et al., 2025): scientific inquiry in which the complexity of AI-generated outputs may exceed the researcher's ability to fully evaluate them. Its benefits are real: AI lowers the barrier to entry, widens exploration, and accelerates output (Noy & Zhang, 2023). Yet AI outputs are growing in complexity and scope faster than researchers' ability to evaluate them, increasingly even for domain specialists (Kaplan et al., 2020). This paper argues that **vibe research drives scientific inquiry shallower unless human judgment is scaffolded**. The resulting gap is uniquely damaging in science: unlike medicine, engineering, or law, the research frontier has no external oracle to catch errors that human judgment misses, so closing the gap requires not stronger AI but deliberately scaffolded judgment.

Existing human–AI collaboration frameworks do not capture these dynamics. They span AI as a tool under direct human control (Shneiderman, 2020), mixed-initiative systems (Horvitz, 1999), and hybrid intelligence (Dellermann et al., 2019), but all presuppose that the human understands enough to evaluate the AI's output. Tool use assumes the human can verify the tool's output, mixed initiative that the human can absorb each handoff of control, and hybrid intelligence that each side can detect the other's errors. Likewise, levels of automation assume meaningful human oversight at each level (Parasuraman et al., 2000), complementary performance requires discriminating correct from incorrect AI advice (Bansal et al., 2021), and appropriate reliance presupposes the domain knowledge

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

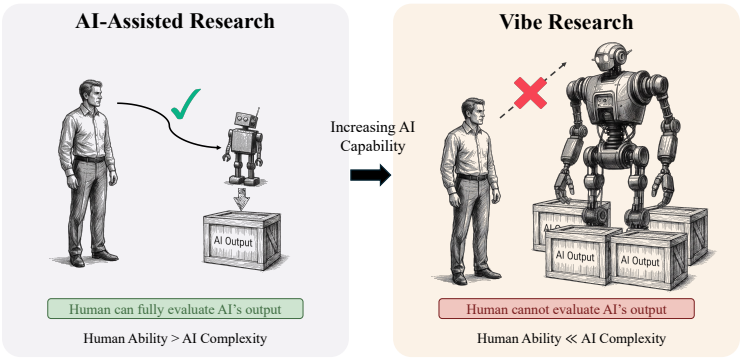


Figure 1: **From AI-assisted research to vibe research.** In AI-assisted research (left), human ability exceeds AI complexity, so researchers can fully evaluate AI outputs and safely absorb their gains. As AI capability increases (right), outputs multiply and outgrow what the researcher can evaluate; they become unverifiable, and the binding constraint shifts from AI ability to human judgment.

to calibrate trust (Lee & See, 2004). When AI outputs systematically exceed human evaluative capacity, this precondition comes under pressure across all paradigms at once. The automation-bias and appropriate-reliance literatures study such failures (Parasuraman & Manzey, 2010; Schemmer et al., 2023), but model reliance as calibrated through observed errors; at the research frontier this error signal never arrives, so miscalibrated reliance becomes a structural condition rather than a correctable bias. Scalable oversight (Bowman et al., 2022) questions the same precondition for the training-time supervision of AI; we address scientific practice, where undetected judgment errors enter the scientific record.

We identify the root cause as *Knowledge-Judgment decoupling* (K-J decoupling): AI's capacity for knowledge production increasingly outpaces human evaluative judgment, so the capability that makes AI useful also undermines the human's ability to verify its outputs. We formalize this with the Asymmetric Coupling Framework (ACF), which models the interaction as two nested loops and explains why the gap is structural and widens with AI capability. Under the ACF, we show theoretically and empirically that human judgment, not AI usage volume, determines the quality of AI-assisted research, and that removing human judgment through autonomous AI research fails on precisely the dimensions that require it. To address K-J decoupling, we propose *decision readiness scaffolding*, through which AI helps researchers build the evaluative capacity needed at specific decision points. Scaffolding substantially reduces the expertise deficit: all 13 participants working outside their expertise produce ideas matching or exceeding the quality baseline they set within it.

2 The Rise of Vibe Research

Definition. We define vibe research as scientific inquiry in which the researcher articulates intent at a high level of abstraction while AI systems handle technical execution (Feng & Liu, 2026). It differs from conventional AI tool use in three ways: (1) the interaction is intent-driven, with the scientist specifying what to investigate and the AI determining how; (2) execution is delegated while judgment is retained, with the researcher exercising domain judgment over every AI-generated artifact (Zhang, 2026); and (3) the researcher's contribution lies in process quality (how questions are framed, outputs evaluated, and refinement steered) rather than in implementation. Vibe research thus sits between AI-assisted research, where the scientist controls code and parameters (Schmidgall et al., 2025), and autonomous AI research with minimal human oversight (Zheng et al., 2025).

Benefits. AI lowers the barrier to entry, letting researchers without specialized programming or statistical expertise run analyses that once required years of training (Noy & Zhang, 2023). It broadens exploration by generating multiple framings and hypotheses at decision points, surfacing options that conventional literature review may miss (Si et al., 2024). Recent work pushes generation further: modeling the distribution of human expertise lets AI anticipate future discoveries and steer toward promising hypotheses that humans are unlikely to reach (Sourati & Evans, 2023), and AI agents are charted as a route to biomedical discovery with human expertise kept in the loop (Gao et al.,

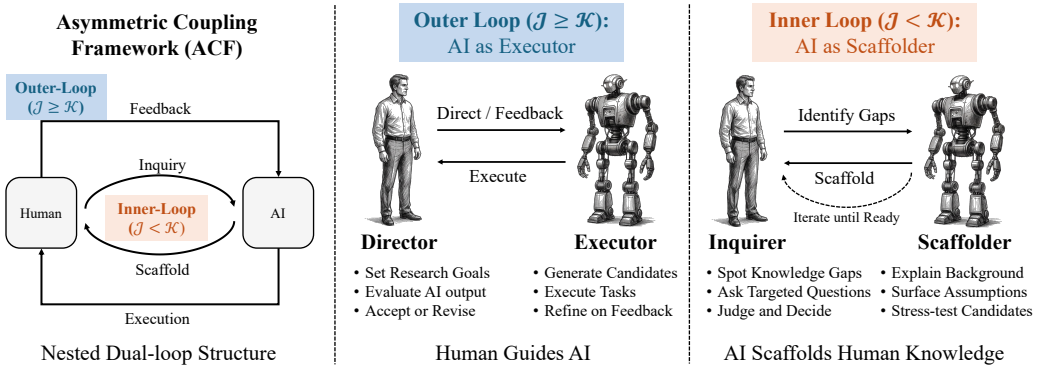


Figure 2: **The Asymmetric Coupling Framework (ACF).** (a) Nested dual-loop structure: the outer loop operates when human judgment suffices ($J \geq K$); the inner loop activates when it does not ($J < K$). (b) Outer loop: the human acts as *director*, setting research goals, evaluating AI output, and accepting or revising it, while AI acts as *executor*, generating candidates, executing tasks, and refining on feedback. (c) Inner loop: the human becomes an *inquirer* who spots knowledge gaps and asks targeted questions, while AI becomes a *scaffolder* that explains background, surfaces assumptions, and stress-tests candidates, iterating until the human is ready to judge and decide.

2024). It accelerates output through rapid iteration and can act as a cognitive challenger, proposing counterexamples and critiques that stress-test the researcher's reasoning (Chiang et al., 2024).

The Trap. The same benefits create a systematic vulnerability. Lowered barriers draw in researchers who lack the capacity to judge AI outputs, increasing the volume of judgments made under incomplete understanding (Dell'Acqua et al., 2023; Mahajan, 2025). Broader exploration yields more candidates per decision point, but evaluative capacity does not scale accordingly (Lahlou, 2025; Parasuraman & Manzey, 2010). Faster iteration leaves less time to separate sound conclusions from plausible but unverified ones. And an AI challenger forces the researcher to evaluate the critique itself, which demands the very capacity it is meant to supplement. Each benefit thus raises the demand for human judgment without raising the supply. Advances in hypothesis generation sharpen this imbalance: sampling more novel ideas enlarges what must be judged while leaving evaluation unchanged, raising the risk that AI-assisted science produces more while understanding less (Messeri & Crockett, 2024); our framework addresses this evaluation-side complement to the generation literature. In science the imbalance is especially dangerous, because no established ground truth at the frontier catches errors that escape human judgment (Messeri & Crockett, 2024). Domain experts can compensate with accumulated understanding in their own fields, but vibe research routinely draws researchers into adjacent or unfamiliar areas where that judgment does not apply (Dell'Acqua et al., 2023; Lin & Sohail, 2026). As AI capability grows, even specialists may face outputs in their own domains that exceed their evaluative capacity: methods too novel to assess, reasoning chains too long to verify, or conclusions resting on literatures no single researcher has surveyed (Romeo & Conti, 2026).

The Vibe Research Trap: The very benefits that make vibe research attractive (lower barriers, broader exploration, faster output) also lead researchers to make judgments they are not equipped to make. The more capable the AI, the more compelling these benefits, and the further researchers venture beyond their evaluative reach.

3 The Knowledge-Judgment Gap

This section examines the structural basis of this evaluative burden and tests its implications.

3.1 Knowledge-Judgment Decoupling

A human expert delivers knowledge and judgment as a coupled package: "use method A, not method B" conveys both what methods exist and which is appropriate. This coupling is earned

through years of failed experiments, peer critique, and accumulated domain intuition (Ericsson et al., 1993), so a less experienced collaborator can rationally trust the package without evaluating each component. AI also provides both knowledge and judgment, recommending methods, suggesting analyses, and offering evaluations, but the two are structurally decoupled.

> **Knowledge-Judgment Decoupling**: AI acquires knowledge by training on text; its judgment emerges from pattern matching over that text rather than from feedback grounded in real-world consequences. Even an AI that never produced factual errors would exhibit K-J decoupling, because its judgment lacks the basis of human expertise: no track record of consequences, no calibrated uncertainty from experience, no domain intuition forged through practice.

The decoupling creates an asymmetric evaluation burden: whether AI provides knowledge or judgment, the human must evaluate it independently, a task that in human–human collaboration can be delegated to the expert. Worse, the decoupling is hidden: human experts signal the boundary between established knowledge and provisional judgment through hedging, deferral, and qualification, whereas AI presents both in the same authoritative register, with no reliable signal separating well-grounded knowledge from pattern-derived inference (Ji et al., 2023).

3.2 The Asymmetric Coupling Framework

The ACF formalizes how K-J decoupling operates in research and why better AI alone cannot resolve it, modeling human–AI collaboration as two nested loops (Figure 2). In the outer loop, AI executes tasks (searching literature, running analyses, generating hypotheses), the human evaluates the outputs (whether the direction is sound, the methodology appropriate, the conclusions warranted), and AI continues on that feedback. This loop presupposes that the human's judgment capacity (J) suffices to evaluate AI's knowledge output (K). When it does not, the outer loop stalls at a decision point and an inner loop activates: AI shifts from execution to scaffolding, providing background knowledge, explaining alternatives, exposing assumptions and risks, and decomposing the evaluation into verifiable components, until the researcher can judge and the outer loop resumes.

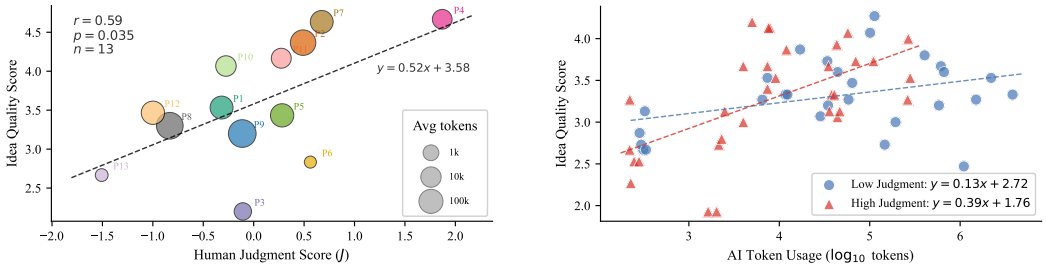
The two loops show why K-J decoupling is structural rather than transient. AI's epistemic capacity grows with scaling in data, model size, and reasoning (Kaplan et al., 2020), whereas human evaluative capacity grows at the pace of experience, requiring time, domain immersion, and repeated failure (Ericsson et al., 1993). As AI improves, its outputs exceed human evaluative capacity more often, and the inner loop must activate more regularly.

> **The Asymmetric Scaling Hypothesis**: Stronger AI does not close the K-J gap but widens it, not because AI is inadequate but because it is increasingly capable. As K grows faster than J, the outer loop stalls more often, and the inner loop becomes a necessary condition, not an optional supplement, for reliable vibe research.

Scope of the Hypothesis. The hypothesis is bounded by the availability of a verifiable signal. Where outputs can be checked against clear answers (code under tests, checkable mathematics), verification both catches errors and serves as a reward that lets models improve on their own (DeepSeek-AI et al., 2025); there, human judgment is not the binding constraint, and the hypothesis does not apply. Where no clear answer exists (which problem to pursue, whether a framing is sound or an interpretation follows), the same absence blocks both error detection and reward-driven self-improvement, so capability gains concentrate in verifiable stages and the gap widens in unverifiable ones. Mixed stages split accordingly: tests verify an implementation but not the adequacy of its specification, and citations verify provenance but not inferential support.

3.3 Implications and Empirical Evidence

We test the ACF's implications in an exploratory study in which 13 AI researchers each produce about five research ideas through free interaction with an AI system (68 ideas; Appendix A); the same participants join the scaffolding experiment of §4.3. Judgment capacity (J) combines two validated self-report instruments, the Self-Efficacy in Research Measure–Brief (Kahn & Scott, 1997) and the Actively Open-Minded Thinking scale (Stanovich & Toplak, 2023), as $J = -(z(\text{SERM-B}) + z(\text{AOT-13}))/2$. The ACF characterizes judgment at a no-ground-truth frontier by two capacities:



(a) Higher J predicts higher quality.

(b) K amplifies J: high-J benefits more from AI.

Figure 3: Empirical evidence for K-J decoupling ($n = 13$ participants, 68 ideas). (a) Participant-level quality against J; bubble size shows average AI token usage, and the dashed line is the least-squares fit. (b) Idea-level quality against AI token usage, split at the median of J, with a linear fit per group. Human judgment (J) predicts research quality more strongly than AI usage does; AI usage (K) provides more benefit when J is higher, consistent with the Asymmetric Scaling Hypothesis.

calibrated self-assessment, and resistance to fluent persuasion, since AI presents all output in one authoritative register (§3.1). Both instruments read these capacities inversely: SERM-B measures confidence in one's research skills, which runs opposite to calibration (Kruger & Dunning, 1999), and AOT-13 measures readiness to revise beliefs on external input, which runs opposite to the filter needed against fluent AI output (Parasuraman & Manzey, 2010). J is therefore the negative composite; the sign is fixed by the construct rather than estimated from the data. Each component correlates negatively with quality on its own, and reversing the orientation changes only the sign of the correlation, not its magnitude or significance (Appendix A.2). AI usage (K) is the mean number of interaction tokens per participant ($\log_{10}$-transformed).

Implication 1: J Determines Quality. The outer loop advances through alternating AI execution and human evaluation, so its reliability depends on J: a high-J researcher can identify promising directions, reject flawed reasoning, and steer productively, whereas a low-J researcher cannot distinguish good outputs from bad, and output quality becomes effectively random with respect to the AI's contribution. Variation in J should therefore predict variation in output quality, even when all researchers use the same AI system. Three independent ranking agents comparatively score each participant's strongest contribution (1–5), and the average serves as participant-level quality. Figure 3a is consistent with the prediction: J correlates significantly with quality ($r = 0.59$, $p = 0.035$, $n = 13$; permutation $p = 0.034$; Spearman $\rho = 0.51$, $p = 0.078$), and the highest-quality participants are not necessarily the heaviest AI users (bubble size).

Implication 2: The Value of K Depends on J. K determines how much material AI generates; J determines whether the researcher can leverage it. Increasing K should therefore benefit high-J researchers, who can evaluate and build on more output, far more than low-J researchers, for whom more output means more material they cannot evaluate. Using idea-level scores (five dimensions averaged over three AI reviewer agents) and a median split on J, Figure 3b is consistent with this prediction: quality rises steeply with token usage in the high-J group (slope $= 0.39$) but much less in the low-J group (slope $= 0.13$). With 13 participants, this contrast is descriptive rather than a formally tested interaction. In this sample, more AI interaction appears to amplify strong judgment rather than compensate for weak judgment. Similarly, across 160,097 GitHub developers, senior programmers captured nearly all AI productivity gains, while early-career programmers saw no significant benefit despite heavier AI use (Daniotti et al., 2026).

Implication 3: Removing J Entirely Fails. If AI handles both execution and evaluation, the outer loop loses its primary verification mechanism. The ACF predicts that surface dimensions such as clarity and presentation, which need no external evaluation, should hold up, whereas judgment-dependent dimensions (whether a contribution is original, addresses a significant problem, and advances the field) should degrade systematically. We compare 12 papers co-authored by humans and AI (vibe research) with 37 papers generated entirely by five autonomous AI systems (Appendix B; Figure 4). The results match the prediction: autonomous systems perform comparably on presentation and clarity but significantly worse on originality ($p < 0.001$), contribution ($p = 0.004$), and

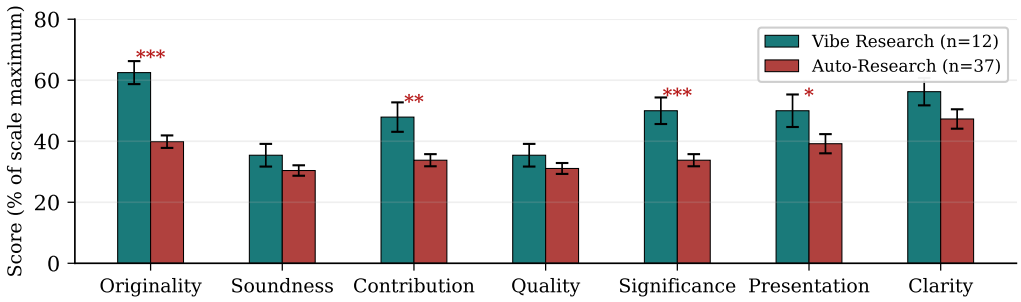


Figure 4: Removing human judgment degrades research quality on judgment-requiring dimensions. Mean review scores of papers from vibe research ($n = 12$) and autonomous AI research systems ($n = 37$), scored by the same automated reviewer (Appendix B) as a percentage of each dimension's maximum; error bars are standard errors. Stars mark one-sided Mann–Whitney U tests: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

significance ($p < 0.001$). Although this comparison evaluates full papers with a standard review rubric rather than ideas, the pattern is consistent with Implications 1 and 2.

4 Scaffolding Human Judgment

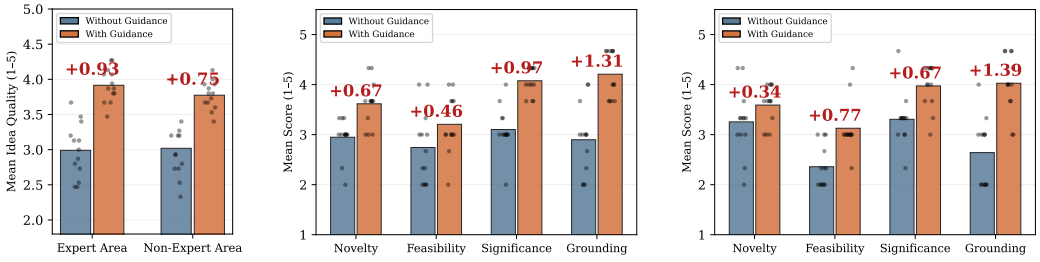
4.1 Decision Readiness Scaffolding

When the outer loop stalls, the ACF's inner loop activates: AI shifts from executing research tasks to building the human's understanding until a judgment can be made. We call this mechanism *decision readiness scaffolding*. Instead of a direct judgment ("use method A"), AI decomposes the evaluation into components the researcher can verify: background on each candidate method, the assumptions each makes, the risks of each choice, and criteria for comparing alternatives. Although AI's high-level judgments may be unreliable (the premise of K-J decoupling), these components are lower-level claims the human can check against textbooks, datasets, or domain knowledge, so scaffolding turns one difficult judgment into a series of easier, verifiable steps. Without it, a researcher facing an output beyond their evaluative capacity can only accept it uncritically, propagating undetected errors, or abandon the direction, forfeiting the benefits of vibe research. This matches §3.3: researchers with low J benefited much less from increased AI usage (K), as expected when no structured mechanism exists to convert K into evaluative understanding.

The concept draws on Vygotsky's Zone of Proximal Development (Vygotsky, 1978), with one key distinction: classical scaffolding aims for durable competence and is removed once the skill is learned (Wood et al., 1976), whereas decision readiness scaffolding aims for point-wise readiness, sufficient understanding for a specific judgment at a specific decision point. A new decision point involving different methods or knowledge may require scaffolding from scratch; this reflects the nature of research, where each new problem demands new understanding and the inner loop must be ready to activate repeatedly.

4.2 From Divergence to Judgment: A Four-Stage Scaffolding Workflow

To operationalize the mechanism, we propose a four-stage workflow in which AI's role shifts from knowledge generation to judgment scaffolding (Appendix C.2): (1) diverge, (2) structure, (3) scaffold and stress-test, and (4) human decides. In Stages 1–2, AI works mainly in the outer loop, opening the problem space with candidate directions and organizing vague options into comparable objects along uniform dimensions such as core assumptions, potential value, and main risks; the human sets boundaries and comparison criteria without committing to a choice. In Stages 3–4, the inner loop takes over: AI becomes a challenger, explaining background knowledge, exposing trade-offs, surfacing hidden assumptions, and stress-testing each candidate from the opposing side. The human builds understanding not by being told which option is best but by seeing where each option's



(a) Overall effect (b) Expert area dimensions (c) Non-expert area dimensions

Figure 5: Effect of *decision readiness scaffolding* on idea quality ($n = 13$). (a) Mean overall quality by expertise area and guidance condition; dots are participants, and red numbers give gains from guidance. (b, c) Per-dimension scores and gains in the expert and non-expert areas.

strengths and weaknesses lie, iterates within Stage 3 until an informed judgment is possible, and then makes the final decision and bears responsibility for it.

The Scaffolding Principle: AI usage can remain high throughout, but AI's role must shift from generating conclusions to building the human's evaluative capacity. Divergence is delegated to AI (K released), understanding is built through scaffolding ($K \rightarrow J$ conversion), and judgment must come from the human (J irreplaceable).

For the computational biologist of §1, the workflow replaces a recommendation with a structured comparison of applicable methods and their assumptions (Stages 1–2), an explanation of where each method's assumptions might break down on her data and what artifacts could result (Stage 3), and her own informed choice (Stage 4): the same researcher, AI, and data, but AI now builds her evaluative capacity instead of providing answers.

4.3 Empirical Observations

Design. Thirteen AI researchers each complete four ideation tasks in a 2×2 design crossing domain expertise (expert vs. non-expert area) with guidance (free AI interaction vs. four-stage scaffolding), yielding 52 ideas; task order is counterbalanced to mitigate learning effects. AI reviewer agents independently score each idea on novelty, feasibility, significance, and grounding (1–5; Appendix C).

Finding 1: Scaffolding Improves Quality Across Conditions and Dimensions. Scaffolding improves idea quality in both the expert area (+0.93, from 2.99 to 3.92; $t = 6.74$, $p < 0.001$) and the non-expert area (+0.75, from 3.02 to 3.77; $t = 7.14$, $p < 0.001$), with a pooled Cohen's $d = 1.94$ (Figure 5). By dimension (Figures 5b, 5c), expert-area gains are largest on significance (+0.97) and novelty (+0.67), the dimensions most closely tied to evaluative judgment, whereas non-expert-area gains are largest on feasibility (+0.77), followed by significance (+0.67) and novelty (+0.34). Scaffolding thus appears to address the most pressing gap first: with domain knowledge, it strengthens the higher-order judgment behind significance and novelty; without it, it first builds the foundational understanding needed for feasible work.

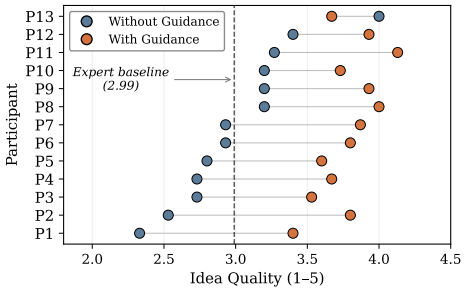


Figure 6: Non-expert area: per-participant quality without (blue) and with (orange) guidance. The dashed line marks the expert baseline (2.99), the mean unguided quality in participants' own areas: 6 of 13 exceed it without guidance, all 13 with guidance.

Finding 2: With Guidance, All Participants Exceed the Unguided Expert-Area Baseline. To test whether scaffolding closes the expertise gap for individuals, we take as expert baseline the mean quality participants achieve in their own area without guidance (2.99). In the non-expert area (Figure 6), only 6 of 13 participants exceed this baseline without guidance, but all 13 do with

guidance: with scaffolding, participants working outside their expertise meet or exceed the quality they achieve within it unaided.

Finding 3: Mechanism-Consistent Signatures. The gain is not explained by length: scaffolded ideas are shorter (794 vs. 872 words), and the guidance–quality association survives partialing out length (partial $r = 0.84$; raw $r = 0.80$). Scaffolding also nearly halves the spread across participants (pooled SD 0.40 to 0.23; Levene $p = 0.010$) and raises the lowest participant-condition score from 2.33 to 3.40. A uniform benefit of structure would shift scores without narrowing their spread, whereas compression is expected if scaffolding supplies evaluative capacity where it is lacking; because regression to the mean can also compress variance and self-reported J does not predict individual lift ($r \leq 0.14$), we treat this as consistent with the mechanism, not as proof. Even with Finding 1 (gains land on each condition's deficit) and Implication 3 (judgment-dependent dimensions degrade without human judgment), the evidence remains indirect: the experiment measures outcomes rather than the change in judgment, and time-on-task was not logged. §5.4 outlines the direct test; Appendix C.4 gives supplementary analyses.

5 Discussion

5.1 Limitations

The empirical findings have limited generalizability. Samples are small (13 participants and 68 ideas in §3.3; 13 participants and 52 ideas in §4.3); both experiments address only ideation, among the most judgment-demanding stages, and the K-J gap may be less severe where verification is easier (e.g., traceable citations, automated tests); and all participants are computer scientists. Larger replications across research stages and disciplines are needed before these patterns can be considered robust. Internally, J relies on self-report scales (SERM-B and AOT-13) subject to self-assessment bias; a behavioral measure is the planned next step (§5.4), and AI reviewer agents judge AI-assisted output, a potential circularity. Time-on-task was not logged in the scaffolding experiment, so part of the gain may reflect effort, although not output length (§4.3).

5.2 What Would Count Against the Position

The position is stated so that it can fail. First, scaffolding is defined ex ante, not by outcomes: an intervention qualifies only if it contains the components of §4.2 (forced divergence, structured comparison, adversarial stress-testing, and a human-owned decision). An intervention lacking them that produces the same gains is therefore a counterexample, not "implicit scaffolding," and failure under the full protocol counts against the claim, not as "insufficient scaffolding." Second, the claim is scoped to verification-poor stages (§3.2); it asserts neither that scaffolding is the only route to quality nor that AI use is harmful. Third, "shallower" is measurable: quality should track judgment rather than usage volume, and degradation should concentrate on judgment-dependent dimensions (novelty, significance, grounding) rather than surface ones (presentation, clarity). The position would thus be weakened if a structure-matched control with generic content matched the gains of full scaffolding, if the full protocol failed to lift judgment-dependent dimensions at verification-poor stages, or if usage volume rather than judgment predicted quality once AI outputs exceed the researcher's expertise.

5.3 Alternative Views

Four objections merit response. (1) AI judgment may eventually match human judgment; even so, the human must still decide whether to trust the AI's judgment in each case, which itself requires evaluative capacity, so better AI judgment shifts the verification problem rather than eliminating it. (2) Scaffolding may seem to reduce to prompt engineering; but prompt engineering optimizes AI output, whereas scaffolding targets human evaluative capacity, so the two are complementary, not equivalent. (3) Science has always trusted tools it does not fully understand; but such tools have well-characterized error models (a mass spectrometer's precision is known and bounded), whereas LLMs fail in unpredictable, content-dependent ways without reliable confidence signals (Ji et al., 2023), making K-J decoupling qualitatively different. (4) The bottleneck may be workflow design (provenance tracking, reproducibility checks, validation tooling) rather than judgment; we see the two as complementary layers. Such tools govern the verifiable part of the pipeline, establishing

where a result came from and whether it can be regenerated, but at frontier decision points (which problem to pursue, whether a framing is sound, whether an interpretation follows) they have no ground truth to check against: perfect provenance tells us where a conclusion came from, not whether it is right. Better workflows narrow the surface on which unaided judgment operates but cannot close the evaluative gap where our position locates the risk.

5.4 A Research Agenda

The ACF separates two levels of claim. At the framework level, wherever AI assists research without ground truth, progress combines divergence opened by AI with convergence by a human scaffolded to decision readiness, and these two elements are load-bearing in any domain; the four-stage workflow is one instantiation for ideation, and which of its stages carry the effect is open to ablation. Four contextual dimensions structure the design space: the *expertise gap* sets where the inner loop starts and how much K must be converted into J (the gain profiles in §4.3 track each condition's deficit); *verifiability* sets how much evaluation can exit to external checks (§3.2); *stakes* set how much inner-loop iteration should precede a decision; and the *domain* changes the content of the loops, not their structure (Appendix D lists open questions for each). We see two community artifacts, in order: first, a *protocol* for AI-assisted work at verification-poor stages, whose core rule is that researchers deliberately check whether they understand what the AI says before accepting it, with the four-stage workflow as a first version to be tested against a structure-matched control; second, a *decision-readiness benchmark* measuring whether researchers can judge frontier outputs, so that tools and training can be compared on the evaluative capacity they build rather than the output they produce. Our next study reads the middle of the causal chain directly: participants blind-judge pre-scored ideas before and after scaffolding, so the change in judgment accuracy is measured rather than inferred, and this behavioral measure replaces the self-report J.

5.5 Broader Implications

Individual Level: Empowerment or Deskilling? Decision readiness scaffolding works in the short term, but does it build capability or create dependence over the long term? Vygotsky's scaffold is removed once the learner internalizes *durable competence*, whereas decision readiness scaffolding targets *point-wise readiness*. As AI grows stronger and scaffolding more frequent, do researchers grow through each episode, or come to depend on scaffolding for any judgment at all? This question requires longitudinal investigation.

Training Level: Research Education. Research judgment is traditionally formed as a byproduct of execution rather than taught directly (Ericsson et al., 1993): tuning hyperparameters builds intuition about model behavior, writing data pipelines reveals data characteristics, and unproductive directions teach what separates promising approaches from dead ends. When AI executes, students may produce competent-looking research faster without forming the judgment to evaluate it; if scaffolded capacity stays point-wise, doctoral training must ask how judgment will form once execution no longer supplies the friction.

Systemic Level: Future Research Workflows. As AI takes on more execution, the researcher's contribution shifts toward evaluative and directional judgment: which questions to pursue, whether results are valid, and how far claims extend. This raises open questions about assessing and attributing individual contributions when execution is shared with AI, with implications for evaluation and peer review. Reproducibility is also at stake (Baker, 2016): when critical decisions emerge from iterative human–AI dialogue rather than deterministic procedures, conventional methods sections may not document the process, and new forms of methodological reporting may be needed.

Beyond Science: Cross-Domain Applicability. K-J decoupling arises wherever AI outputs may exceed the decision-maker's evaluative capacity: physicians judging AI-recommended treatments (Rajpurkar et al., 2022), lawyers assessing AI-drafted arguments, and policymakers evaluating AI-generated recommendations. These domains have partial verifiable anchors (diagnostic procedures, case law, historical data and simulations) that may make scaffolding easier to implement; science is the extreme case, because the thing to be discovered is the thing to be verified. The core structure is the same: if the Asymmetric Scaling Hypothesis holds, AI's knowledge production outpaces human judgment, and the gap is structural.

AI USE STATEMENT

In this work, we used generative AI tools only for language polishing, improving the grammar, wording, and clarity of text drafted by the authors. We have not used generative AI tools for research ideation, for the design of the proposed method or the experimental protocol, for implementation, for the generation or analysis of results, for the creation of figures and tables, or for literature discovery and citation; all of these are the original work of the authors. We have reviewed all AI-assisted work, checking every suggested revision against the authors' original text and the underlying results, and we take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

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

Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–16, New York, NY, USA, 2021. ACM. doi: 10.1145/3411764.3445717.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 41–48, 2009.

Daniil A. Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023. doi: 10.1038/s41586-023-06792-0.

Samuel R. Bowman, Jeeyoon Hyun, Ethan Perez, Edwin Chen, Craig Pettit, Scott Heiner, Kamilė Lukošiūtė, Amanda Askell, Andy Jones, Anna Chen, et al. Measuring progress on scalable oversight for large language models. *arXiv preprint arXiv:2211.03540*, 2022.

Andrés M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D. White, and Philippe Schwaller. ChemCrow: Augmenting large-language models with chemistry tools. *Nature Machine Intelligence*, 6:525–535, 2024. doi: 10.1038/s42256-024-00832-8.

Chun-Wei Chiang, Zhuoran Lu, Zhuoyan Li, and Ming Yin. Enhancing AI-assisted group decision making through LLM-powered devil's advocate. In *Proceedings of the 29th International Conference on Intelligent User Interfaces (IUI)*, pp. 103–119, 2024.

Simone Daniotti, Johannes Wachs, Xiangnan Feng, and Frank Neffke. Who is using ai to code? global diffusion and impact of generative ai. *Science*, pp. eadz9311, 2026.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and Xiaokang Zhang. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *Nature*, 645:633–638, 2025. doi: 10.1038/s41586-025-09422-z.

Fabrizio Dell'Acqua, Edward McFowland III, Ethan R Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R Lakhani. Navigating the jagged technological frontier: Field experimental evidence of the effects of ai on knowledge worker productivity and quality. *Harvard business school technology & operations mgt. Unit working paper*, (24-013), 2023.

Dominik Dellermann, Philipp Ebel, Matthias Söllner, and Jan Marco Leimeister. Hybrid intelligence. *Business & Information Systems Engineering*, 61(5):637–643, 2019. doi: 10.1007/s12599-019-00595-2.

K. Anders Ericsson, Ralf Th. Krampe, and Clemens Tesch-Römer. The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, 100(3):363–406, 1993.

Yebo Feng and Yang Liu. A visionary look at vibe researching. *arXiv preprint arXiv:2604.00945*, 2026.

Leisa Reinecke Flynn and Ronald E Goldsmith. A short, reliable measure of subjective knowledge. *Journal of Business Research*, 46(1):57–66, 1999.

Shane Frederick. Cognitive reflection and decision making. *Journal of Economic Perspectives*, 19(4): 25–42, 2005.

Shanghua Gao, Ada Fang, Yepeng Huang, Valentina Giunchiglia, Ayush Noori, Jonathan Richard Schwarz, Yasha Ektefaie, Jovana Kondic, and Marinka Zitnik. Empowering biomedical discovery with AI agents. *Cell*, 187(22):6125–6151, 2024. doi: 10.1016/j.cell.2024.09.022.

Eric Horvitz. Principles of mixed-initiative user interfaces. In *Proceedings of the 1999 CHI Conference on Human Factors in Computing Systems*, pp. 159–166, New York, NY, USA, 1999. ACM. doi: 10.1145/302979.303030.

Kexin Huang, Serena Zhang, Hanchen Wang, Yuanhao Qu, Yingzhou Lu, Yusuf Roohani, Ryan Li, Lin Qiu, Gavin Li, Junze Zhang, et al. Biomni: A general-purpose biomedical ai agent. *biorxiv*, 2025.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Delong Chen, Wenliang Dai, Ho Shu Chan, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023. doi: 10.1145/3571730.

Jeffrey H Kahn and Norman A Scott. Predictors of research productivity and science-related career goals among counseling psychology doctoral students. *The Counseling Psychologist*, 25(1):38–67, 1997.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Justin Kruger and David Dunning. Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, 77(6):1121–1134, 1999. doi: 10.1037/0022-3514.77.6.1121.

Salem Lahlou. Mitigating societal cognitive overload in the age of ai: Challenges and directions. *arXiv preprint arXiv:2504.19990*, 2025.

John D. Lee and Katrina A. See. Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1):50–80, 2004. doi: 10.1518/hfes.46.1.50_30392.

Zhicheng Lin and Aamir Sohail. Recalibrating academic expertise in the age of generative ai. *Patterns*, 7(1), 2026.

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.

Chris Lu, Cong Lu, Robert Tjarko Lange, Yutaro Yamada, Shengran Hu, Jakob Foerster, David Ha, and Jeff Clune. Towards end-to-end automation of AI research. *Nature*, 651(8107):914–919, 2026. doi: 10.1038/s41586-026-10265-5.

Sachit Mahajan. The democratization dilemma: When everyone is an expert, who do we trust? *Humanities and Social Sciences Communications*, 12(1):1–5, 2025.

Lisa Messeri and M. J. Crockett. Artificial intelligence and illusions of understanding in scientific research. *Nature*, 627(8002):49–58, 2024. doi: 10.1038/s41586-024-07146-0.

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

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 and Dietrich H Manzey. Complacency and bias in human use of automation: An attentional integration. *Human factors*, 52(3):381–410, 2010.

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.

William Peebles and Saining Xie. Scalable diffusion models with transformers. *arXiv preprint arXiv:2212.09748*, 2023.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben Allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The FineWeb datasets: Decanting the web for the finest text data at scale. *arXiv preprint arXiv:2406.17557*, 2024.

Pranav Rajpurkar, Emma Chen, Oishi Banerjee, and Eric J. Topol. AI in health and medicine. *Nature Medicine*, 28(1):31–38, 2022. doi: 10.1038/s41591-021-01614-0.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019.

Giuseppe Romeo and Daniela Conti. Exploring automation bias in human–ai collaboration: a review and implications for explainable ai. *Ai & Society*, 41(1):259–278, 2026.

Ranjan Sapkota, Konstantinos I Roumeliotis, and Manoj Karkee. Vibe coding vs. agentic coding: Fundamentals and practical implications of agentic ai. *arXiv preprint arXiv:2505.19443*, 2025.

Max Schemmer, Niklas Kuehl, Carina Benz, Andrea Bartos, and Gerhard Satzger. Appropriate reliance on AI advice: Conceptualization and the effect of explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces (IUI '23)*, pp. 410–422, New York, NY, USA, 2023. ACM. doi: 10.1145/3581641.3584066.

Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Michael Moor, Zicheng Liu, and Emad Barsoum. Agent laboratory: Using LLM agents as research assistants. *arXiv preprint arXiv:2501.04227*, 2025.

Ben Shneiderman. Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction*, 36(6):495–504, 2020. doi: 10.1080/10447318.2020.1741118.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can LLMs generate novel research ideas? A large-scale human study with 100+ NLP researchers. *arXiv preprint arXiv:2409.04109*, 2024.

Jamshid Sourati and James A. Evans. Accelerating science with human-aware artificial intelligence. *Nature Human Behaviour*, 7(10):1682–1696, 2023. doi: 10.1038/s41562-023-01648-z.

Keith E Stanovich and Maggie E Toplak. Actively open-minded thinking and its measurement. *Journal of Intelligence*, 11(2):27, 2023.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

Lev S. Vygotsky. *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press, Cambridge, MA, 1978.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

David Wood, Jerome S. Bruner, and Gail Ross. The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry*, 17:89–100, 1976. doi: 10.1111/j.1469-7610.1976.tb00381.x.

Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. *arXiv preprint arXiv:2212.11565*, 2023.

Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy Liang, Quoc V Le, Tengyu Ma, and Adams Wei Yu. DoReMi: Optimizing data mixtures speeds up language model pretraining. In *Advances in Neural Information Processing Systems*, 2023.

Yongjun Zhang. Vibe researching as wolf coming: Can AI agents with skills replace or augment social scientists? *arXiv preprint arXiv:2602.22401*, 2026.

Tianshi Zheng, Zheye Deng, Hong Ting Tsang, Weiqi Wang, Jiaxin Bai, Zihao Wang, and Yangqiu Song. 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 (EMNLP)*, 2025.

A K-J Decoupling Study: Detailed Methodology and Experiments

A.1 Study Design, Participants, and Task

This exploratory observational study tests the ACF's core prediction: AI usage volume (K) cannot compensate for insufficient human judgment ability (J). Thirteen participants (P1–P13) in AI research, ranging from doctoral students to researchers with publication records, each produced approximately five original research ideas through free interaction with Claude Opus 4.6. No constraints were placed on interaction strategy or volume. All ideas fell within AI/ML research and included a title, motivation, proposed method, and expected contribution. Participants worked independently. Twelve participants produced 5 ideas each; one (P9) produced 8, yielding 68 ideas total.

A.2 Measuring Human Judgment Ability (J)

The ACF identifies two facets of evaluative capacity that are theoretically relevant to judgment quality in AI-assisted research: (1) calibrated self-assessment of one's own research abilities, and (2) deliberateness in belief revision under external input. We selected one validated instrument for each facet:

- **Self-Efficacy in Research Measure—Brief** (SERM-B (Kahn & Scott, 1997)): 12 items on a 0–9 Likert scale assessing confidence in specific research skills (experimental design, hypothesis formulation, statistical interpretation, scientific writing, and oral defense). Score: mean of all 12 items. This instrument targets facet (1): researchers who rate their abilities more conservatively—consistent with the calibration asymmetry documented by Kruger & Dunning (1999)—may exercise more careful judgment when evaluating AI output.
- **Actively Open-Minded Thinking Scale** (AOT-13 (Stanovich & Toplak, 2023)): 13 items on a 1–6 Likert scale measuring disposition toward evidence-based belief revision. Eight items are reverse-coded ($7 - \text{raw}$). Score: mean of all 13 items after reverse coding. This instrument targets facet (2): lower scores indicate a disposition to scrutinize arguments rather than reflexively accept them, which may serve as a stronger filter when evaluating AI-generated proposals.

The composite judgment score is defined as:

$$J = -\big(z(\text{SERM-B}) + z(\text{AOT-13})\big) \,/\, 2$$

where $z(\cdot)$ denotes standardization across the 13 participants ($r = 0.59$, $p = 0.035$).

Two additional instruments were administered but do not map onto either facet. The Cognitive Reflection Test (CRT) (Frederick, 2005) measures the tendency to override intuitive responses with deliberative reasoning—a general cognitive trait rather than a facet of research-specific evaluative capacity; it also exhibited a ceiling effect in this sample (12 of 13 participants scored 3/3). The Subjective Knowledge Scale (SKS) (Flynn & Goldsmith, 1999) measures self-perceived domain knowledge ("how much do you know about AI"), which corresponds to K (knowledge) rather than J (judgment) in the ACF framework—and its lack of association with idea quality ($r = 0.12$, $p = 0.69$) is itself consistent with K-J decoupling. Table 1 reports correlations for all candidate composites as a robustness check.

Table 1: Robustness check: correlations between alternative J composites and participant-level idea quality ($n = 13$). The theoretically motivated composite is in boldface.

J Definition	Pearson r	p
SERM-B only	-0.44	0.129
AOT-13 only	-0.52	0.067
SKS only	$+0.12$	0.688
SERM-B + SKS	-0.26	0.387
SERM-B(rev) + AOT-13(rev)	$+0.59$	**0.035**
SERM-B(rev) + AOT-13(rev) + SKS	$+0.52$	0.069

Orientation and robustness of the J correlation. The direction of J follows from its definition (§3.3): SERM-B reads confidence, the inverse of calibrated self-assessment, and AOT-13 reads readiness to revise beliefs on external input, the inverse of resistance to fluent persuasion. The minus sign therefore labels the axis rather than adding a degree of freedom: the composite without reversal correlates with quality at $r = -0.586$ ($p = 0.035$), identical in magnitude to J. Both components are individually negative (Table 1), so the direction is not produced by combining them, and no search over sign combinations is involved. A permutation test (100,000 permutations) gives $p = 0.034$, so the parametric p-value is not an artifact of $n = 13$. In leave-one-out analysis the correlation stays positive in all 13 folds ($r \in [0.45, 0.69]$; $p < 0.05$ in 8 of 13 folds, maximum $p = 0.143$), so no single participant drives the result. The rank correlation points in the same direction but is marginal (Spearman $\rho = 0.51$, $p = 0.078$). By contrast, AI usage does not predict quality ($r = 0.34$, $p = 0.26$; Spearman $\rho = 0.19$, $p = 0.54$).

A.3 Measuring AI Usage (K)

AI usage can vary along two dimensions: *quantity* (how much interaction) and *quality* (how capable the model). We operationalize K as quantity—specifically, the mean number of interaction tokens per participant—because this choice isolates the variable of interest while holding confounds constant. All participants interacted with the same model (Claude Opus 4.6); thus, model capability, output style, hallucination rate, and instruction-following behavior are held fixed, and the only dimension of K that varies is interaction volume. Had we instead manipulated model capability (e.g., comparing a frontier model against a weaker one), differences in idea quality could not be attributed to K alone, since interaction dynamics, output format, and error profiles would co-vary with model identity.

This design choice reflects the paper's central claim: we do not argue that AI is unhelpful—stronger models likely do produce better raw material—but rather that, *given the same person and the same AI*, more interaction cannot compensate for insufficient judgment. Token count is the appropriate proxy for testing this specific prediction.

Token counts were extracted from conversation logs. Values ranged from approximately 240 to 1,340,000 tokens across participants (nearly four orders of magnitude); a $\log_{10}$ transformation was applied to address the right-skewed distribution.

A.4 Evaluating Idea Quality

Quality was assessed at two levels of granularity.

Idea-level scoring. Each idea was independently evaluated by three AI reviewer agents adopting complementary personas—Methodologist (technical rigor), Visionary (novelty and impact), and Pragmatist (feasibility and executability)—on five dimensions (1–5 scale):

- **Novelty**: Does the core insight go beyond obvious combinations? 5 = fundamentally new concept; 4 = non-obvious insight beyond "apply X to Y"; 3 = competent but predictable; 2 = repackaging; 1 = already exists.
- **Feasibility**: Can this be executed in 6 months by 2–3 researchers with standard compute? 5 = precise path with de-risking; 4 = feasible with minor uncertainties; 3 = significant assumptions required; 2 = major risks; 1 = infeasible.
- **Significance**: If successful, how much would it matter? 5 = opens new direction with clear "why now"; 4 = meaningful advance with specific impact; 3 = limited scope, generic motivation; 2 = marginal; 1 = no impact.
- **Clarity**: Is the proposal precise and coherent? 5 = crystal clear, reproducible plan; 4 = clear with minor gaps; 3 = vague in important places; 2 = significant confusion; 1 = incoherent.
- **Grounding**: Does the author show genuine understanding of the problem space? 5 = deep understanding, specific failure modes, honest limitations; 4 = good awareness; 3 = generic, formulaic; 2 = superficial; 1 = ungrounded.

The complete system prompt provided to each reviewer agent is reproduced in Figure 7. The placeholder `{idea_text}` is replaced with the full text of each research idea.

Reviewer Agent System Prompt

```
You are a senior AI/ML researcher serving as a reviewer for a
top-tier venue (NeurIPS/ICML/ICLR). Your task is to evaluate the
quality of a research idea on five dimensions.

Pay close attention to signs of deep understanding vs.
surface-level AI-generated proposals.  A high-quality idea shows
evidence of the author understanding trade-offs, anticipating
failure modes, and making non-obvious choices -- not just
describing a plausible-sounding system.
```

Research Idea to Evaluate

```
{idea_text}
```

Evaluation Dimensions

```
Rate each dimension on a scale of 1-5 and provide a brief
justification (2-3 sentences).
```

1. Novelty (1–5): Does this idea offer a genuinely new insight? 5 = fundamentally new concept; 4 = non-obvious insight beyond "apply X to Y"; 3 = competent but predictable; 2 = repackaging; 1 = already exists.

2. Feasibility (1–5): Can this be executed in 6 months by 2-3 researchers? 5 = precise execution path with de-risking; 4 = feasible, minor uncertainties; 3 = significant assumptions required; 2 = major risks; 1 = infeasible.

3. Significance (1–5): If successful, how much would it matter? 5 = opens new direction, clear "why now"; 4 = meaningful advance; 3 = limited scope; 2 = marginal; 1 = no impact.

4. Clarity (1–5): Is the proposal precise and coherent? 5 = crystal clear, reproducible; 4 = clear with minor gaps; 3 = vague in places; 2 = confused; 1 = incoherent.

5. Grounding (1–5): Does the author show genuine understanding? 5 = deep understanding, specific failure modes, honest limitations; 4 = good awareness; 3 = generic; 2 = superficial; 1 = ungrounded.

Output Format

```
Respond in JSON format only:
{"novelty_score": int, "novelty_justification": str,
"feasibility_score": int, "feasibility_justification": str,
"significance_score": int, "significance_justification":
str, "clarity_score": int, "clarity_justification": str,
"grounding_score": int, "grounding_justification": str,
"overall_comment": str}
```

Figure 7: System prompt used by each AI reviewer agent for idea-level evaluation. Three agents (Methodologist, Visionary, Pragmatist) each receive an additional persona-specific preamble before this shared rubric.

Participant-level ranking. Absolute scoring tends to cluster around the scale midpoint (central tendency bias), yielding insufficient between-participant discrimination for the J–quality correlation analysis. To address this, we supplement idea-level scoring with a comparative ranking protocol.

Three independent ranking agents each performed the following procedure: (1) read all ideas from all 13 participants in their entirety; (2) for each participant, identify their single strongest idea; (3) produce a forced ranking of all 13 participants from best to worst, assigning scores on a 1–5 scale with meaningful gaps between tiers. Each agent operated independently without access to the other agents' rankings or to any survey data. The final participant-level quality measure is the mean of the three agents' scores. This comparative protocol mirrors how conference program committees evaluate submissions: reviewers read a pool of papers and rank them relative to each other, rather than

Ranking Agent Prompt

```
You are an experienced ML researcher reviewing research idea
portfolios.  Rank 13 participants by the quality of their BEST
research idea.

Read each participant's ideation file.  Each file contains ideas
with titles and full descriptions.
```

For each participant:

1. Read ALL their ideas
2. Pick the BEST one
3. Note which idea you picked and why

```
Then RANK all 13 participants' best ideas against each other.
Consider:
• Is the core insight genuinely novel or just applying X to Y?
• Is the execution plan concrete and realistic?
• Does the author show deep understanding of trade-offs and
limitations?
• Is the problem important and well-motivated?

Assign scores on a 1-5 scale based on the ranking, but with real
gaps between tiers (not all clustered at 3-4).  Use the full
range.
```

Output Format

```
JSON: {"rankings":  [{"rank":  1, "author":  str,
"best_idea_title":  str, "score":  float, "rationale":  str}, ...,
{"rank":  13, ...}]}
```

Figure 8: Prompt used by each of the three independent ranking agents for participant-level quality assessment. Each agent reads all 13 participants' idea portfolios before producing a forced ranking.

assigning scores in isolation. Forced ranking ensures that the agents commit to quality distinctions even when absolute differences are subtle. Inter-ranker agreement was substantial: pairwise Spearman correlations between the three agents ranged from $\rho = 0.68$ to $\rho = 0.82$, indicating that the quality ordering is robust across independent evaluators. The ranking prompt is reproduced in Figure 8.

A.5 Per-Participant Data

Table 2 reports the raw survey scores, computed J composite, mean AI token usage (K), and ranking-based quality for each participant. SERM-B scores range from 5.83 to 7.83 (mean 6.88); AOT-13 scores range from 3.31 to 5.77 (mean 4.69). Token usage spans nearly four orders of magnitude (240 to 1.3M), reflecting substantial natural variation in how participants chose to interact with AI. Quality scores from the three independent rankers range from 2.20 to 4.67. The highest-quality participant (P4, Quality = 4.67) has the highest J ($+1.87$) but moderate token usage (7,000), while the participant with the most tokens (P9, $K = 1.3$M) achieves only mid-range quality (3.20)—illustrating the K-J decoupling pattern at the individual level.

A.6 Illustrative Examples

To calibrate the evaluation scale, we present the best idea from three participants selected systematically: the highest-J (P4), median-J (P1), and lowest-quality (P3) participants. Full idea texts are reproduced below; evaluation summaries follow each idea.

Table 2: Per-participant data for the K-J decoupling study. J = composite judgment score; K = mean tokens; Quality = averaged ranking score from 3 independent rankers.

ID	SERM-B	AOT-13	J	K (tokens)	Quality
P1	6.33	5.77	-0.32	47,020	3.53
P2	6.67	4.23	$+0.49$	256,800	4.37
P3	7.17	4.46	-0.11	2,119	2.20
P4	5.83	3.31	$+1.87$	7,000	4.67
P5	6.08	5.23	$+0.28$	59,400	3.43
P6	6.58	4.23	$+0.56$	240	2.83
P7	6.83	3.77	$+0.67$	44,720	4.63
P8	7.75	4.77	-0.83	621,000	3.30
P9	6.67	5.08	-0.11	1,342,737	3.20
P10	6.92	5.00	-0.27	10,980	4.07
P11	6.92	4.23	$+0.27$	7,880	4.17
P12	7.50	5.31	-1.00	60,220	3.47
P13	7.83	5.62	-1.51	311	2.67

P4 | High J ($+1.87$) | Quality = 4.67 | Behavioral Divergence Monitoring

Scores: N=3.7 F=3.0 S=4.3 C=5.0 G=5.0

Motivation. Reward hacking is a persistent failure mode in reinforcement learning from human feedback: a policy discovers behaviors that score highly under the proxy reward while violating the underlying human intent. Current mitigation practice is largely post-hoc—human evaluators inspect outputs after training, identify problematic patterns, and redesign the reward function, repeating the cycle at significant cost. This work asks: can reward hacking be detected *online*, during training, before it becomes severe enough to corrupt the final policy?
Key Insight. When a policy begins exploiting the reward proxy, its behavior on a fixed reference dataset diverges from its behavior on proxy-optimized rollouts. This behavioral divergence is detectable before the proxy reward itself shows anomalies.
Proposed System. A three-module architecture: (1) a *divergence tracker* that continuously monitors behavioral statistics on held-out reference data vs. training rollouts; (2) a *severity classifier* that distinguishes benign policy improvement from reward-exploiting divergence; (3) a *graded intervention* mechanism that can slow learning rates, increase KL penalties, or trigger human review depending on severity.
Evaluation. Experiments on three RLHF benchmarks with known reward hacking failure modes. Metrics: detection latency (how many training steps before detection), false positive rate, and final policy quality under intervention vs. baseline.
Limitations. Detection depends on the reference dataset being representative; distribution shift in the reference set could produce false positives. The severity classifier requires labeled examples of reward hacking, which are expensive to obtain.

Evaluation summary: High scores on clarity (5.0) and grounding (5.0) reflect the concrete three-module architecture, specific failure mode analysis, and honest limitations. Significance (4.3) is strong—reward hacking detection is a critical open problem. Novelty (3.7) is moderate: the behavioral divergence signal is a genuine insight, but the monitoring architecture uses established techniques.

P1 | Median J (-0.32) | Quality = 3.53 | HARP

Scores: N=3.3 F=2.0 S=4.0 C=4.7 G=3.3

Problem. Autonomous AI research agents suffer from unchecked errors, low novelty, and a binary control paradigm (fully autonomous or fixed checkpoints). Researchers want AI assistance but not full delegation. The core question: at which points in the research pipeline should an AI agent request human input to maximize quality while minimizing effort?
Approach (HARP). Formulates the problem as a contextual bandit. State space includes pipeline stage, semantic uncertainty (via multiple generations $\rightarrow$ clustering $\rightarrow$ entropy), decision criticality (learned classifier predicting downstream impact), remaining human effort budget, and interaction history. Action space: $\{\texttt{ask}, \texttt{proceed}\}$. Reward: $R = Q_{\text{final}} - \lambda \cdot T_{\text{human}}$, tracing the quality–effort Pareto frontier.
Key Components. (A) Semantic uncertainty estimation via generation diversity; (B) decision criticality classifier trained on historical revision impact; (C) efficient query interface presenting top-K options with confidence scores.
Novelty Claims. First learned intervention policy for scientific research agents; joint uncertainty–criticality signal; research-specific state representation encoding pipeline stage semantics.
Experimental Plan. ScienceAgentBench (32 ML problems); baselines include fully autonomous, fixed checkpoints, random, uncertainty-only, and criticality-only; human study with 20–30 ML researchers; ablations on each signal component.
Risks. RL policy may be unstable (fallback: learned threshold policy). Criticality classifier requires training data from historical revision logs. Human study recruitment timeline is tight.

Evaluation summary: Strong clarity (4.7) and significance (4.0)—the problem is important and the proposal is well-structured with clear formulation. Feasibility (2.0) is low: the contextual bandit requires training data that may not exist in sufficient quantity, and the human study scope is ambitious. Novelty (3.3) is moderate: the formulation is clean but the individual components (uncertainty estimation, criticality classification, contextual bandits) are established. Grounding (3.3) engages with related work but does not deeply analyze why prior approaches fail.

P3 | Low Quality | Quality = 2.20 | Geodesic Flow Alignment

Scores: N=3.3 F=2.0 S=3.3 C=3.0 G=2.0

Motivation. Current single-step models (Consistency Models, Rectified Flow) attempt to "flatten" the trajectory from noise to data. However, the image manifold is highly non-Euclidean. Standard distillation forces the model to learn a "straight line" in a space where the data distribution is curved, leading to texture blurring, mode collapse, and accumulated approximation error.
Proposed Approach. Geodesic Flow Alignment (GFA): instead of minimizing Euclidean distance between predicted and target trajectories, minimize the *geodesic distance* on the data manifold. Train a curvature estimator network alongside the flow model; use Riemannian gradient corrections during single-step inference.
Expected Results. Higher FID and improved texture fidelity compared to Consistency Models and Rectified Flow at single-step inference.

Evaluation summary: The mathematical framing (Riemannian geometry on the data manifold) is conceptually appealing, earning moderate novelty (3.3). However, the proposal lacks concrete methodology for computing geodesics in high-dimensional pixel space, does not specify the curvature estimator architecture, provides no experimental plan beyond "higher FID," and does not engage with related work beyond surface-level descriptions. The low grounding score (2.0) reflects the absence of failure mode analysis and the gap between mathematical motivation and practical implementation. At 608 words, it is substantially shorter and less developed than the other two examples.

A.7 Limitations

Sample size and statistical power. The study involves 13 participants and 68 ideas. With $n = 13$ at the participant level, statistical power is limited: the observed correlation ($r = 0.59$, $p = 0.035$) is significant but would not survive a stringent Bonferroni correction if treated as one of multiple comparisons. All conclusions should be interpreted as exploratory evidence for the K-J decoupling hypothesis rather than confirmatory tests. Replication with a larger and more diverse sample—ideally spanning multiple disciplines and career stages—is necessary before the empirical patterns reported here can be considered robust.

Measurement of J. Human judgment capacity is operationalized through two self-report scales (SERM-B and AOT-13), which are inherently subject to self-assessment bias. The reversed scoring direction—lower self-assessed ability predicts higher idea quality—is theoretically motivated by the calibrated self-assessment literature, but it could also reflect individual differences in response style (e.g., cultural norms around modesty) rather than genuine differences in evaluative capacity. A stronger design would supplement self-report instruments with behavioral measures, such as having participants evaluate a set of pre-scored research ideas without AI assistance, providing a direct assessment of judgment quality independent of self-perception. We state this behavioral measure as the next study (§5.4).

Measurement of idea quality. Using AI reviewer agents to evaluate AI-assisted output introduces a potential circularity: the evaluator shares architectural similarities with the tool used to produce the output, which could systematically bias quality assessments. The two-level evaluation protocol (idea-level absolute scoring plus participant-level comparative ranking by independent agents) partially mitigates this concern by introducing evaluator diversity and forcing comparative judgments. Nevertheless, human expert blind evaluation—in which domain specialists assess idea quality without knowing which participant produced each idea or how much AI was used—would provide substantially stronger validation. Future work should include such calibration to quantify the extent of AI-evaluator bias.

Measurement of K. Token count is a coarse proxy for AI usage that conflates qualitatively different interaction patterns. A participant who spends 100,000 tokens on deep back-and-forth critique with the AI is using AI differently from one who generates 100,000 tokens of uncritical brainstorming. The current design cannot distinguish these modes. Finer-grained interaction analysis—such as coding each conversational turn for whether the participant challenged, accepted, or redirected the AI's output—would provide a richer characterization of how AI usage relates to idea quality.

Generalizability. All participants are computer science researchers working on AI/ML topics, and all used the same model (Claude Opus 4.6). The extent to which the K-J decoupling pattern generalizes to other disciplines (where domain knowledge structures differ), other AI tools (where interaction affordances vary), or other research stages beyond ideation (where independent verification may be more readily available) remains an open question.

B Benchmarking Autonomous Research Systems: Methodology and Results

B.1 Scope and Task

This evaluation quantifies, under a uniform reviewing protocol, the gap between the paper-level outputs of current autonomous research systems and the standard of a top-tier ML venue. It complements the K-J decoupling and scaffolding studies (Appendices A and C), which examined human-AI co-ideation; the pipelines evaluated here attempt to remove the human from the loop entirely.

We evaluate 49 papers drawn from six autonomous research systems (Table 3): Dr-Claw, AutoResearchClaw, EvoScientist, FARS, and both generations of The AI Scientist. The systems differ in architecture and target subdomain; all treat the produced manuscripts as submission-ready outputs rather than intermediate artifacts. Papers for Dr-Claw, AutoResearchClaw, EvoScientist, and AI Scientist V2 are evaluated exhaustively; for FARS (166 released) and AI Scientist V1 (269 released), 10 papers each are sampled. The AI Scientist V1 sample is stratified across the four backbone LLMs used by the system (Claude, DeepSeek, GPT-4o, Llama 3.1) and the three task domains covered (diffusion, grokking, nanogpt).

B.2 Review Protocol

All manuscripts are evaluated using the automated reviewer from The AI Scientist (Lu et al., 2024) (specifically, the `perform_review.py` pipeline), which implements a modified NeurIPS reviewer form with numerical scores for Soundness, Presentation, Contribution, Originality, Quality, Clarity,

Table 3: Scope of the benchmark evaluation. "Reviewed" counts the papers passed through the reviewer agent; "Released" is the total manuscript count made available by each system.

System	Domain focus	Released	Reviewed
Dr-Claw	AI/ML (mixed)	12	12
AutoResearchClaw	Statistics / scientific ML	8	8
EvoScientist	AI/ML (mixed)	6	6
AI Scientist V2	AI/ML (mixed)	3	3
FARS	LLM safety / alignment	166	10
AI Scientist V1	Diffusion / grokking / nanoGPT	269	10
Total			**49**

Significance (each 1–4), an Overall score (1–10), a Confidence score (1–5), and a binary Accept/Reject decision.

Model and parameters. Reviews are produced by `claude-sonnet-4-6` at temperature 0.75, with a 4,096-token response budget. Each paper receives a single review rather than the ensemble-plus-reflection protocol used in the original AI Scientist paper. The reviewer uses the `reviewer_system_prompt_neg` setting ("Be critical and cautious... If a paper is bad or you are unsure, give it bad scores and reject it"), identical to the original AI Scientist evaluation.

Artifact neutralization. Several papers contain self-identifying disclaimers (e.g., "VibeLab Test Artifact," "AI-Scientist Generated Preprint," explicit provenance statements). These are stripped during text preprocessing, and the system prompt instructs the reviewer to ignore any residual provenance markers.

PDF ingestion. Manuscripts are extracted to plain text; text exceeding 80,000 characters is truncated. Equations, tables, and figures are retained only insofar as they survive PDF-to-text conversion.

B.3 Aggregate Results

Across 49 papers and six systems, the automated reviewer accepted zero (Table 4). Mean Overall scores range from 1.00 (AI Scientist V2) to 2.60 (FARS), all well below the NeurIPS Weak Accept threshold of 6. Soundness, which captures whether the paper's core technical claims are adequately supported, averages below 2 ("Fair") for every system.

Table 4: Per-system mean scores under the shared NeurIPS-style rubric. Overall is on a 1–10 scale; Soundness, Presentation, and Contribution are on a 1–4 scale. No paper in any system received an Accept decision.

System	n	Overall	Soundness	Presentation	Contribution
FARS	10	2.60	1.70	2.60	1.70
Dr-Claw	12	2.33	1.42	2.00	1.92
EvoScientist	6	2.00	1.17	1.33	1.83
AI Scientist V1	10	1.30	1.00	1.00	1.10
AutoResearchClaw	8	1.12	1.00	1.38	1.00
AI Scientist V2	3	1.00	1.00	1.00	1.00
Accept rate	49		0 / 49		

B.4 Cross-System Failure Modes

The reviewer's free-text Weaknesses fields converge on a small set of recurring problems that recur across systems of substantially different design.

Placeholder artifacts in the final PDF. Many manuscripts contain literal placeholder text in place of central content. Loss-function definitions are replaced by "[picture intentionally omitted]" in

Dr-Claw SyncLite; equations and figures are similarly omitted across AutoResearchClaw and AI Scientist V1 papers; AI Scientist V2 manuscripts contain broken citation keys rendered as "?" for backbone models such as ResNet-18 and standard datasets such as CIFAR-10.

Claimed experiments that were not run. Dr-Claw PhysCode4D states that Tables 1–5 report "predicted estimates" rather than measured values. AutoResearchClaw papers report a "logged scalar benchmark metric" that the paper itself acknowledges is "dimensionless in the artifact" and never operationally defined. EvoScientist Adaptive Evidential Meta-Learning describes a main comparison table in prose but omits the table itself.

Claims that contradict the reported numbers. AI Scientist V1 ScaleDiffuse claims "consistent improvements" while one of four datasets shows a degradation (KL $0.095 \rightarrow 0.109$). FARS canary-controlled experiments compare a 13% interleaving method against a 5% fixed baseline and report the gap as a method effect rather than a dosage effect. AI Scientist V2 Label Noise reports accuracy as "$\sim$85%" in text while the accompanying figure shows $\sim$75%.

Trivial or ungrounded novelty claims. Multiple AI Scientist V1 papers apply established techniques (layer aggregation, curriculum learning, multi-scale processing) to toy-scale settings and describe the result as a "significant advancement." Dr-Claw MonoVL proposes a unified token-space VLM design that the paper's own introduction acknowledges "is not new," then claims architectural novelty. The reviewer consistently rates Originality and Significance at 1–2 for these papers.

Missing ablations, statistical tests, or baseline fairness. FARS Deflated-RankICIR reports headline gains whose 90% confidence intervals include zero. Dr-Claw ditMem (the single highest-scoring paper in the benchmark, Overall $= 4$) is faulted for proposing a benchmark (MultiEditBench) that it also uses to evaluate itself without independent validation.

B.5 Illustrative Reviews

To calibrate the aggregate scores in concrete terms, we reproduce condensed reviews for the benchmark's highest- and lowest-scoring papers: Dr-Claw's *ditMem* (Overall $= 4$) and AI Scientist V1's *ScaleDiffuse* (Overall $= 1$).

Dr-Claw | ditMem | Overall = 4 (highest in benchmark)

Scores: Overall=4 Soundness=2 Presentation=3 Contribution=3 Decision = Reject

Reviewer summary. A multi-turn interactive video editing framework built on Diffusion Transformers augmented with a Persistent Spatial-Temporal Memory module, a Memory-Gated Consistency loss, and a new benchmark (MultiEditBench) for multi-turn editing consistency.
Strengths cited. Addresses an under-explored problem; the PSTM architecture is clearly motivated; the paper is transparent about failure cases including configurations where the proposed method underperforms baselines; comprehensive ablations isolate each component; a user study with 30 participants corroborates quantitative findings.
Weaknesses cited. Training data are pseudo multi-turn sequences from per-frame InstructPix2Pix plus post-hoc smoothing, leaving a domain gap; the MultiEditBench benchmark is proposed and evaluated by the same authors without independent validation; baselines are applied sub-optimally by feeding outputs iteratively through single-turn editors; FID-VID (76.8) is worse than RAVE (72.1) and AnyV2V (74.3), indicating the memory conditioning degrades per-frame quality; the memory-update threshold τ is not specified, analyzed, or ablated; CEA improvement is marginal (0.289 vs. 0.281) and the paper acknowledges it may not survive statistical testing.

AI Scientist V1 | ScaleDiffuse | Overall = 1 (lowest)

Scores: Overall=1 Soundness=1 Presentation=1 Contribution=1 Decision = Reject

Reviewer summary. An "Adaptive Multi-Scale Diffusion" method that processes inputs at five fixed scales with learnable weights, evaluated on four 2D toy datasets against a single baseline using KL divergence.
Strengths cited. The multi-scale idea is reasonable; an ablation examines component contributions; computational overhead is honestly reported at $\approx 2\times$; scale-weight evolution offers some interpretability.
Weaknesses cited. Sections 3.1–3.4, which constitute the entire method description, are missing or replaced with placeholder images; multiple equations and figures are marked "picture intentionally omitted"; the method degrades performance on the *moons* dataset (KL $0.095 \rightarrow 0.109$) while the abstract claims "consistent improvements"; evaluation is limited to four 2D toy datasets despite claims of broader applicability; best-case improvement is 3.15% KL reduction at $2\times$ compute; the introduction contains duplicated paragraphs; statistical significance is claimed ($p < 0.01$) but the test is never specified; the core technique is not novel.

B.6 Limitations

Single-review variance. Each paper receives one review from one model at temperature 0.75. The original AI Scientist protocol uses ensembles (multiple reviewers, mean aggregation) and reflection (multi-round self-revision) to reduce variance; both are omitted here to keep the evaluation tractable. Per-paper scores should be interpreted with a confidence band of roughly ± 1 on the Overall scale, and aggregate patterns across systems are more robust than any individual number.

Negative reviewer prompt. The system prompt instructs the reviewer to default toward rejection under uncertainty ("If a paper is bad or you are unsure, give it bad scores and reject it"), matching the setting used in the AI Scientist paper's own self-evaluation. This systematically depresses scores relative to a neutrally-prompted reviewer, and the 0/49 accept rate should be read against this backdrop rather than as evidence that autonomous output fails a maximally permissive bar.

PDF-to-text fidelity. Equations, tables, and figures that do not survive PDF extraction are invisible to the reviewer. Some Weaknesses citations of the form "Figure X is missing" could therefore be extraction artifacts rather than submission defects. Inspection of the source PDFs indicates that the bulk of such citations, particularly those involving the "[picture intentionally omitted]" marker, correspond to genuine omissions, but false positives at the case-by-case level cannot be ruled out.

Truncation at 80,000 characters. Appendix content in the longest submissions may be truncated. This primarily affects supplementary material rather than main-paper claims, but where a headline result is defended only in an appendix, the reviewer may not see that defense.

Sampling for FARS and AI Scientist V1. Exhaustive review was infeasible for the two systems that released > 100 papers. The AI Scientist V1 sample is stratified across the four backbone LLMs and three tasks covered by the system; the FARS sample is the first 10 papers, not stratified. A stratified FARS sample could shift its mean in either direction.

Evaluator circularity. As with the K-J study (Appendix A.7), using an AI reviewer to evaluate AI-generated output carries a circularity risk. A blinded human expert evaluation would provide stronger calibration of the failure modes reported in §B.4.

C Scaffolding: Detailed Methodology and Experiments

C.1 Design, Participants, and Task

This experiment tests the ACF's prescription: structured scaffolding that shifts AI's role from answer provider to judgment builder should improve research idea quality, particularly when the researcher lacks domain expertise.

The same 13 participants from the K-J decoupling study (Appendix A) each completed four additional ideation tasks in a 2×2 within-subjects design crossing domain expertise (expert area vs. non-expert

area) with guidance condition (free AI interaction vs. four-stage scaffolding). In the *expert-area* conditions, participants ideated within their own research specialization; in the *non-expert-area* conditions, they ideated in a topic chosen from another participant's domain. In the *guidance* conditions, participants followed the four-stage scaffolding protocol described below (§C.2); in the *control* conditions, they interacted freely with the same AI model (Claude Opus 4.6) without any prescribed structure. Task order was counterbalanced across participants to mitigate learning effects. Each task produced one final research idea, yielding 52 ideas total (13×4).

Each idea was independently evaluated by three AI reviewer agents (Methodologist, Visionary, Pragmatist) on five dimensions—novelty, feasibility, significance, clarity, and grounding—using the same 1–5 rubric and prompt described in Appendix A.4.

C.2 Four-Stage Guidance Protocol

The guidance condition operationalizes decision readiness scaffolding (§4.1) as a structured four-stage workflow in which AI's role progressively shifts from knowledge provider to judgment scaffolder. Each stage's output serves as the next stage's input.

Stage 1: Diverge. AI serves as generator: opening the problem space, providing candidate directions, possible framings, and approaches from related fields. The human sets boundaries (what to investigate, what to exclude) but does not choose. This corresponds to the outer loop of the ACF. The core principle is to avoid premature convergence, as the option that appears most promising at first glance may simply reflect AI's tendency toward mainstream framing.

Stage 2: Structure. AI shifts from generator to organizer: transforming vague candidates into structured, comparable objects, each expanded along uniform dimensions such as core question, key assumptions, potential value, main risks, and differentiation from alternatives. The human defines the comparison dimensions. This stage marks the transition from outer loop to inner loop, as the human begins to recognize which aspects they cannot evaluate and where scaffolding is needed.

Stage 3: Scaffold and Stress-Test. This is the inner loop's core. AI's role inverts to challenger: explaining background knowledge, filling information gaps, exposing trade-offs, surfacing hidden assumptions, identifying failure modes, and stress-testing each candidate from the opposing side. The human builds deep understanding of each candidate through this scaffolding—not by being told which option is best, but by examining for themselves where each option's strengths and weaknesses lie. AI usage can remain high in this stage, but what increases should be scaffolding and critique, not AI generating conclusions on the human's behalf.

Stage 4: Human Decides. The human, having been scaffolded, makes the choice and bears decisional responsibility. AI provides only targeted refinement. If understanding remains insufficient, the process returns to Stage 3, which constitutes the inner loop's iteration mechanism. The critical discipline is that the human must actively judge rather than accept AI's implicit ranking.

In the control condition (free AI interaction), participants received no stage structure and interacted with the same AI model in an unconstrained conversational format.

C.3 Illustrative Comparison

To illustrate how scaffolding changes idea quality in practice, we present the expert-area and non-expert-area ideas of participant P7 (median guidance lift among all participants, excluding co-authors). For each domain, we first describe the key difference between the two ideas, then reproduce the full texts with scores.

Expert area (Medical VLMs). Without guidance, the participant proposed an 8-type hallucination taxonomy requiring 8,000 annotations across 5 models, a hybrid detection system, and a controlled mitigation study—essentially an entire research program compressed into one proposal, with no discussion of which component to prioritize if resources run short. With guidance, the same participant narrowed to a single phenomenon—*confabulation* (misdiagnosing unfamiliar pathology as a common condition)—and designed a concrete escalation protocol that degrades gracefully if detection proves

difficult. The scores reflect this shift: grounding rose from 3.0 to 4.0, clarity from 3.0 to 4.7, and novelty from 3.0 to 4.3.

P7 | Expert Area | Without Guidance | Overall = 3.00

Scores: N=3.0 F=2.0 S=4.0 C=3.0 G=3.0

Medical VLMs hallucinate in fundamentally different ways—fabricating findings absent from the image, misattributing observations to wrong locations, inventing specific measurements from modalities that cannot support them, and generating unsupported temporal claims about change from nonexistent prior studies—yet existing benchmarks collapse these distinct failure modes into a single "hallucination" label. We propose MedHallBench 2.0, an 8-type hallucination taxonomy validated through iterative pilot annotation with a deterministic decision tree to ensure inter-annotator reliability (target $\kappa > 0.7$ for ≥ 6 types). The benchmark pairs 8,000 sentence-level annotations across 5 open-source medical VLMs with instance-level clinical severity scoring, a hybrid detection system providing automated evaluation for 5 reliably detectable types and honest acknowledgment that 3 subtle types require human judgment, and a controlled mitigation diagnosis applying DPO, contrastive decoding, and grounding-based generation to the same base model.

P7 | Expert Area | With Guidance | Overall = 4.07

Scores: N=4.3 F=3.0 S=4.3 C=4.7 G=4.0

When medical vision-language models encounter rare conditions absent from their training data, they do not abstain—they *confabulate*, confidently misassigning unfamiliar pathology to the nearest common diagnosis. This failure mode is distinct from hallucination (fabricating findings not present in the image) and is not addressed by existing mitigations like DPO or contrastive decoding, which target generation quality rather than knowledge gaps. We propose to first characterize this phenomenon empirically across multiple medical VLMs, establishing confabulation rates on rare conditions and demonstrating that standard hallucination mitigations fail to reduce them. Building on this characterization, we develop a concept-level out-of-knowledge detector that identifies when an input falls outside the model's competence boundary, triggering a structured escalation protocol—visual observations without diagnostic labels, closest known differentials, and a recommendation for specialist review. The work is designed to degrade gracefully: the characterization stands as an independent contribution even if detection proves difficult.

Non-expert area (Music Generation). Without guidance, the participant—a medical VLM researcher with no music generation background—proposed measuring attribute coupling in audio codecs and building a multi-stream codec with three disentanglement losses, but could not specify how to obtain ground-truth attribute labels, what success would look like, or which baselines to compare against. With guidance, the participant shifted to a different problem entirely: generating synthetic video-music training pairs to replace copyrighted data, motivated by the Udio/Suno copyright settlements of 2025–2026. The scaffolded proposal names specific tools (CogVideoX-5B, SymMV), defines a concrete experiment (can 50K synthetic pairs match 900 real pairs?), and contributes a domain gap taxonomy. Grounding rose from 3.0 to 4.7—the largest single-dimension improvement in this pair—reflecting the shift from a vague technical sketch to a proposal anchored in specific resources and real-world constraints.

P7 | Non-Expert Area | Without Guidance | Overall = 3.27

Scores: N=3.7 F=2.3 S=3.3 C=4.0 G=3.0

Musical attributes like pitch, rhythm, timbre, and dynamics are physically coupled in real instrument acoustics—the accent pattern that makes funk feel like funk *is* a loudness contour, and a violin's timbre changes with every pitch. Every neural audio codec in use today (EnCodec, DAC, SoundStream, DisCodec) learns entangled representations where these attributes are mixed across codebook levels, making fine-grained attribute control impossible at the representation level. We present the first empirical measurement of these attribute couplings in music audio across 10,000 tracks, and propose a Multi-Stream Codec with attribute-specific encoder heads trained with three disentanglement losses—attribute prediction probes, mutual information minimization, and swap consistency.

P7 | Non-Expert Area | With Guidance | Overall = 4.13

Scores: N=4.0 F=3.0 S=4.3 C=4.7 G=4.7

The era of training video-to-music models on scraped copyrighted data is ending—major label settlements with Udio and Suno in 2025–2026 have made this clear—yet no open alternative exists for generating large-scale video-music training pairs with controlled cross-modal alignment. We propose SynthV2M, a fully open pipeline that extracts rich musical attributes from licensed seed music, translates them into temporally structured video generation prompts via a learned Alignment Specification Language, and generates synchronized video using open-source models (CogVideoX-5B) at zero API cost. The central experiment tests whether 50K synthetic pairs can match the training value of ∼900 real-world pairs from SymMV, with mixed-training experiments probing whether synthetic data is most valuable as a supplement to scarce real data. The work contributes a systematic domain gap taxonomy characterizing where synthetic-trained models fail on real video, and a reproducible copyright audit methodology.

C.4 Supplementary Analyses

The analyses below re-analyze the data of this experiment; no new data were collected.

Non-parametric tests. Wilcoxon signed-rank tests agree with the paired t-tests in §4.3: $W = 1$, $p < 0.001$ in both the expert-area and non-expert-area conditions.

Variance compression and floor raising. The between-participant SD of overall quality, pooled across expertise conditions, falls from 0.40 without guidance to 0.23 with guidance (Levene $W = 7.19$, $p = 0.010$), and the lowest participant-condition score rises from 2.33 to 3.40. In the non-expert area, guided quality meets or exceeds each participant's own unguided expert-area score for all 13 participants (12 strictly above, 1 tie).

Output length. Guided ideas are shorter than unguided ones (mean 794 vs. 872 words). Across the 52 ideas, length is only weakly associated with quality ($r = 0.26$, $p = 0.067$), and the association between guidance condition and quality is unchanged after partialing out length (raw $r = 0.80$; partial $r = 0.84$, $p < 10^{-13}$). At the participant level, the change in length correlates with the change in quality ($r = 0.81$, $p < 0.001$), even though mean length falls while every lift is positive; length therefore cannot explain the main effect, but it is not fully orthogonal to quality at the individual level.

Null results. Self-reported J does not predict individual lift ($r = 0.14$ in the expert area and $r = 0.11$ in the non-expert area, both $p > 0.6$), and J does not correlate with quality within the unguided conditions of this experiment (r between -0.10 and 0.05). The predicted pattern that lower-J participants gain more therefore does not appear with the self-report measure. Both nulls may be limited by measurement: this experiment uses absolute idea-level scoring, which compresses scores toward the center (the reason the K-J study uses forced ranking; Appendix A.4), and guided scores cluster near the ceiling (SD 0.20), which limits detectable variation in lift. Regression to the mean can also compress variance between paired measurements, so we read the variance compression above as consistent with the mechanism rather than as a direct test of it.

C.5 Limitations

Sample size and domain selection. The experiment involves 13 participants producing 52 ideas (13×4 conditions). While the within-subjects design increases statistical power relative to a between-subjects comparison, the sample remains small. The counterbalancing of task order mitigates but does not eliminate potential learning effects across conditions. Non-expert topics were drawn from other participants' specializations rather than from a standardized pool, so the degree of domain distance varied across participants and may contribute to variance in the scaffolding effect.

Evaluation blinding. The AI evaluator agents were not blinded to guidance condition, and in practice blinding may be infeasible: scaffolded ideas tend to exhibit structural signatures—explicit trade-off analysis, staged reasoning, named failure modes—that are largely absent from unscaffolded ideas. If evaluators implicitly associate these features with higher quality, the measured scaffolding effect could be inflated. A stronger design would have human evaluators score ideas after stripping

structural formatting, or would use a matched-format protocol in which unscaffolded ideas are also post-edited into a comparable structure before evaluation.

Effort confound. Participants in the guidance condition followed a four-stage protocol that may have increased the total time and cognitive effort invested in each idea relative to the free-interaction control. The observed quality improvement could therefore reflect greater effort rather than the scaffolding mechanism itself. A controlled comparison would equalize time-on-task across conditions, or measure and report time as a covariate. Time-on-task was not logged in this experiment. Output length does not account for the effect (§C.4), but a structure-matched, time-matched control remains necessary, and it doubles as a falsification test (§5.2).

Generalizability. The experiment tests scaffolding only at the ideation stage of AI/ML research. Whether the four-stage protocol remains effective at other research stages—experiment design, data analysis, manuscript writing—where the nature of judgment differs, and in other disciplines, remains an open question.

Mechanism. The experiment demonstrates that scaffolding improves quality but does not isolate which of the four stages contributes most. The observed improvement could stem primarily from the stress-testing in Stage 3, from the scope narrowing in Stage 2, or from the mere imposition of any structured process. Ablation studies removing individual stages would clarify the active ingredients. A structure-matched control arm, with the same stages and time budget but the judgment-building content replaced by generic interaction, would test the last possibility directly (§5.4).

D Research Agenda: Open Questions by Design Dimension

Table 5 expands the design space of §5.4. Each row is a contextual dimension derived from the ACF; the middle column states what the dimension sets in the framework, and the right column lists open questions that research on scaffolding should answer. The final row collects questions about measurement that cut across all dimensions.

Table 5: Open questions for decision readiness scaffolding, organized by the ACF's contextual dimensions.

Dimension	What it sets in the ACF	Open questions
Expertise gap	Where the inner loop starts; the size of the $K \rightarrow J$ conversion	How should scaffolding content adapt to the gap (foundations first vs. higher-order judgment)? Can the gap be estimated before a decision point, for instance from the researcher's own questions? When does the gap exceed what point-wise scaffolding can close?
Verifiability	How much evaluation can exit to external checks	How can mixed stages be decomposed into verifiable and judgment-bound parts? Which proxy checks (tests, citation tracing) are mistaken for verification of the claim itself?
Stakes	How much inner-loop iteration precedes a decision	How much scaffolding is enough before committing? How should the stopping rule of the inner loop scale with the cost of an error?
Domain	The content of the loops, not their structure	Does the loop structure transfer beyond computer-science ideation to experimental, clinical, or social-science research? Which domain artifacts (protocols, datasets, instruments) should scaffolding surface?
Workflow instantiation	Which elements of the four-stage workflow carry the effect	Which stages are load-bearing when ablated? Does a structure-matched control with generic content reproduce the gains?
Measurement (cross-cutting)	How judgment and its change are observed	Can judgment be measured behaviorally, for example by blind-judging pre-scored ideas? How closely do AI evaluators agree with format-stripped human expert blind evaluation? Does repeated scaffolding build durable competence or dependence over time?

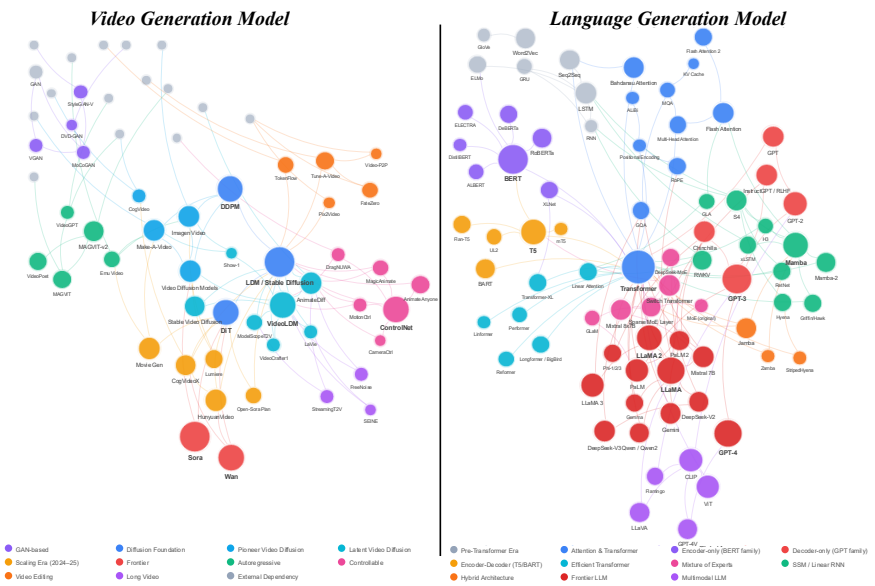


Figure 9: Citation network graphs constructed by the automated literature survey module. **Left**: Video Generation (60 nodes, 57 edges, 11 categories, 2014–2025). **Right**: Language Generation (82 nodes, 101 edges, 11 categories, 1986–2025). Node size indicates citation count; colors denote thematic categories; edges represent citation relationships.

E Illustrative Case Studies: Detailed Methodology and Artifacts

This appendix provides methodological details and extended results for three case studies of the vibe research pipeline: automated survey, iterative idea generation, and semantic analysis of human versus machine ideas. All artifacts were produced by a single vibe research system operating on Claude Opus 4.6.

E.1 Automated Literature Survey

Construction protocol. For each of the two survey topics (Video Generation and Language Generation), the system receives only a topic name and a brief scope description. It then autonomously (1) retrieves relevant publications from open-access repositories, (2) extracts pairwise citation relationships, (3) assigns each paper to a thematic category via keyword matching combined with embedding-based clustering, and (4) organizes the resulting corpus into a directed citation graph where nodes represent papers and edges represent citation dependencies. Each node is annotated with publication year, venue, and a categorical label. No manual curation is applied.

Video Generation citation network. The system produces a citation graph containing 60 nodes and 57 directed edges spanning the period 2014–2025 (Figure 9, left). The nodes are organized into 11 thematic categories: GAN-based (4), diffusion foundation (3), pioneer video diffusion (4), latent video diffusion (7), scaling era 2024–25 (6), frontier (2), autoregressive (5), controllable (6), video editing (5), long video (3), and external dependency (15). The graph captures key architectural lineages: from DDPM to LDM/Stable Diffusion to DiT and subsequently to frontier systems such as Open-Sora Plan. The highest-importance nodes (as measured by citation in-degree within the graph) are LDM/Stable Diffusion, Sora, DDPM, DiT, and Wan.

Language Generation citation network. The Language Generation graph contains 82 nodes and 101 directed edges spanning the period 1986–2025 (Figure 9, right), also organized into 11 categories: pre-Transformer (7), attention & Transformer (11), encoder-only/BERT family (7), decoder-only/GPT family (6), encoder-decoder (5), efficient Transformers (6), Mixture of Experts (6), SSM/linear RNN



Figure 10: Idea iteration trajectories for the two case studies. **Top**: RADS evolves over 4 iterations from a rough hypothesis about data ordering to a complete algorithm with 7B-scale validation. **Bottom**: ditMem evolves over 5 iterations from naive memory concatenation to a full multi-turn video editing framework with persistent spatial-temporal memory. Maturity percentages indicate progression from initial concept to publication-ready contribution.

(10), hybrid architecture (3), frontier LLM (16), and multimodal LLM (5). The highest-importance nodes are Transformer (in-degree importance 7), BERT (6), GPT-3 (6), LLaMA (5.5), and Mamba (5). The graph traces branching trajectories from the Transformer through encoder-only and decoder-only lineages, through efficiency-focused branches (Efficient Transformers, SSMs), and into the current frontier LLM generation.

Strengths and limitations of the automated survey. The system achieves broad coverage within minutes, identifying papers across subfields that a manual search might require weeks to assemble. The citation graph reveals architectural lineages and thematic clusters that are difficult to discern from reading papers individually. The automatic categorization produces groupings that align well with standard expert-level organization of both fields. However, the system exhibits recency bias, over-representing papers from the last two years and under-representing foundational work before 2018. Coverage is uneven across venues: workshop papers and technical reports are underrepresented, while highly-cited survey papers are overweighted. The categorical labels occasionally conflate distinct research threads (e.g., grouping all efficient architectures regardless of whether they address computational cost, memory, or latency). The graph captures citation *structure* but not citation *semantics*: a citation that builds upon prior work is indistinguishable from one that criticizes or merely acknowledges it.

E.2 Iterative Idea Generation

Process description. The system ingests the citation graph from the survey phase together with a stream of recent publications to identify active research fronts. It generates an initial idea by detecting gaps or under-explored intersections in the graph, then refines the idea through multiple iterations. At each iteration, the human researcher may direct the system to read additional papers in targeted subfields, redirect the hypothesis, or impose domain-specific constraints. The primary human intervention is guiding which literature the system should engage with at each stage.

Case Study A: RADS (Reasoning-Aware Data Scheduling). RADS evolves over four iterations from a general survey to a complete algorithm with 7B-scale validation (Figure 10, top). Table 6

summarizes each step; maturity percentages are estimated by the system based on the fraction of a publication-ready contribution that each iteration represents.

Table 6: Iteration trajectory for RADS. Each row lists the new components introduced at that step.

Step	Stage	Maturity	New components
0	General survey	12%	Literature landscape scan of data-centric pretraining (Penedo et al., 2024; Xie et al., 2023), curriculum learning (Bengio et al., 2009), LLM reasoning (Wei et al., 2022); identified gap: the "when" axis of data scheduling is unexplored ("what data" and "how much data" have been studied, but not temporal ordering)
1	Iteration 1	32%	Hypothesis that front-loading vs. back-loading reasoning data yields different robustness profiles; two schedules (early vs. late); small-scale pilot on GSM8K
2	Iteration 2	55%	Formal definitions of reasoning density and reasoning momentum (EWMA of reasoning exposure); probing classifiers at intermediate checkpoints; expansion to four schedules
3	Iteration 3	78%	RADS algorithm: exponentially increasing reasoning density with periodic consolidation phases; 1.3B-parameter model trained; long-horizon evaluation
4	Iteration 4	100%	Scale to 7B parameters; ablation study; post-hoc validation on LLaMA-3.1-8B; noise sensitivity analysis; final result: $+3.8\%$ average reasoning accuracy, 52% reduction in long-horizon degradation

The survey step identifies a structural gap in the literature: prior work addresses *what* data to use and *how much*, but not *when* specific data categories should be introduced during pretraining. Iterations 1–2 formalize this gap into testable hypotheses and introduce the reasoning momentum concept (an exponentially weighted moving average borrowed from time-series analysis). Iterations 3–4 move from observation to prescription, producing a concrete scheduling algorithm and scaling the experiments.

Case Study B: ditMem (Multi-Turn Video Editing with Persistent Memory). ditMem evolves over five iterations from naive memory concatenation to a full multi-turn video editing framework (Figure 10, bottom). Table 7 summarizes each step. Iteration 1 reveals a concrete failure mode (artifacts from naive concatenation) that motivates the architectural redesign in Iteration 2. The progression from Iteration 3 (loss design) to Iteration 4 (benchmark creation) to Iteration 5 (ablation and user study) mirrors the structure of a typical venue submission.

E.3 Semantic Similarity Between Human and Machine Ideas

Corpus construction. Two corpora of research ideas are assembled in the domain of language generation. The first corpus comprises 401 research ideas extracted from published papers in NLP and machine learning venues (ACL, EMNLP, NeurIPS, ICML) from the pre-LLM era (2018–2021), prior to the widespread availability of large language models for research assistance. Each idea is represented as a one-paragraph natural-language description of the core hypothesis, proposed method, and expected contribution. The second corpus consists of 75 research ideas generated by the vibe research system using the iterative refinement pipeline described above, restricted to topics that were active during the same period to prevent the system from exploiting post-2021 knowledge.

The 401 human-generated ideas span 19 topical categories (Table 8). The 75 LLM-generated ideas are labeled as a single category for visualization purposes.

Embedding and projection. All 476 idea descriptions are encoded into vector representations using Sentence-BERT (Reimers & Gurevych, 2019) with the `all-mpnet-base-v2` model (768 dimensions). t-SNE (van der Maaten & Hinton, 2008) with perplexity 30 and 1,000 iterations is applied to project the embeddings into two dimensions for visualization (Figure 11).

Table 7: Iteration trajectory for ditMem.

Step	Stage	Maturity	New components
0	General survey	10%	Literature scan of diffusion-based video editing (Tune-A-Video (Wu et al., 2023), FateZero, TokenFlow), DiT architectures (Peebles & Xie, 2023), and memory mechanisms in video object segmentation; identified gap: all editing methods operate single-turn with no memory of prior edits
1	Iteration 1	22%	Naive key-value memory with concatenation; produces artifacts and unbounded memory growth
2	Iteration 2	40%	Switch from U-Net to DiT backbone; Persistent Spatial-Temporal Memory (PSTM) module with structured memory banks; gated residual fusion (initialized $\alpha = 0$)
3	Iteration 3	60%	Memory-Gated Consistency (MGC) loss for temporal smoothness in memory-attended regions; memory compression via weighted clustering; training pipeline
4	Iteration 4	82%	MultiEditBench benchmark (200 source videos, 840 multi-turn editing sequences); four new metrics (CTPS, TC, CEA, FID-VID); comparison against five baselines (TokenFlow, RAVE, FateZero, AnyV2V, DiT-Naive)
5	Iteration 5	100%	Ablation study (5 variants); user study with 30 participants; failure case analysis; transparent trade-off reporting; final result: CTPS 0.098 vs. baseline 0.163, 40% slower consistency degradation over editing turns

Table 8: Distribution of 401 human-generated ideas across 19 topical categories, assigned via k-means clustering on Sentence-BERT embeddings with manual label assignment to each cluster centroid.

Category	Count	Category	Count
Decoder Models	30	Multimodal LLM	20
State Space Models	30	Reasoning & Agents	20
Misc. NLP Architecture	30	Scaling & Training	20
Attention Mechanisms	25	Encoder-Decoder	18
Encoder Models	25	Positional Encoding	15
Efficient Transformers	25	Hybrid Architectures	15
Mixture of Experts	25	Retrieval & Knowledge	15
Open LLMs	25	Vision Architectures	15
Transformer Core	20	Compression	15
		Tokenization	13

Results. In the resulting projection, LLM-generated ideas (star markers) are distributed throughout the same regions as human-generated ideas (circle markers), with no discernible systematic separation by origin. LLM-generated ideas appear in clusters corresponding to Attention Mechanisms, State Space Models, Efficient Transformers, Decoder Models, and other categories, rather than forming an isolated cluster of their own.

Representative LLM-generated ideas include "topological attention using persistent homology of token relationships," "fractal transformer with self-similar nested architecture," "evolutionary state space model with parameter mutation and selection during training," and "diffusion-based token refinement iteratively denoising all output tokens simultaneously." These descriptions occupy the same embedding neighborhoods as human-generated ideas addressing similar technical concepts, suggesting that the iterative refinement pipeline produces ideas of comparable conceptual substance as measured by distributional proximity in embedding space.

Limitations of the similarity analysis. Semantic co-location in embedding space does not imply equivalence in quality, feasibility, or scientific impact. Sentence-BERT encodes surface-level semantic content; two ideas may be close in embedding space because they use similar vocabulary while differing substantially in depth of formalization or experimental grounding. The t-SNE projection is

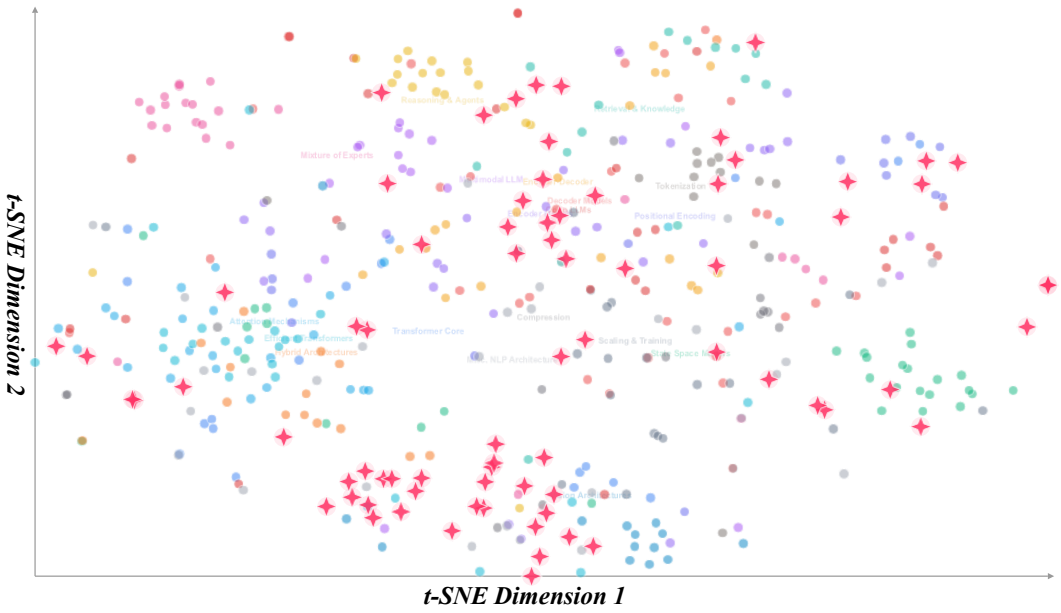


Figure 11: t-SNE visualization of 476 research ideas embedded via Sentence-BERT (Reimers & Gurevych, 2019). Circle markers: 401 human-generated ideas from the pre-LLM era (2018–2021). Star markers: 75 LLM-generated ideas produced by the vibe research system. Colors indicate topical categories (19 categories for human ideas). The spatial overlap between the two populations indicates that LLM-generated ideas occupy the same semantic regions as human ideas, with no systematic separation by origin.

a nonlinear dimensionality reduction that preserves local but not global structure; apparent distances between distant clusters should not be interpreted as meaningful similarity measures. The human corpus is restricted to NLP/ML venues from 2018–2021, and the LLM-generated ideas are restricted to the same topical scope; the comparison does not extend to other disciplines or time periods.

E.4 Academic Writing

The two case-study papers (RADS and ditMem) are carried through to full manuscript drafting by the vibe research system, targeting ICLR 2026 and ECCV 2026 respectively (Figure 12).

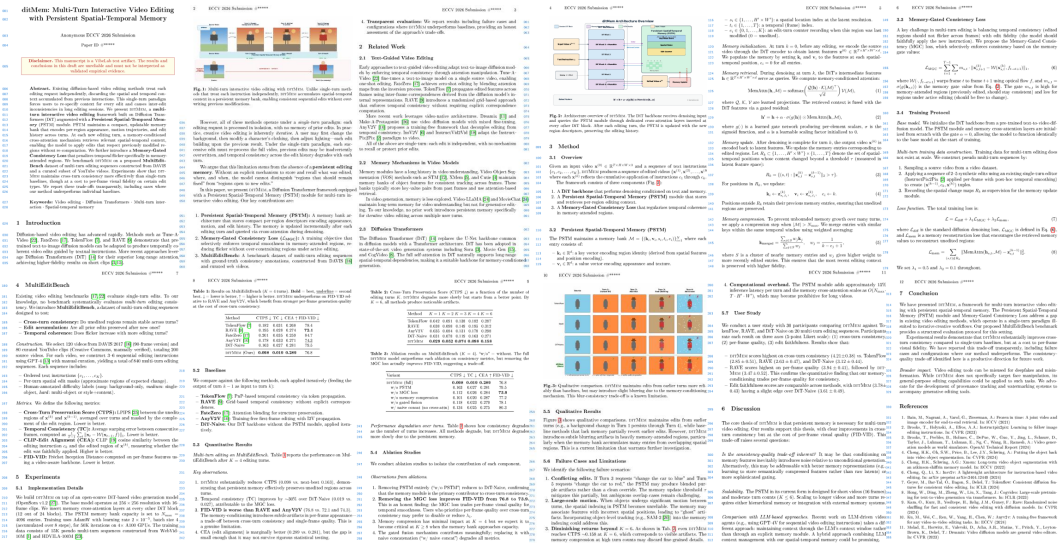
Both manuscripts follow the structural conventions of their target venues: abstract, introduction, related work, method, experiments, ablation, conclusion. Internal consistency is maintained from abstract through conclusion, and results are presented with quantitative specificity (e.g., RADS reports per-benchmark accuracy deltas; ditMem reports per-metric scores against each baseline).

Two limitations are common to both manuscripts. First, the writing exhibits a uniform register that lacks the rhetorical variation found in the strongest human-written papers; transitions between sections and motivational framing are competent but formulaic. Second, the system does not spontaneously identify and address reviewer objections during writing (e.g., the ditMem manuscript proposes and evaluates MultiEditBench without discussing the self-benchmarking concern flagged in the independent review reported in Appendix B).

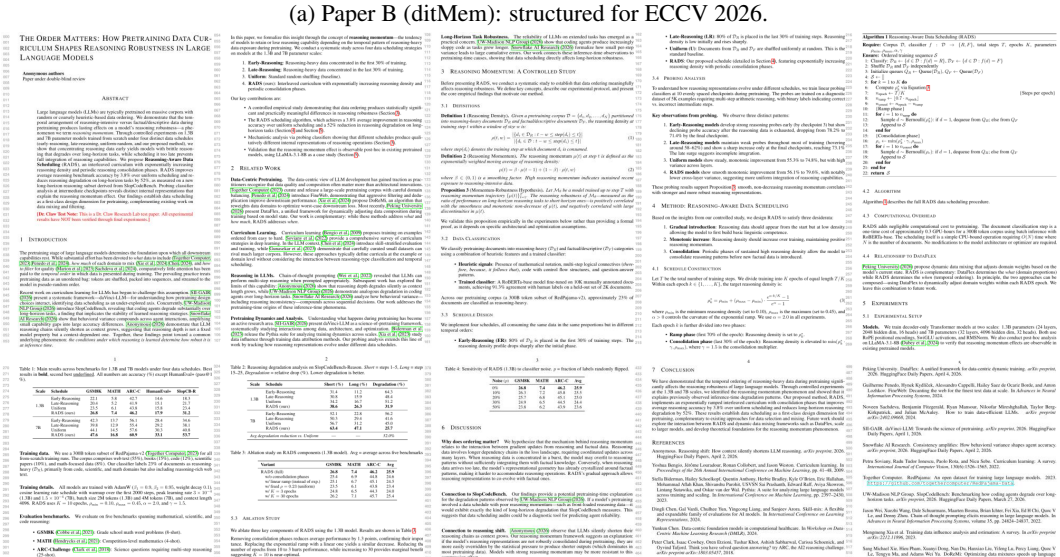
E.5 Limitations

Case study selection. Both case studies (RADS and ditMem) are produced by the same vibe research system and fall within AI/ML. Whether the iterative refinement process, compositional patterns, and writing quality generalize to other systems or disciplines remains untested.

Human baseline representativeness. The human-authored ideas used in the composition analysis and the embedding comparison are drawn from published papers (2018–2022) in NLP/ML venues.



(a) Paper B (ditMem): structured for ECCV 2026.



(b) Paper A (RADS): structured for ICLR 2026.

Figure 12: Full manuscripts produced by the vibe research system for the two case studies, targeting different subdomains (computer vision and NLP) and venues (ECCV and ICLR).

These represent finished contributions rather than the full space of ideas that human researchers considered and discarded. The comparison therefore understates the breadth of human ideation.

Embedding model limitations. Sentence-BERT encodes surface-level semantic content. Two ideas may appear close in embedding space because they share vocabulary (e.g., "attention" and "transformer") while differing in technical depth. The 19-category clustering is coarse; finer-grained categories might reveal systematic differences between human and machine ideas that the current analysis does not capture.

Maturity percentages. The maturity estimates in Tables 6 and 7 are assigned by the system itself and are not calibrated against an external standard. They should be read as rough indicators of progression rather than precise measurements.