# IDEAPRISM: BEYOND GROUND-TRUTH MATCHING FOR OPEN-ENDED HYPOTHESIS EVALUATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

AI scientists generate open-ended hypothesis sets in which several scientifically plausible explanations may fit the same evidence, including mechanisms absent from a predefined target hypothesis. Many existing evaluations, however, score outputs by matching them to designated discoveries or by asking human or LLM judges for direct quality assessments. Direct ratings introduce evaluator dependence, while target matching can miss alternative hypotheses and obscure the breadth of a generated set. We introduce IDEAPRISM, a framework for open-ended evaluation of scientific hypothesis generation. IDEAPRISM characterizes hypothesis sets along three dimensions. Exploration Novelty measures how rarely a mechanism appears in a fixed reference population, with scores adjusted for estimated reference coverage. Diversity measures the effective number of scientific directions represented in a hypothesis set. Value measures the change in a fixed reasoner's accuracy on hidden-evidence questions when conditioned on a hypothesis. We evaluate IDEAPRISM on 61 topics spanning 13 scientific subfields. Controlled experiments assess reference-coverage estimation, effective-direction recovery, and hypothesis-conditioned prediction. We then apply the framework to ten AI-scientist scaffolds. The resulting profiles show that improvements in reference-relative exploration or predictive utility do not necessarily coincide with broader exploration across scientific directions. IDEAPRISM complements target-recovery benchmarks by providing an operational, multidimensional characterization of generated hypothesis spaces.

## 1 INTRODUCTION

Scientific hypothesis generation is inherently open-ended. Given the same observations, multiple mechanisms may remain scientifically plausible, and a useful hypothesis need not reproduce the explanation that later appeared in the literature. This creates a mismatch with closed-ended evaluation: a benchmark built around one designated discovery can determine whether that answer was recovered, but cannot by itself characterize plausible alternatives outside the target set or the breadth of a generated hypothesis set.

Existing evaluations largely follow two paradigms. Rating-based approaches ask humans or LLMs to judge properties such as novelty or feasibility, making the resulting scores dependent on evaluator judgments (Si et al., 2025; Guo et al., 2025). Automatic novelty judgments may disagree with expert assessments (Schopf & Färber, 2026), and recent audits report both evaluator disagreement and convergence in generated hypothesis spaces (Bao et al., 2026). Target-based benchmarks instead measure whether generated hypotheses recover designated discoveries (Liu et al., 2026; Majumder et al., 2025). They provide a clear recovery criterion, but hypotheses outside the designated target set receive no credit, even when they offer distinct explanations compatible with the available evidence. A single target also captures only one direction when several explanations remain plausible. A complementary line of work grounds individual properties in observable quantities, for example using generation frequencies to operationalize originality (Bangash et al., 2025) or prediction tasks to measure hypothesis utility (Liu et al., 2025). This raises a broader question: *how can we evaluate open-ended scientific hypotheses without reducing them to direct quality ratings or recovery of a designated answer?*



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.

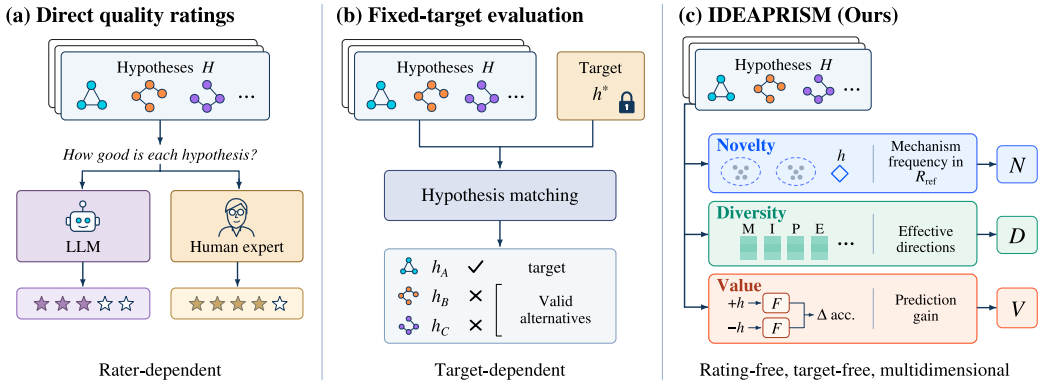


**Figure 1:** Comparison of existing hypothesis-evaluation paradigms and IDEAPRISM. (a) Direct quality ratings ask human or LLM judges to score generated hypotheses, making evaluation dependent on the rater. (b) Fixed-target evaluation scores hypotheses by matching them against a designated target hypothesis $h^*$, tying credit to a predefined answer. (c) IDEAPRISM evaluates the same hypothesis set along three separate dimensions: exploration novelty, effective-direction diversity, and hypothesis-conditioned predictive utility.

We introduce IDEAPRISM, a framework that avoids direct scalar quality ratings for evaluating scientific hypotheses generated by AI scientists (Figure 1). IDEAPRISM reports three complementary measurements: **Exploration Novelty** measures how rarely a hypothesis mechanism appears in a fixed reference population of direct LLM generations, with scores adjusted for estimated reference coverage. **Diversity** measures the effective number of scientific directions represented by a hypothesis set across mechanisms, interventions, predictions, and proposed experiments. **Value** measures how much conditioning on a hypothesis changes a fixed reasoner's accuracy on questions derived from evidence hidden from the generating system.

The three measurements are computed from reference frequencies, structured agreement, and paired accuracy changes; LLMs provide matching, extraction, and question answering. We evaluate each estimator in controlled experiments designed around its intended operational quantity and then evaluate AI-scientist outputs. In a nutshell, our contributions are summarized as follows:

**(1) Open-ended hypothesis evaluation.** We formulate scientific hypothesis evaluation beyond designated-target recovery, treating generated outputs as a set of possible scientific directions.

**(2) Operational measurements and validation.** We introduce measurements of reference-relative novelty, scientific-direction diversity, and predictive value on 61 topic packs spanning 13 scientific subfields. Controlled experiments test reference coverage, direction recovery, and hypothesis-conditioned prediction (§3, §4, §5).

**(3) Evaluation of AI-scientist hypothesis sets.** We use IDEAPRISM to characterize how different AI-scientist workflows change exploration, directional breadth, and predictive utility (§6).

## 2 Related Work

**AI scientist systems and broader scientific benchmarks.** AI scientist systems combine language models with retrieval, review, experimentation, or evolutionary search to support scientific research (Lu et al., 2024; Yamada et al., 2025; Baek et al., 2025; Su et al., 2025; Schmidgall et al., 2025; Gottweis et al., 2026; Lyu et al., 2026). Several methods focus specifically on literature-guided hypothesis generation, using prior scientific work to identify research inspirations and propose new hypotheses (Wang et al., 2024; Zhou et al., 2024; Yang et al., 2026; Ueda et al., 2025). Existing benchmarks evaluate broader scientific capabilities: ScienceAgentBench focuses on data-driven tasks through executable programs (Chen et al., 2025), MLE-bench evaluates machine-learning engineering against competition outcomes (Chan et al., 2025), EXP-Bench evaluates research experiments (Kon et al., 2026), and AstaBench covers a broader suite of scientific research tasks (Bragg et al., 2026). IDEAPRISM focuses on the hypothesis-generation stage and evaluates the resulting hypothesis sets.

**Table 1:** Comparison with prior hypothesis-evaluation benchmarks. ✓ indicates a reported property. No direct ratings denotes scoring without direct human or LLM quality ratings; No fixed target denotes scoring without matching a designated hypothesis.

| Study / benchmark | Scope and evaluation | | | | Scoring design | |
|---|---|---|---|---|---|---|
| | Multi-domain | Novelty assessment | Set diversity | Predictive gain | No direct ratings | No fixed target |
| IdeaBench (KDD 2025) | ✓ | ✓ | ✗ | ✗ | ✗ | ✗ |
| MuseScorer (EMNLP 2025) | ✗ | ✓ | ✗ | ✗ | ✓ | ✓ |
| RINoBench (LREC 2026) | ✗ | ✓ | ✗ | ✗ | ✗ | ✓ |
| ResearchBench (ACL Findings 2026) | ✓ | ✗ | ✗ | ✗ | ✗ | ✗ |
| DiscoveryBench (ICLR 2025) | ✓ | ✗ | ✗ | ✗ | ✗ | ✗ |
| ProjectionBench (arXiv preprint, 2026) | ✗ | ✓ | ✗ | ✗ | ✓ | ✗ |
| LiveIdeaBench (Nature Communications, 2026) | ✓ | ✓ | ✓ | ✗ | ✗ | ✓ |
| AgentIdeaBench (arXiv preprint, 2026) | ✓ | ✓ | ✓ | ✗ | ✗ | ✓ |
| **IDEAPRISM (ours)** | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

**Hypothesis evaluation.** Prior work evaluates generated research ideas through direct quality judgments, comparison with predefined targets, or operational proxies for individual properties. IdeaBench combines novelty and feasibility rankings with target-idea overlap, while expert studies directly rate research ideas and analyze generation redundancy (Guo et al., 2025; Si et al., 2025). RINoBench benchmarks automated novelty judgments against expert-derived scores and justifications (Schopf & Färber, 2026). LLM-based judgments can exhibit presentation and self-preference biases (Zheng et al., 2023; Panickssery et al., 2024), and automatic novelty judgments may disagree with expert assessments (Schopf & Färber, 2026; Bao et al., 2026).

Target-based benchmarks evaluate recovery of designated discoveries or source-paper claims (Liu et al., 2026; Majumder et al., 2025; Lew et al., 2026). Open-ended benchmarks measure properties such as originality and redundancy (Bangash et al., 2025; Ruan et al., 2026; Mo et al., 2026). IDEAPRISM combines reference-relative novelty, set-level diversity, and predictive gain in one evaluation protocol (Table 1).

## 3 THE IDEAPRISM BENCHMARK

### 3.1 TASK AND INFORMATION BOUNDARY

Each topic is represented as $T_i = (Q_i, D_i^{\text{vis}}, R_i^{\text{ref}}, D_i^{\text{hid}})$, where $Q_i$ is the research question, $D_i^{\text{vis}}$ contains observations visible to the AI scientist, $R_i^{\text{ref}}$ defines the reference-generator population, and $D_i^{\text{hid}}$ contains findings reserved for evaluation. Given $(Q_i, D_i^{\text{vis}})$, each AI scientist produces a hypothesis set $H_i$. Retrieval is restricted to the supplied observation summaries, and scaffolds may search these inputs and revise their proposals before returning the final hypothesis set. Evaluation covers hypothesis generation; the reference sample and hidden findings remain evaluator-side.

The three measurements operate at different comparison levels. Exploration Novelty compares each $h \in H_i$ against a sample from $R_i^{\text{ref}}$; Diversity compares scientific components across hypotheses within $H_i$; and Value measures how conditioning on $h$ changes predictions on questions derived from $D_i^{\text{hid}}$. These comparisons yield an $(N, D, V)$ profile for each hypothesis set (Figure 2).

### 3.2 CONSTRUCTION AND ADMISSION

The benchmark contains 61 topic packs spanning 13 scientific subfields covering computational, life, physical, and environmental sciences. Each pack is anchored at an evidence cutoff: pre-cutoff observations are visible to the generating system, while later findings are reserved for evaluation. Topics are selected so that the research question concerns an unresolved mechanism at the cutoff. Visible and hidden evidence are supported by separate citations.

Each pack contains 12 mechanically scored hidden-evidence questions, at least six of which require transfer beyond a verbatim reported finding. Value additionally applies a reasoner-specific headroom screen to determine topic eligibility (§4.3), whereas Exploration Novelty and Diversity are evaluated on the 51 held-out packs. Appendix A provides the full topic catalog, and Appendix D reports the Value cohorts.

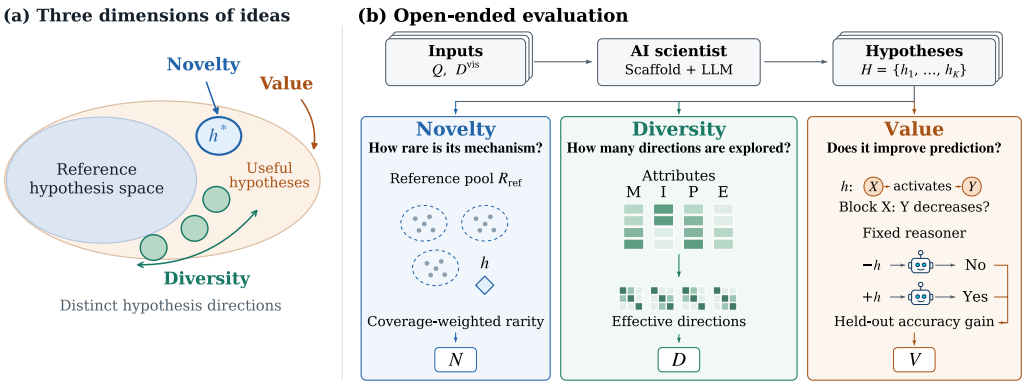


**Figure 2:** IDEAPRISM scores open-ended hypothesis sets from visible inputs. **(a)** Exploration Novelty is reference-relative, Diversity counts directions, and Value is predictive utility. The regions are conceptual, not score thresholds. **(b)** Each AI-scientist generates a set of hypotheses. The three evaluation branches measure exploration novelty using coverage-scaled reference counts, diversity using effective direction counts, and value using paired prediction gain.

### 3.3 DEVELOPMENT AND HELD-OUT EVALUATION

We developed the measurements on 10 AI topics and froze all definitions, prompts, and scoring procedures before evaluating the remaining 51 held-out topics (Appendix A). All system comparisons are restricted to these 51 held-out topics, with metric-specific eligibility applied for Diversity and Value.

## 4 OPERATIONALIZING OPEN-ENDED EVALUATION

### 4.1 EXPLORATION NOVELTY

Exploration Novelty $N(h)$ assigns higher scores to candidate mechanisms that occur less frequently in a fixed reference population of direct LLM generations. For each topic, let $\mathcal{R} = \{u_1, \ldots, u_n\}$ denote reference hypotheses generated directly from the research question and visible observations. The system study samples 30 hypotheses independently from each of five model families ($n = 150$), with equal representation across families. This reference population is fixed across evaluated systems. Thus, $N$ measures rarity relative to this fixed generator population. A high $N(h)$ indicates that the mechanism is rare relative to the reference population, capturing exploration beyond commonly generated mechanisms. E1 evaluates coverage of this reference population (§5.1).

A two-pass LLM matcher groups equivalent reference mechanisms into modes $B_1, \ldots, B_b$; human validation is reported in Appendix F. For mode $B_j$, let $r_j$ denote its reference count and $N_r$ the number of modes observed exactly $r$ times. We estimate the unseen mass using Good–Turing estimation (Good, 1953; McAllester & Schapire, 2000):

$$r_j = \sum_{a=1}^{n} \mathbf{1}[u_a \in B_j], \qquad N_r = \sum_{j=1}^{b} \mathbf{1}[r_j = r], \qquad \hat{P}_{\text{unseen}} = \frac{N_1}{n}, \quad n = \sum_{j=1}^{b} r_j. \tag{1}$$

The reference modes are fixed before evaluated hypotheses are matched. If a candidate $h$ matches mode $B_j$, we set $r(h) = r_j$; otherwise, $r(h) = 0$. Its Exploration Novelty score is

$$N(h) = \left(1 - \hat{P}_{\text{unseen}}\right) \frac{1}{r(h) + 1}. \tag{2}$$

Within a topic, the coverage term $1 - \hat{P}_{\text{unseen}}$ is shared by all candidates, so their ordering depends only on $1/(r(h) + 1)$: unseen mechanisms rank above singleton modes, which in turn rank above more frequent reference modes, with equal counts tied. Across topics, the coverage term rescales all candidate scores, discounting rarity when the reference population is estimated to be less well covered.

For a generated hypothesis set $H$ of size $K$, we report

$$N_{\text{mean}}(H) = \frac{1}{K} \sum_{h \in H} N(h), \qquad N_{\max}(H) = \max_{h \in H} N(h). \tag{3}$$

Here, $N_{\text{mean}}$ summarizes reference-relative novelty across the set, whereas $N_{\max}$ captures its most reference-rare hypothesis.

### 4.2 Diversity: effective scientific directions

Diversity measures the effective number of sufficiently specified scientific directions represented in a hypothesis set. Given $H = \{h_1, \ldots, h_K\}$, we compute a separate effective direction count for each component $f \in \mathcal{F} = \{M, I, P, E\}$: Mechanism, Intervention/Method, Prediction, and Experiment.

**Step 1: Extract scientific attributes.** An LLM maps each hypothesis to structured attributes for Mechanism ($M$: cause, locus, effect sign), Intervention/Method ($I$: method family, target), Prediction ($P$: observable, direction), and Experiment ($E$: design, comparison). Slot values are resolved against topic-local vocabularies constructed from the visible inputs, with open-set values canonicalized within each topic (Appendix E).

For component $f$, hypotheses with all required slots specified form the scorable set

$$\mathcal{I}_f = \{i : z_{ifs} \neq \perp \; \forall s \in \mathcal{S}_f\}, \qquad m_f = |\mathcal{I}_f|. \tag{4}$$

Hypotheses with missing slots are excluded from the corresponding component.

**Step 2: Construct component-agreement kernels.** Let $\mathcal{V}_{fs}^{+}$ denote the union of the frozen base vocabulary and canonicalized open-set values for slot $s$. For each retained hypothesis, encode its slot value as a one-hot vector and stack these vectors in $X_{fs}$. We define the component kernel as the average slot-wise agreement:

$$K_f = \frac{1}{|\mathcal{S}_f|} \sum_{s \in \mathcal{S}_f} X_{fs} X_{fs}^{\top}, \qquad K_f(i,j) = \frac{1}{|\mathcal{S}_f|} \sum_{s \in \mathcal{S}_f} \mathbf{1}[z_{ifs} = z_{jfs}]. \tag{5}$$

As an average of Gram matrices, $K_f$ is positive semidefinite by construction.

**Step 3: Compute effective direction counts.** Let $\mu_{f\ell} \geq 0$ be the eigenvalues of $K_f$, and normalize them as

$$\lambda_{f\ell} = \frac{\mu_{f\ell}}{\sum_{r=1}^{m_f} \mu_{fr}}, \tag{6}$$

$$D_f(H) = \exp\left( - \sum_{\ell=1}^{m_f} \lambda_{f\ell} \log \lambda_{f\ell} \right). \tag{7}$$

Identical component records yield one effective direction, whereas mutually distinct records ($K_f = I$) yield $m_f$ directions. For ideal direction blocks with proportions $p_j$, $D_f$ reduces to the order-one effective count $\exp(H(p))$ (Hill, 1973), assigning 1.38 effective directions to a $(9, 1)$ split and 2 to a balanced $(5, 5)$ split.

**Step 4: Report direction count, dispersion, and coverage.** For each component, we report the effective direction count $D_f$, its normalized dispersion among scorable hypotheses, and specification coverage:

$$D(H) = (D_M, D_I, D_P, D_E), \qquad \widetilde{D}_f = \frac{D_f - 1}{m_f - 1}, \qquad C_f = \frac{m_f}{K}. \tag{8}$$

An axis requires $m_f \geq 2$. For $m_f \geq 2$, $C_f$ records specification coverage.

The same extraction and scoring pipeline is used in E2 and in the system evaluation. Numerical details and the optional four-axis aggregate are given in Appendix E.

### 4.3 VALUE: CONDITIONAL PREDICTIVE UTILITY

Value measures whether a hypothesis improves prediction of evidence hidden from the generating system. A fixed reasoner $F$ answers the same hidden-evidence questions with and without hypothesis $h$. Let $S^{-h}$ and $S^{+h}$ denote the corresponding accuracies. We define

$$\hat{V}(h) = S^{+h} - S^{-h}. \tag{9}$$

To retain prediction headroom, we evaluate Value on the reasoner-specific topic cohort

$$\mathcal{T}_F = \{t : S_t^{-h}(F) \leq 0.85\}. \tag{10}$$

E3 applies this screen separately to GPT-4.1 and Gemini-2.5-Flash.

For a hypothesis set $H_t$ on a retained topic, we report

$$V_{\text{mean}}(H_t) = \frac{1}{K} \sum_{h \in H_t} \hat{V}(h), \qquad V_{\max}(H_t) = \max_{h \in H_t} \hat{V}(h). \tag{11}$$

System-level results average these summaries equally across retained topics. Prompting, missing-response handling, and statistical tests are specified in Appendix D.

## 5 METRIC VALIDATION

The three measurements operationalize different aspects of the open-ended evaluation view. We therefore validate the specific quantity each estimator is intended to recover: finite-reference coverage for Exploration Novelty, effective direction counts for Diversity, and conditional prediction gain for Value. Appendix G specifies the synthetic and real-data protocols.

### 5.1 E1: REFERENCE-COVERAGE CALIBRATION

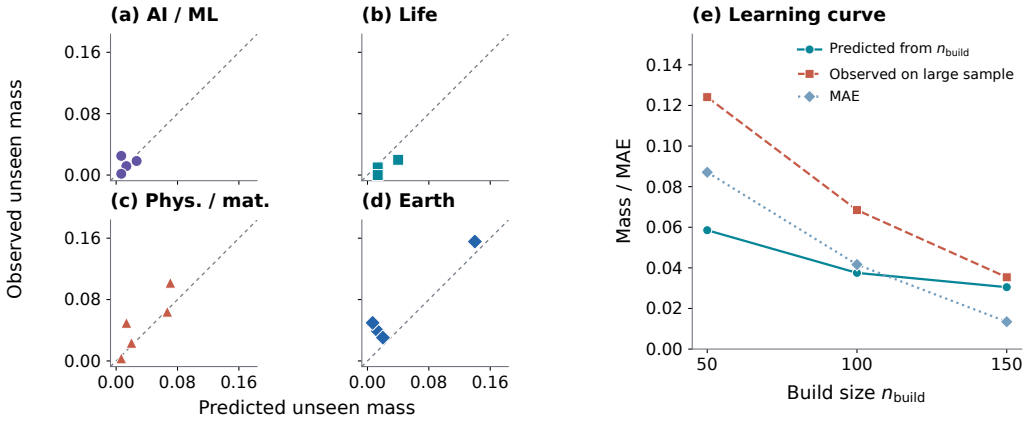


**Figure 3:** E1: in-pool unseen-mass calibration. (a–d) Predicted unseen mass $\hat{P}_{\text{unseen}}$ from $n_{\text{build}} = 150$ versus the unseen share of an independent sample; the dashed line is $y = x$. (e) Unseen-mass estimation error as the build sample grows.

E1 evaluates whether the Good–Turing estimate recovers the unseen mass of the five-family reference population used for scoring. For each topic, the build sample contains 30 hypotheses from each of five families ($n = 150$), and the estimated unseen mass is compared with the unseen share of a larger independent sample from the same generator population (Figure 3).

At the operating pool size, the estimate closely tracks the empirical unseen share, with MAE $0.008$ on AI topics and $0.017$ on science topics. Estimation error also decreases as the build sample grows from 50 to 150 references. Leaving out one reference family changes the estimated unseen mass by a median of $0.003$ on AI topics and $0.013$ on science topics.

We also test reference-pool sensitivity on actual AI-scientist outputs. On the 25 held-out topics with complete outputs from all 11 systems, mean Spearman correlation with the full $n = 150$ ranking is $0.963$ at $n = 30$ and $0.994$ at $n = 120$. The full analysis is reported in Appendix H.1.

### 5.2 E2: Effective direction recovery

We construct 5,120 controlled sets from 320 published-paper claims spanning eight topics. Each set contains ten claims arranged in one of five direction patterns: $(10)$, $(9,1)$, $(5,5)$, $(4,3,3)$, or $(4,2,2,2)$, with 32 repetitions per configuration. We score each set with the same Diversity pipeline used for system evaluation. Author-provided direction labels are withheld until scoring eligibility is fixed and are then used to compute the target effective count $G^{\text{eff}}$ (Appendix G).

All 5,120 sets are scorable. Overall, the estimator recovers the controlled direction counts with pooled MAE 0.523 and accuracy within one effective direction of 0.820 (Table 2). Recovery is strongest for Prediction and remains accurate for Mechanism and Intervention, while Experiment is more challenging and performs similarly to the strongest clustering baseline.

**Table 2:** E2 recovery versus $G^{\text{eff}}$ ($K$=10; 5,120 sets from 320 papers). $\text{Acc}_{\pm 1}$ requires $|\widehat{G}_{\text{eff}} - G_{\text{eff}}| \leq 1$. Errors use the items retained after complete-signature filtering. Gold labels enter only then. Full judge results are in Table 15.

| | MAE↓ | | | | $\text{Acc}_{\pm 1}$ ↑ | | | |
|---|---|---|---|---|---|---|---|---|
| Method | $M$ | $I$ | $P$ | $E$ | $M$ | $I$ | $P$ | $E$ |
| TF-IDF cluster | 0.86 | 0.75 | 0.84 | **0.71** | 0.68 | 0.74 | 0.65 | **0.72** |
| Embed. MiniLM | 1.28 | 2.69 | 0.43 | 1.64 | 0.56 | 0.21 | 0.79 | 0.46 |
| NLI semantic entropy | 1.24 | 1.41 | 1.18 | 1.07 | 0.57 | 0.51 | 0.58 | 0.62 |
| LLM Judge (GPT-4.1) | 1.21 | 3.42 | 1.56 | 3.22 | 0.56 | 0.05 | 0.47 | 0.15 |
| LLM Judge (Gemini-2.5-Flash) | 1.74 | 3.46 | 2.25 | 3.30 | 0.49 | 0.11 | 0.38 | 0.21 |
| **Ours** | **0.49** | **0.55** | **0.32** | 0.73 | **0.82** | **0.83** | **0.92** | 0.71 |

### 5.3 E3: Conditional predictive utility

We validate Value using controlled hypothesis treatments: Supported (A), Partial (B), Irrelevant (C), and Misleading (D).

**Two-reasoner validation.** GPT-4.1 and Gemini-2.5-Flash independently evaluate the same 51 held-out topic packs at $T = 0$. Each pack contains 12 multiple-choice questions derived from hidden papers dated on or after 2025-05-01, including six transfer questions. For each reasoner, we first measure the no-hypothesis baseline and retain topics with $S^{-h} \leq 0.85$ to ensure sufficient headroom. A closed-book check is additionally used to monitor prior exposure to the hidden evidence.

Table 3(a) summarizes the treatment results. Both reasoners have mean order $A > B > C > D$. Gemini-2.5-Flash's 95% interval for Partial includes zero. Irrelevant is negative on both reasoners, below the no-hypothesis baseline. Additional intervals and controls are reported in Appendices F and D. Gemini-2.5-Flash is the fixed Value reasoner in the system study.

**Scientific-content ablation.** We compare the intact Supported hypothesis with four perturbations using Qwen-Plus-2025-04-28: removing the prediction, removing the mechanism, shuffling the words, and replacing the hypothesis with length-matched meaningless text. Table 3(b) reports paired differences in Value between the intact and perturbed hypotheses.

**Table 3:** E3 validation of conditional predictive utility. (a) Mean paired $V$ for Supported (A), Partial (B), Irrelevant (C), and Misleading (D). (b) Scientific-content ablations with Qwen-Plus-2025-04-28, reporting $\Delta V = V_{\text{ablation}} - V_{\text{full}}$. Full A–D intervals are in Table 12.

**(a) Treatment validation**

| Reasoner | A | B | C | D |
|---|---|---|---|---|
| GPT-4.1 | +0.175 | +0.052 | −0.115 | −0.440 |
| Gemini-2.5-Flash | +0.218 | +0.030 | −0.111 | −0.381 |

**(b) Scientific-content ablation**

| Ablation | $\Delta V$ |
|---|---|
| No prediction | −0.102 |
| No mechanism | −0.146 |
| Shuffled words | −0.384 |
| Length-matched nonsense | −0.555 |

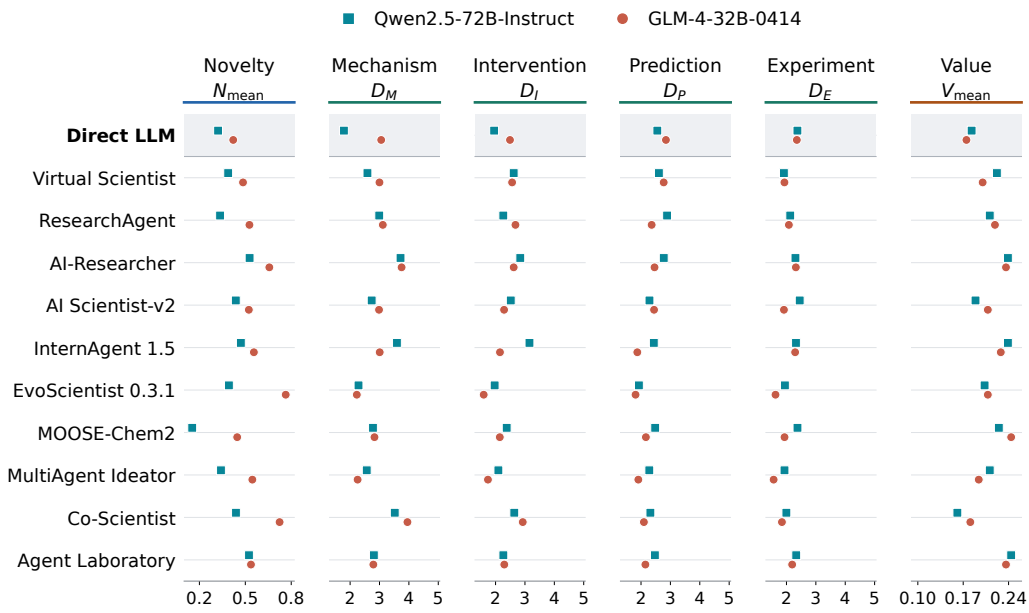


**Figure 4: Hypothesis profiles across 22 systems.** Each row is a workflow; color and marker identify its generator. Direct LLM is shaded. Panels show Novelty, four effective direction counts, and Value. Value uses Gemini-2.5-Flash throughout.

All four perturbations reduce Value, with substantially larger drops for shuffled and meaningless text than for removing a single scientific component.

Value also has higher adjacent-pair accuracy point estimates than direct LLM judging in all six matched comparisons across GPT-4.1 and Gemini-2.5-Flash. Full results are in Appendix D.1.

## 6 Evaluating AI-Scientist Scaffolds

### 6.1 Systems and evaluation protocol

We evaluate 22 AI-scientist systems formed by ten scaffolds and Direct LLM with Qwen2.5-72B-Instruct and GLM-4-32B-0414. The scaffolds are ResearchAgent (Baek et al., 2025), InternAgent 1.5 (Feng et al., 2026), Virtual Scientist (Su et al., 2025), AI-Researcher (Tang et al., 2026), AI Scientist-v2 (Yamada et al., 2025), EvoScientist 0.3.1 (Lyu et al., 2026), MOOSE-Chem2 (Yang et al., 2026), MultiAgent Research Ideator (Ueda et al., 2025), Co-Scientist (Gottweis et al., 2026), and Agent Laboratory (Schmidgall et al., 2025).

Each system receives the same visible inputs $(Q, D^{\text{vis}})$ and produces ten hypotheses per topic. Generation and review use the corresponding generator; all Value scores use Gemini-2.5-Flash. Single-idea workflows are repeated to obtain ten hypotheses; multi-idea workflows retain native selection. Direct LLM uses independent calls disjoint from the reference pool. Novelty is averaged over 51 held-out topics; Diversity and Value use their metric-specific eligible cohorts. Details are in Appendix C.

### 6.2 Comparing multidimensional system profiles

**System strengths differ across dimensions.** Under Qwen, AI-Researcher has the highest point estimate for Novelty and Mechanism, while Agent Laboratory has the highest point estimate for Value. Under GLM, EvoScientist has the highest point estimate for Novelty, Co-Scientist for Mechanism and Intervention/Method, and MOOSE-Chem2 for Value; Direct has the highest point estimate for Prediction and Experiment. The profiles also differ within a workflow: Co-Scientist is

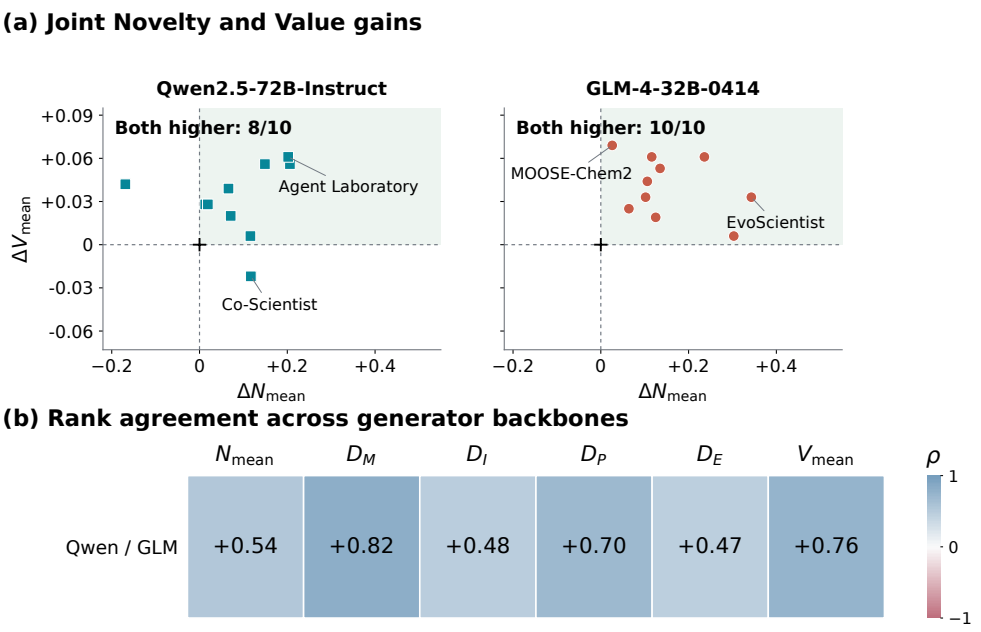


**Figure 5: Joint gains and scaffold rank agreement.** (a) Scaffold-minus-Direct changes in $N_{\text{mean}}$ and $V_{\text{mean}}$; the shaded quadrant marks gains in both. (b) Spearman correlation between Qwen and GLM scaffold rankings for each metric.

strong on Mechanism and Intervention/Method but weaker on Value, while EvoScientist reaches the highest GLM Novelty with lower Diversity on several components.

**Novelty and Value often rise together, while Diversity changes are component-specific.** Both metrics increase for 8/10 Qwen scaffolds and all ten GLM scaffolds (Figure 5a), with mean changes of $(+0.079, +0.031)$ and $(+0.156, +0.040)$ in $(\Delta N_{\text{mean}}, \Delta V_{\text{mean}})$. Under Qwen, all ten scaffolds increase Mechanism and Intervention/Method diversity, but only 3/10 increase Prediction and 2/10 Experiment. Under GLM, all ten fall below Direct on Prediction and Experiment.

**Rank agreement varies by metric.** Qwen and GLM rank the scaffolds similarly on Mechanism ($\rho = 0.82$) and Value ($\rho = 0.76$), with lower agreement on Novelty ($0.54$), Intervention/Method ($0.48$), and Experiment ($0.47$) (Figure 5b). Prediction lies between these groups ($\rho = 0.70$). Mechanism and Value show the strongest cross-generator rank agreement.

## 7 Limitations

The benchmark covers a finite set of topics and evaluates hypotheses before experimental execution; broader domains and downstream outcomes remain outside its scope. Exploration Novelty is reference-relative rather than a measure of absolute scientific novelty, while Value captures predictive utility rather than broader scientific importance or impact.

## 8 Conclusion

Scientific hypothesis generation is open-ended, while many existing evaluations rely on direct quality judgments or recovery of a designated answer. IdeaPrism provides a complementary way to characterize generated hypothesis sets through reference-relative exploration, scientific-direction diversity, and hypothesis-conditioned predictive utility. Across 61 topic packs, controlled studies validate the corresponding estimators, while the system study reveals workflow differences that are not captured by a single quality score or target-recovery criterion. IdeaPrism complements these evaluation paradigms by providing a structured view of open-ended hypothesis sets.

## AI USE STATEMENT

Generative AI tools assisted with measurement and simulation code, manuscript editing, literature lookup and offline result checks. LLMs drafted the conceptual question packs and treatment hypotheses described in Appendix D. They also serve as the evaluated agents, reference generators, matchers, parsers, and Value reasoners. We retain model responses, extracted attributes, and evaluation settings for reproducibility. The authors are responsible for the final text, citations, experimental claims, and interpretation.

## ETHICS STATEMENT

IDEAPRISM evaluates hypotheses and proposed discriminating experiments; it does not execute those experiments. Topic evidence comes from published literature. Evaluation access separates the visible inputs from hidden findings and answer keys. One annotator with a Ph.D. in computer science reviewed Value question validity and treatment interpretation. Two annotators with Ph.D.s in computer science reviewed Diversity slot canonicalization on 200 AI items; agreement is $0.86$. Two annotators with Ph.D.s in computer science reviewed Novelty mode assignments on 100 pairs from AI topics; agreement with the matcher is $0.82$. Details are specified in Appendix F. These audits validate the scoring components but do not enter benchmark scores. Appendix I describes the evaluation boundary.

## REPRODUCIBILITY STATEMENT

The accompanying implementation contains the three scorers, seeded synthetic studies, topic cards, conceptual question packs, and model outputs. Appendix E specifies prompts and evaluation settings; Appendix C describes the AI-scientist algorithms, task adaptations, and scoring conditions. System results use the scoring conditions in Appendix C. Annotation instructions are provided in Appendix F.

## REFERENCES


Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. Researchagent: Iterative research idea generation over scientific literature with large language models. In *Proceedings of the 2025 conference of the nations of the Americas chapter of the association for computational linguistics: human language technologies (volume 1: long papers)*, pp. 6709–6738, 2025.

Ali Sarosh Bangash, Krish Veera, Ishfat Abrar Islam, and Raiyan Abdul Baten. Musescorer: Idea originality scoring at scale. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 19947–19965, 2025.

Honglin Bao, Siyang Wu, Xiao Liu, Sida Li, Shiyun Cao, and James A Evans. Contemporary ai lacks the imagination to diverge or negate in science. *arXiv preprint arXiv:2606.08251*, 2026.

Jonathan Bragg, Mike D'Arcy, Nishant Balepur, Dan Bareket, Bhavana Dalvi Mishra, Sergey Feldman, Dany Haddad, Jena Hwang, Peter Jansen, Varsha Kishore, et al. Astabench: Rigorous benchmarking of ai agents with a scientific research suite. In *International conference on learning representations*, volume 2026, pp. 110136–110223, 2026.

Jun Shern Chan, Neil Chowdhury, Oliver Jaffe, James Aung, Dane Sherburn, Evan Mays, Giulio Starace, Kevin Liu, Leon Maksin, Tejal Patwardhan, et al. Mle-bench: Evaluating machine learning agents on machine learning engineering. In *International Conference on Learning Representations*, volume 2025, pp. 50466–50494, 2025.

Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, et al. Scienceagentbench: Toward rigorous assessment of language agents for data-driven scientific discovery. In *International Conference on Learning Representations*, volume 2025, pp. 96934–96990, 2025.

Shiyang Feng, Runmin Ma, Xiangchao Yan, Yue Fan, Yusong Hu, Songtao Huang, Shuaiyu Zhang, Zongsheng Cao, Tianshuo Peng, Jiakang Yuan, et al. Internagent-1.5: A unified agentic framework for long-horizon autonomous scientific discovery. *arXiv preprint arXiv:2602.08990*, 2026.

Irving J Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3-4):237–264, 1953.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Petar Sirkovic, Artiom Myaskovsky, Grzegorz Glowaty, Felix Weissenberger, Alessio Orlandi, Dan Popovici, et al. Accelerating scientific discovery with co-scientist. *Nature*, pp. 1–3, 2026.

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

Mark O Hill. Diversity and evenness: a unifying notation and its consequences. *Ecology*, 54(2): 427–432, 1973.

Patrick Tser Jern Kon, Qiuyi Ding, Jiachen Liu, Xinyi Zhu, Jingjia Peng, Jiarong Xing, Yibo Huang, Yiming Qiu, Jayanth Srinivasa, Myungjin Lee, et al. Exp-bench: Can ai conduct ai research experiments? In *International Conference on Learning Representations*, volume 2026, pp. 120555–120604, 2026.

Andrew J Lew, Yuan Cao, and Markus J Buehler. Projectionbench: Evaluating scientific hypothesis generation in llms under progressive information disclosure. *arXiv preprint arXiv:2605.30284*, 2026.

Haokun Liu, Yangqiaoyu Zhou, Mingxuan Li, Chenfei Yuan, and Chenhao Tan. Literature meets data: A synergistic approach to hypothesis generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 245–281, 2025.

Yujie Liu, Zonglin Yang, Tong Xie, Jinjie Ni, Ben Gao, Yuqiang Li, Shixiang Tang, Wanli Ouyang, Erik Cambria, and Dongzhan Zhou. Researchbench: Benchmarking llms in scientific discovery via inspiration-based task decomposition. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 13187–13207, 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.

Yougang Lyu, Xi Zhang, Xinhao Yi, Yuyue Zhao, Shuyu Guo, Wenxiang Hu, Jan Piotrowski, Jakub Kaliski, Jacopo Urbani, Zaiqiao Meng, et al. Evoscientist: Towards multi-agent evolving ai scientists for end-to-end scientific discovery. *arXiv preprint arXiv:2603.08127*, 2026.

Bodhisattwa Prasad Majumder, Harshit Surana, Dhruv Agarwal, Bhavana Dalvi Mishra, Abhijeetsingh Meena, Aryan Prakhar, Tirth Vora, Tushar Khot, Ashish Sabharwal, and Peter Clark. Discoverybench: Towards data-driven discovery with large language models. In *International Conference on Learning Representations*, volume 2025, pp. 4556–4579, 2025.

David A McAllester and Robert E Schapire. On the convergence rate of good-turing estimators. In *COLT*, pp. 1–6, 2000.

Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.

Yunxiang Mo, Tianshi Zheng, Yisen Gao, Rui Wang, Newt Nguyen Kim Hue Nam, Kelvin Kiu Wai Tam, Jiaxin Bai, Yangqiu Song, Ginny Wong, and Simon See. Agentideabench: Benchmarking scientific ideation in the agent era. *arXiv preprint arXiv:2609.07611*, 2026.

Arjun Panickssery, Samuel R Bowman, and Shi Feng. Llm evaluators recognize and favor their own generations. *Advances in Neural Information Processing Systems*, 37:68772–68802, 2024.

Kai Ruan, Xuan Wang, Jixiang Hong, Peng Wang, Yang Liu, and Hao Sun. Evaluating llms' divergent thinking capabilities for scientific idea generation with minimal context. *Nature communications*, 17(1):3625, 2026.

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. *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 5977–6043, 2025.

Tim Schopf and Michael Färber. Is this idea novel? an automated benchmark for judgment of research ideas. *arXiv preprint arXiv:2603.10303*, 2026.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. In *International Conference on Learning Representations*, volume 2025, pp. 94003–94092, 2025.

Haoyang Su, Renqi Chen, Shixiang Tang, Zhenfei Yin, Xinzhe Zheng, Jinzhe Li, Biqing Qi, Qi Wu, Hui Li, Wanli Ouyang, et al. Many heads are better than one: Improved scientific idea generation by a llm-based multi-agent system. In *Proceedings of the 63rd annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 28201–28240, 2025.

Jiabin Tang, Lianghao Xia, Zhonghang Li, and Chao Huang. Ai-researcher: Autonomous scientific innovation. *Advances in Neural Information Processing Systems*, 38:9481–9520, 2026.

Keisuke Ueda, Wataru Hirota, Kosuke Takahashi, Takahiro Omi, Kosuke Arima, and Tatsuya Ishigaki. Exploring the design of multi-agent llm dialogues for research ideation. In *Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 322–337, 2025.

Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. Scimon: Scientific inspiration machines optimized for novelty. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 279–299, 2024.

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

Zonglin Yang, Wanhao Liu, Ben Gao, Yujie Liu, Wei Li, Tong Xie, Lidong Bing, Wanli Ouyang, Erik Cambria, and Dongzhan Zhou. Moose-chem2: Exploring llm limits in fine-grained scientific hypothesis discovery via hierarchical search. *Advances in Neural Information Processing Systems*, 38:89045–89076, 2026.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Yangqiaoyu Zhou, Haokun Liu, Tejes Srivastava, Hongyuan Mei, and Chenhao Tan. Hypothesis generation with large language models. In *Proceedings of the 1st Workshop on NLP for Science (NLP4Science)*, pp. 117–139, 2024.

Appendices A–B document benchmark construction and model roles. Appendix C contains the full system results and workflow protocol. Appendix D specifies the Value cohorts and matched comparisons. Appendices E and F record prompts, parser settings, and independent annotation. Appendices G–H provide validation and sensitivity analyses, and Appendix I summarizes scope and dual-use considerations.

## APPENDIX CONTENTS

**A Benchmark Topics and Splits** ..... **15**

**B Model Roles and Temporal Boundaries** ..... **16**

**C System Evaluation and Full Results** ..... **17**

C.1 Aggregation and Within-Backbone Comparisons ..... 17

C.2 Absolute Profiles by Generator Backbone ..... 17

C.3 Scaffold Changes from the Direct LLM Baseline ..... 18

C.4 Directional Dispersion and Specification Coverage ..... 18

C.5 Workflow Adaptation and Scoring Protocol ..... 21

**D Value Protocol and Reasoner Cohorts** ..... **22**

D.1 Value versus Direct Judgments with Matched Models ..... 23

**E Prompts and Evaluation Settings** ..... **24**

**F Independent Annotation** ..... **27**

**G Validation Protocols** ..... **28**

**H Reference-Coverage Sensitivity** ..... **29**

H.1 Reference-Pool Size Sensitivity on System Outputs ..... 29

**I Ethics and Scope** ..... **30**

# A BENCHMARK TOPICS AND SPLITS

This appendix records the topic composition, held-out split, and benchmark scale used throughout the paper.

**Table 4:** Catalog coverage. Each topic card records an evidence cutoff and later findings.

| Macro-domain | Subfields | Time-stamp source | Packs | Evidence cutoff |
|---|---|---|---|---|
| AI/ML | 4 | arXiv | 26 | Topic-specific |
| Life Sciences | 3 | bioRxiv / journals | 11 | Topic-specific |
| Physics & Materials | 4 | arXiv / journals | 18 | Topic-specific |
| Earth & Environment | 2 | ESSOAr / IPCC cycles | 6 | Topic-specific |

**Table 5:** Benchmark scale and configuration.

| Component | Scale / configuration |
|---|---|
| Topic suite | 61 packs spanning 13 scientific subfields |
| Reference pool | 5 families $\times$ 30 hypotheses per topic; 9,150 total |
| Evaluation unit | 10 hypotheses per system/topic; 12 questions per pack |
| System evaluation | 11 systems $\times$ 2 generator backbones $\times$ 51 held-out topics |

**Topic cards and question packs.** A catalog entry records the research question, freeze date, visible observations, and later findings. Generators receive the research question and visible observations; the freeze date defines the evidence boundary, and later findings remain evaluator-side. The conceptual evaluation retains 61 packs: 26 AI/ML, 11 life science, 18 physics/materials, and six earth/environment. Topic freeze dates and model knowledge or release dates serve different purposes and are recorded separately (Appendix B).

**Subfields.** The 13 subfields are LLMs and NLP, vision and multimodality, agents and retrieval, and alignment and safety; molecular and cellular biology, neuroscience, and immunology and microbiology; condensed matter physics, materials science, chemistry, and astrophysics and cosmology; climate science and ecology.

**Development topics.** Ten AI topics informed method development: lost-in-the-middle, VLM image–text conflict, RLHF diversity collapse, long-horizon agent error accumulation, position bias in LLM judges, length bias in RLHF, watermark robustness to paraphrase, benchmark-contamination detection, side effects of activation steering, and chain-of-thought faithfulness. Reported system scores exclude these ten topics and use the other 51.

## B Model Roles and Temporal Boundaries

This appendix records generator, reference, and reasoner roles together with the temporal assumptions used for hidden-evidence evaluation. Hidden findings used to score $V$ are dated on or after 2025-05-01. We define temporal eligibility using provider-documented knowledge cutoffs where available and public model snapshots preceding the hidden findings otherwise. Table 6 records both the configured dates and the corresponding evidence type.

The selected Value reasoners are `gpt-4.1` and `gemini-2.5-flash`, evaluated separately. Their official documentation states knowledge cutoffs of 2024-06-01 and January 2025, respectively (OpenAI model documentation; Google model documentation, checked 2026-09-20). The reported model identifiers are `gpt-4.1-2025-04-14` and `gemini-2.5-flash`. The E3 scientific-content ablation additionally uses the `qwen-plus-2025-04-28` snapshot; reference generation and system Value evaluation use the models listed for those roles in Table 6.

`gpt-4.1`, `gemini-2.5-flash`, `DeepSeek-V3`, `Qwen2.5-72B-Instruct`, and `GLM-4-32B-0414` are the five $R^{\text{ref}}$ families. Qwen2.5-72B-Instruct and GLM-4-32B-0414 serve as generator backbones in Appendix C; Gemini-2.5-Flash is the fixed system Value reasoner, and DeepSeek-V3 is reference-only. For open-weight models without a provider-stated knowledge cutoff, Table 6 uses the public release date as the available snapshot date and labels this evidence type explicitly. The GLM repository, for example, dates the 32B-0414 release to 2025-04-14.

The five-family reference pool contains 30 hypotheses per family and topic, $n = 150$, totaling 9,150 hypotheses across 61 topics. Two-pass matching assigns all reference hypotheses to modes. Scoring a system on a topic requires a valid mode assignment for each of its ten hypotheses. E1 calibrates coverage on these codebooks. The later-trained `deepseek-v4.1-flash` model is used only as the Novelty matcher for evaluator-side mode assignment; it does not generate reference hypotheses or answer Value questions.

**Table 6: Model-date assumptions and reference families.** Hidden findings used for Value are dated $\geq$2025-05-01. $G$ denotes a generator in the main system study; $G_{\text{sens}}$ denotes generation used only for the fixed-output reference-pool sensitivity analysis; $V$ denotes a Value reasoner, and $R_{\text{ref}}$ a reference family.

| Model | Role | Documented date | Evidence type |
|---|---|---|---|
| gpt-4.1 | $G_{\text{sens}}$; $V$; $R_{\text{ref}}$ | 2024-06-01 | OpenAI cutoff |
| gemini-2.5-flash | $V$; $R_{\text{ref}}$ | 2025-01 | Google cutoff |
| DeepSeek-V3 | $R_{\text{ref}}$ | 2024-12-26 | public release |
| Qwen2.5-72B-Instruct | $G$; $R_{\text{ref}}$ | 2024-09-19 | public release |
| GLM-4-32B-0414 | $G$; $R_{\text{ref}}$ | 2025-04-14 | public release |
| qwen-plus-2025-04-28 | E3 ablation | 2025-04-28 | snapshot identifier |

# C System Evaluation and Full Results

This section expands the 22-system comparison in Section 6. We report complete profiles by generator backbone, differences from Direct LLM, and Diversity specification diagnostics, followed by workflow details. Gemini-2.5-Flash is the fixed Value reasoner throughout.

## C.1 Aggregation and Within-Backbone Comparisons

Novelty is averaged over the 51 held-out topics. Diversity and screened Value use their respective eligible topic cohorts. Within each topic, $N_{\text{mean}}$ and $V_{\text{mean}}$ summarize the hypothesis set, while $N_{\max}$ and $V_{\max}$ select its highest-scoring hypothesis; these quantities are then averaged across topics.

For system-level metric $X$, the matched scaffold effect is

$$\delta^X_{s,b} = \overline{X}_{s,b} - \overline{X}_{\text{Direct},b}, \qquad \bar{\delta}^X_b = \frac{1}{10}\sum_{s=1}^{10} \delta^X_{s,b}, \tag{12}$$

where $X \in \{N_{\text{mean}}, D_M, D_I, D_P, D_E, V_{\text{mean}}\}$. Table 7 reports $\bar{\delta}^X_b$ and the number of scaffolds with $\delta^X_{s,b} > 0$.

**Table 7: Within-backbone differences from Direct LLM.** Entries are the mean of ten scaffold-minus-Direct differences in system point estimates; parentheses count positive differences out of ten. The columns retain their original units.

| | Novelty | Diversity | | | | Value |
|---|---|---|---|---|---|---|
| Backbone | $\Delta N_{\text{mean}}$ | $\Delta D_M$ | $\Delta D_I$ | $\Delta D_P$ | $\Delta D_E$ | $\Delta V_{\text{mean}}$ |
| Qwen2.5-72B | +0.079 (9) | +1.168 (10) | +0.523 (10) | −0.107 (3) | −0.202 (2) | +0.031 (9) |
| GLM-4-32B | +0.156 (10) | −0.067 (3) | −0.191 (4) | −0.643 (0) | −0.382 (0) | +0.040 (10) |

## C.2 Absolute Profiles by Generator Backbone

Tables 8–9 provide the values underlying Figure 4, including $N_{\max}$. Bold entries identify the largest point estimate in each column within a backbone, including ties. The leading workflow varies across dimensions: for example, under Qwen, AI-Researcher leads Novelty and Mechanism, ResearchAgent leads Prediction, and Agent Laboratory leads Value.

**Table 8:** Absolute Qwen2.5-72B-Instruct system profiles. Value uses the fixed Gemini-2.5-Flash reasoner. Bold entries mark column maxima, including ties.

| System | $N_{\text{mean}}$ | $N_{\max}$ | $D_M$ | $D_I$ | $D_P$ | $D_E$ | $V_{\text{mean}}$ |
|---|---|---|---|---|---|---|---|
| Direct LLM | 0.322 | 0.511 | 1.790 | 1.951 | 2.555 | 2.373 | +0.183 |
| Virtual Scientist | 0.388 | 0.736 | 2.590 | 2.617 | 2.609 | 1.913 | +0.222 |
| ResearchAgent | 0.335 | 0.628 | 2.991 | 2.263 | **2.887** | 2.128 | +0.211 |
| AI-Researcher | **0.528** | **0.963** | **3.716** | 2.841 | 2.774 | 2.306 | +0.239 |
| AI Scientist-v2 | 0.438 | 0.660 | 2.734 | 2.518 | 2.288 | **2.451** | +0.189 |
| InternAgent 1.5 | 0.471 | 0.928 | 3.594 | **3.154** | 2.436 | 2.324 | +0.239 |
| EvoScientist 0.3.1 | 0.393 | 0.476 | 2.286 | 1.969 | 1.929 | 1.948 | +0.203 |
| MOOSE-Chem2 | 0.153 | 0.442 | 2.778 | 2.379 | 2.482 | 2.375 | +0.225 |
| MultiAgent Ideator | 0.341 | 0.484 | 2.566 | 2.096 | 2.282 | 1.933 | +0.211 |
| Co-Scientist | 0.439 | 0.811 | 3.520 | 2.637 | 2.319 | 1.997 | +0.161 |
| Agent Laboratory | 0.524 | 0.698 | 2.809 | 2.265 | 2.473 | 2.332 | **+0.244** |

**Table 9:** Absolute GLM-4-32B-0414 system profiles. Value uses the fixed Gemini-2.5-Flash reasoner. Bold entries mark column maxima, including ties.

| System | $N_{\text{mean}}$ | $N_{\max}$ | $D_M$ | $D_I$ | $D_P$ | $D_E$ | $V_{\text{mean}}$ |
|---|---|---|---|---|---|---|---|
| Direct LLM | 0.421 | 0.696 | 3.058 | 2.491 | **2.848** | **2.353** | +0.175 |
| Virtual Scientist | 0.485 | 0.702 | 2.999 | 2.563 | 2.772 | 1.931 | +0.200 |
| ResearchAgent | 0.527 | 0.737 | 3.113 | 2.677 | 2.362 | 2.082 | +0.219 |
| AI-Researcher | 0.657 | **0.963** | 3.752 | 2.617 | 2.460 | 2.320 | +0.236 |
| AI Scientist-v2 | 0.523 | 0.887 | 2.984 | 2.290 | 2.444 | 1.914 | +0.208 |
| InternAgent 1.5 | 0.556 | 0.738 | 3.007 | 2.150 | 1.874 | 2.295 | +0.228 |
| EvoScientist 0.3.1 | **0.764** | 0.848 | 2.228 | 1.596 | 1.814 | 1.627 | +0.208 |
| MOOSE-Chem2 | 0.447 | 0.735 | 2.829 | 2.144 | 2.166 | 1.934 | **+0.244** |
| MultiAgent Ideator | 0.546 | 0.626 | 2.255 | 1.741 | 1.909 | 1.563 | +0.194 |
| Co-Scientist | 0.724 | **0.963** | **3.951** | **2.925** | 2.098 | 1.846 | +0.181 |
| Agent Laboratory | 0.537 | 0.890 | 2.791 | 2.297 | 2.149 | 2.194 | +0.236 |

### C.3 Scaffold Changes from the Direct LLM Baseline

Figure 6 shows the individual scaffold-minus-Direct contrasts summarized in Table 7.

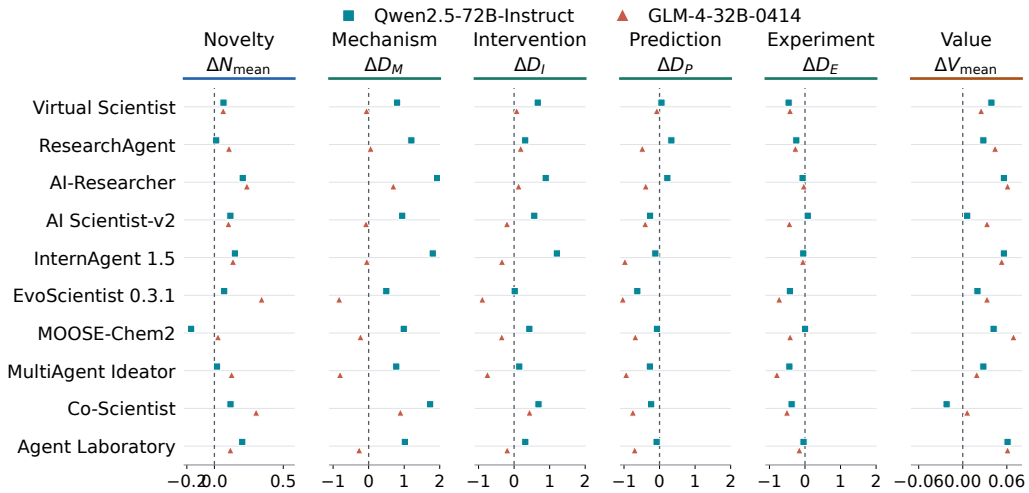


**Figure 6: Changes relative to the same-backbone Direct baseline.** Each point is a difference of system means as defined in Eq. 12; zero denotes the Direct baseline. Colors and markers match Figure 4. Novelty, the four Diversity components, and Value retain their own units.

### C.4 Directional Dispersion and Specification Coverage

For component $f$, let $m_f$ be the number of hypotheses with complete direction signatures among the $K = 10$ outputs. We report

$$\widetilde{D}_f = \frac{D_f - 1}{m_f - 1} \quad (m_f \geq 2), \qquad C_f = \frac{m_f}{K}. \tag{13}$$

Normalized dispersion is zero when the retained signatures describe one effective direction and one at the maximum $D_f = m_f$. Specification coverage measures how much of the generated set supports the component measurement. We average these quantities after computing them per topic; coverage is reported separately from dispersion.

Figure 7 shows both quantities for the two backbones. Qwen's largest mean scaffold-minus-Direct increases in normalized Diversity are in Mechanism (+0.134) and Intervention/Method (+0.105); GLM has lower mean dispersion than Direct on all four components. Raw counts and normalized dispersion can order workflows differently: under Qwen, AI Scientist-v2 has the largest $D_E$ (2.451), while ResearchAgent has the largest $\widetilde{D}_E$ (0.483). Tables 10–11 provide the full-precision summaries and $V_{\max}$.

**(a) Normalized Diversity $\widetilde{D}_f$**

| | Qwen2.5-72B-Instruct | | | | GLM-4-32B-0414 | | | |
|---|---|---|---|---|---|---|---|---|
| | M | I | P | E | M | I | P | E |
| **Direct LLM** | **0.09** | **0.14** | **0.19** | **0.35** | **0.24** | **0.31** | **0.25** | **0.41** |
| Virtual Scientist | 0.19 | 0.28 | 0.21 | 0.29 | 0.23 | 0.33 | 0.25 | 0.35 |
| ResearchAgent | 0.22 | 0.20 | 0.26 | 0.48 | 0.25 | 0.32 | 0.21 | 0.40 |
| AI-Researcher | 0.30 | 0.36 | 0.23 | 0.45 | 0.32 | 0.35 | 0.18 | 0.53 |
| AI Scientist-v2 | 0.19 | 0.21 | 0.20 | 0.48 | 0.23 | 0.24 | 0.21 | 0.31 |
| InternAgent 1.5 | 0.29 | 0.38 | 0.20 | 0.35 | 0.23 | 0.22 | 0.16 | 0.30 |
| EvoScientist 0.3.1 | 0.15 | 0.18 | 0.13 | 0.20 | 0.15 | 0.08 | 0.11 | 0.28 |
| MOOSE-Chem2 | 0.20 | 0.20 | 0.19 | 0.39 | 0.22 | 0.22 | 0.15 | 0.29 |
| MultiAgent Ideator | 0.17 | 0.20 | 0.17 | 0.36 | 0.14 | 0.17 | 0.13 | 0.21 |
| Co-Scientist | 0.28 | 0.27 | 0.17 | 0.38 | 0.36 | 0.42 | 0.19 | 0.35 |
| Agent Laboratory | 0.22 | 0.21 | 0.22 | 0.27 | 0.21 | 0.25 | 0.16 | 0.44 |

**(b) Specification coverage $C_f$**

| | M | I | P | E | M | I | P | E |
|---|---|---|---|---|---|---|---|---|
| **Direct LLM** | **1.00** | **0.72** | **0.94** | **0.50** | **0.97** | **0.66** | **0.90** | **0.39** |
| Virtual Scientist | 0.98 | 0.74 | 0.94 | 0.37 | 0.97 | 0.68 | 0.90 | 0.35 |
| ResearchAgent | 1.00 | 0.70 | 0.92 | 0.41 | 0.96 | 0.72 | 0.85 | 0.37 |
| AI-Researcher | 1.00 | 0.70 | 0.89 | 0.37 | 0.95 | 0.62 | 0.90 | 0.32 |
| AI Scientist-v2 | 0.99 | 0.72 | 0.88 | 0.33 | 0.97 | 0.65 | 0.87 | 0.34 |
| InternAgent 1.5 | 0.99 | 0.72 | 0.90 | 0.37 | 0.99 | 0.62 | 0.80 | 0.33 |
| EvoScientist 0.3.1 | 0.99 | 0.69 | 0.90 | 0.39 | 0.94 | 0.69 | 0.90 | 0.35 |
| MOOSE-Chem2 | 0.99 | 0.70 | 0.94 | 0.36 | 0.97 | 0.70 | 0.88 | 0.31 |
| MultiAgent Ideator | 1.00 | 0.65 | 0.92 | 0.31 | 0.98 | 0.70 | 0.90 | 0.37 |
| Co-Scientist | 0.99 | 0.68 | 0.90 | 0.27 | 0.93 | 0.66 | 0.80 | 0.21 |
| Agent Laboratory | 0.93 | 0.66 | 0.84 | 0.35 | 0.95 | 0.67 | 0.85 | 0.35 |




**Figure 7: Diversity dispersion and specification coverage.** Rows retain the system order of Figure 4. The upper panels show topic-averaged normalized Diversity $\widetilde{D}_f$; the lower panels show specification coverage $C_f$. M/I/P/E denote Mechanism, Intervention/Method, Prediction, and Experiment. Cells use a common $[0, 1]$ color scale and display two decimal places; Tables 10–11 retain three.

**Table 10:** Qwen2.5-72B-Instruct normalized-Diversity, specification-coverage, and $V_{\max}$ diagnostics.

| System | $\widetilde{D}_M$ | $\widetilde{D}_I$ | $\widetilde{D}_P$ | $\widetilde{D}_E$ | $C_M$ | $C_I$ | $C_P$ | $C_E$ | $V_{\max}$ |
|---|---|---|---|---|---|---|---|---|---|
| Direct LLM | 0.088 | 0.142 | 0.193 | 0.347 | 0.996 | 0.715 | 0.935 | 0.504 | +0.278 |
| Virtual Scientist | 0.187 | 0.275 | 0.208 | 0.293 | 0.977 | 0.735 | 0.935 | 0.365 | +0.361 |
| ResearchAgent | 0.222 | 0.201 | 0.258 | 0.483 | 0.996 | 0.700 | 0.923 | 0.408 | +0.306 |
| AI-Researcher | 0.304 | 0.359 | 0.231 | 0.448 | 0.996 | 0.704 | 0.892 | 0.369 | +0.417 |
| AI Scientist-v2 | 0.194 | 0.210 | 0.202 | 0.479 | 0.992 | 0.719 | 0.881 | 0.327 | +0.278 |
| InternAgent 1.5 | 0.294 | 0.379 | 0.197 | 0.346 | 0.985 | 0.719 | 0.900 | 0.369 | +0.361 |
| EvoScientist 0.3.1 | 0.145 | 0.176 | 0.130 | 0.202 | 0.992 | 0.688 | 0.904 | 0.388 | +0.278 |
| MOOSE-Chem2 | 0.199 | 0.202 | 0.188 | 0.388 | 0.992 | 0.696 | 0.935 | 0.358 | +0.306 |
| MultiAgent Ideator | 0.174 | 0.197 | 0.165 | 0.364 | 1.000 | 0.654 | 0.919 | 0.308 | +0.278 |
| Co-Scientist | 0.284 | 0.265 | 0.172 | 0.379 | 0.992 | 0.681 | 0.904 | 0.265 | +0.250 |
| Agent Laboratory | 0.220 | 0.205 | 0.224 | 0.266 | 0.931 | 0.658 | 0.835 | 0.346 | +0.389 |

**Table 11:** GLM-4-32B-0414 normalized-Diversity, specification-coverage, and $V_{\max}$ diagnostics.

| System | $\widetilde{D}_M$ | $\widetilde{D}_I$ | $\widetilde{D}_P$ | $\widetilde{D}_E$ | $C_M$ | $C_I$ | $C_P$ | $C_E$ | $V_{\max}$ |
|---|---|---|---|---|---|---|---|---|---|
| Direct LLM | 0.242 | 0.310 | 0.253 | 0.408 | 0.965 | 0.658 | 0.904 | 0.392 | +0.278 |
| Virtual Scientist | 0.233 | 0.329 | 0.251 | 0.351 | 0.973 | 0.681 | 0.900 | 0.354 | +0.278 |
| ResearchAgent | 0.248 | 0.322 | 0.214 | 0.396 | 0.962 | 0.723 | 0.854 | 0.365 | +0.306 |
| AI-Researcher | 0.321 | 0.354 | 0.177 | 0.533 | 0.954 | 0.619 | 0.896 | 0.323 | +0.278 |
| AI Scientist-v2 | 0.229 | 0.242 | 0.209 | 0.310 | 0.969 | 0.646 | 0.873 | 0.335 | +0.278 |
| InternAgent 1.5 | 0.229 | 0.217 | 0.157 | 0.299 | 0.985 | 0.623 | 0.796 | 0.327 | +0.278 |
| EvoScientist 0.3.1 | 0.150 | 0.080 | 0.106 | 0.279 | 0.938 | 0.688 | 0.900 | 0.350 | +0.306 |
| MOOSE-Chem2 | 0.216 | 0.218 | 0.147 | 0.288 | 0.969 | 0.696 | 0.877 | 0.312 | +0.333 |
| MultiAgent Ideator | 0.141 | 0.166 | 0.125 | 0.208 | 0.981 | 0.700 | 0.904 | 0.369 | +0.306 |
| Co-Scientist | 0.355 | 0.421 | 0.188 | 0.350 | 0.931 | 0.662 | 0.804 | 0.212 | +0.333 |
| Agent Laboratory | 0.212 | 0.248 | 0.163 | 0.436 | 0.954 | 0.669 | 0.850 | 0.354 | +0.333 |

### C.5 Workflow Adaptation and Scoring Protocol

**Common task interface.** Every workflow receives the benchmark question and visible observations $(Q, D^{\text{vis}})$. Generation, critique, and auxiliary model calls use the same paired backbone. The evaluation covers hypothesis generation before experimental execution. Retrieval is restricted to the visible corpus; hidden evidence and question answers remain evaluator-only. We use the authors' implementations where available and a reproduction for Co-Scientist (Gottweis et al., 2026). Source revisions are documented with the released code. Single-idea workflows are repeated to obtain ten outputs; multi-idea workflows use their native selection.

**Direct LLM and reference construction.** Direct LLM makes ten independent one-hypothesis calls per topic and backbone, without review or revision. Its hypothesis prompt matches reference generation, but evaluated Direct outputs and reference samples are disjoint. The five-family pool supplies a common Novelty reference. Novelty uses the full codebook with $n = 150$ reference hypotheses.

**Diversity and Value.** Diversity applies the open-set component parser, complete-signature rule, and positive-semidefinite kernel of Section 4.2 to the full hypothesis text, including Direct LLM outputs. Per-topic base vocabularies are built from $Q + D^{\text{vis}}$ before scoring. Explicit open-set values are canonicalized within topic and slot; unstated required attributes reduce specification coverage. Value uses the fixed Gemini-2.5-Flash reasoner, no-hypothesis baseline, headroom screen, and hidden question packs. Appendix D gives the conditioning and screening protocol, also validated with GPT-4.1 in E3.

**Search and revision procedures.** We preserve each workflow's native search, critique, revision, and selection structure while adapting inputs to the benchmark setting. MOOSE-Chem2 retains its hierarchical beam search, recombination, and stopping procedure; MultiAgent Research Ideator retains proposal–critique–revision dialogues; Virtual Scientist retains multi-scientist discussion and novelty checking; ResearchAgent retains iterative problem, method, and experiment development; AI-Researcher and AI Scientist-v2 retain their native candidate selection, reflection, and finalization procedures; InternAgent 1.5 retains idea generation, reflection, evolution, ranking, and refinement; and EvoScientist 0.3.1 retains role delegation, planning, and memory.

Task-specific adaptations restrict retrieval and auxiliary agents to $(Q, D^{\text{vis}})$, isolate memory across topics and backbones, and omit components that require unavailable experimental outcomes or external task-specific corpora. Further implementation details are provided with the released code.

**Generation budgets and response handling.** We preserve each workflow's native search and call structure rather than equalizing computation. Responses allow up to 8,192 output tokens unless the native workflow uses a lower cap; ResearchAgent is raised from 1,000 to 8,192 to avoid truncation. Runs use a 90-minute budget, extended to six hours for ResearchAgent and InternAgent, with at most two run attempts and three format retries. Gemini Value requests retry truncation at 8,192, 16,384, and 32,768 tokens. Incomplete runs and ambiguous responses are left unscored.

Minor compatibility fixes preserve workflow semantics, candidate counts, and selection procedures; implementation details are documented with the released code.

# D VALUE PROTOCOL AND REASONER COHORTS

**Question packs and treatments.** Each of the 61 topic packs contains three hidden findings from papers dated on or after 2025-05-01 and 12 conceptual multiple-choice questions, split evenly between recall and transfer. The packs comprise 26 AI/ML, 11 life science, 18 physics/materials topics (including five chemistry), and six earth/environment topics. E3 uses the 51 held-out packs, with a separate headroom screen for each reasoner. The construction model, `deepseek-chat`, drafts Supported (A), Partial (B), Irrelevant (C), and Misleading (D) hypotheses from $Q + D^{\text{hid}}$; $D^{\text{hid}}$ is unavailable to the evaluated systems. Appendix F describes the annotation protocol.

**Reasoner protocol.** GPT-4.1 and Gemini-2.5-Flash evaluate the same held-out packs and A–D treatments at temperature zero. Each reasoner uses its own closed-book answers, no-hypothesis baseline, and headroom screen ($S^{-h} \leq 0.85$). Neither uses search or external tools. Responses initially allow 8,192 output tokens; truncation triggers retries at 16,384 and 32,768 tokens with the prompt unchanged. Abstentions are scored incorrect and unresolved responses as missing. The system study uses the same question packs, Gemini baseline screen, and response-validity rules.

For paired Value,

$$\hat{V} = (n_{01} - n_{10})/m,$$

where $n_{01}$ and $n_{10}$ count questions correct only with and only without the hypothesis, respectively. Exact McNemar tests use these discordant counts (McNemar, 1947). All reported results use raw $\hat{V}$.

**Table 12:** Mean Value and 95% topic-bootstrap intervals for the A–D treatments. Each reasoner uses its own baseline-screened cohort.

| Reasoner | A | B | C | D |
|---|---|---|---|---|
| GPT-4.1 | +0.175 | +0.052 | −0.115 | −0.440 |
| 95% CI | [0.14, 0.21] | [0.02, 0.08] | [−0.20, −0.04] | [−0.51, −0.37] |
| Gemini-2.5-Flash | +0.218 | +0.030 | −0.111 | −0.381 |
| 95% CI | [0.17, 0.26] | [−0.01, 0.07] | [−0.17, −0.05] | [−0.43, −0.33] |

Both reasoners yield the mean ordering A>B>C>D. Adjacent one-sided Wilcoxon tests with Holm correction are significant for all three GPT-4.1 contrasts ($p = 0.000256$, 0.002, and 0.000142) and all three Gemini contrasts ($p = 3.2 \times 10^{-7}$, 0.000248, and $6.18 \times 10^{-7}$). Gemini's confidence interval for Partial includes zero. GPT-4.1 produces four explicit abstentions across the treatment responses, while Gemini produces none.

**Common-cohort sensitivity.** The two baseline screens share 17 retained topics. On this common cohort, GPT-4.1 gives mean A/B/C/D values of $+0.191/+0.059/-0.103/-0.431$, while Gemini-2.5-Flash gives $+0.260/+0.034/-0.152/-0.358$. The A>B>C>D ordering is preserved for both reasoners.

**Scientific-content ablation.** Using `qwen-plus-2025-04-28`, we perturb the Supported hypothesis by removing its prediction, removing its mechanism, shuffling its words, or replacing it with length-matched meaningless text. Each paired contrast uses the packs for which both the intact and perturbed conditions are available. Table 13 reports the absolute scores; all paired changes relative to the intact hypothesis are negative.

**Table 13:** Absolute Value under scientific-content ablations.

| Condition | Mean $V$ | 95% CI |
|---|---|---|
| Full Supported | +0.323 | [+0.284, +0.363] |
| No prediction | +0.222 | [+0.177, +0.266] |
| No mechanism | +0.175 | [+0.140, +0.211] |
| Shuffled words | −0.061 | [−0.107, −0.016] |
| Length-matched nonsense | −0.232 | [−0.275, −0.191] |

### D.1 VALUE VERSUS DIRECT JUDGMENTS WITH MATCHED MODELS

We compare Value with direct significance judgments using matched models, questions, visible evidence, and authored hypotheses. Direct judges receive $Q + D^{\text{vis}} + h$, without treatment labels, hidden findings, or answer keys, at temperature zero. The comparison uses the cohorts fixed by the corresponding Value baseline screen, and Value and direct judging are evaluated on identical hypothesis items. Intervals use 20,000 paired topic-bootstrap resamples (seed 20260921), with ties counted as unsuccessful.

**Table 14:** Adjacent-pair accuracy for Value and matched direct judgments. Entries show accuracy with 95% topic-bootstrap intervals.

| Reasoner | Method | A>B | B>C | C>D |
|---|---|---|---|---|
| GPT-4.1 | Value | 0.85 [0.70, 1.00] | 0.75 [0.55, 0.95] | 0.90 [0.75, 1.00] |
| | Direct judge | 0.60 [0.40, 0.80] | 0.60 [0.40, 0.80] | 0.55 [0.35, 0.75] |
| Gemini-2.5-Flash | Value | 0.81 [0.68, 0.92] | 0.62 [0.46, 0.78] | 0.84 [0.70, 0.95] |
| | Direct judge | 0.43 [0.27, 0.59] | 0.32 [0.19, 0.49] | 0.38 [0.22, 0.54] |

# E PROMPTS AND EVALUATION SETTINGS

This section records the prompt strings and parser settings used in the calibration, Diversity, and Value studies. The Novelty mode matcher uses the official `deepseek-v4.1-flash` model at temperature zero. The Diversity slot parser also runs at temperature zero; reference generation uses the single-shot template. AI-scientist prompts and task adaptations are described in Appendix C.

### REFERENCE GENERATOR

```
SYSTEM = "You are a research scientist proposing hypotheses. Respond with a single JSON
     object and nothing else."

SINGLE_PROMPT:
{context}

Propose ONE scientific hypothesis that answers the research question.

Return JSON with exactly these keys:
- "mechanism": one sentence stating the core causal explanation (what actually causes the
     phenomenon).
- "hypothesis": two or three sentences elaborating the mechanism.
- "prediction": one sentence stating an observation that should hold if the hypothesis is
     true.
```

### MODE MATCHER: PASS-2 ASSIGNMENT

```
MATCH_SYSTEM = "You are an expert annotator who merges scientific hypotheses into equivalence
      classes called modes. Respond with a single JSON object and nothing else."

MATCH_PROMPT:
Research question: {question}

Two hypotheses belong to the SAME mode if they propose the same core causal mechanism for the
      phenomenon, regardless of wording, level of detail, terminology, proposed remedies, or
     predictions. They belong to DIFFERENT modes if the causal mechanism differs: a different
      cause, a different locus (e.g. training data vs. architecture vs. decoding), or a
     different direction of effect. A hypothesis that combines two mechanisms belongs to the
     mode of its primary mechanism, unless the interaction itself is the causal claim.
{contrasts}
Existing modes: {registry}
Hypotheses to assign: {batch}

For each hypothesis, either assign an existing mode id (e.g. "M12") or a new mode key ("N1",
     "N2", ...). Hypotheses in this batch that share the same new mechanism must share the
     same new key. For every new key, write one canonical sentence describing the mechanism,
     general enough to cover paraphrases of it.

Return JSON: {"assignments": {"H1": "M3", ...}, "new_modes": {"N1": "canonical sentence",
     ...}}

MERGE_PROMPT (codebook pass): identify groups of modes that describe the SAME mechanism (same
      cause, same locus, same direction of effect). Do not merge modes that differ in cause
     or locus even if they often co-occur. Return {"merges": [["M3","M7"], ...]}.
```

**Diversity slot parser used in E2 and system evaluation.** E2 validates the same open-set component parser and effective-direction construction used for the main benchmark. The parser uses `deepseek-chat` ($T = 0$, 400-token limit) with a separate prompt for each component. Each prompt provides the scientific question, one component claim, and the topic-local base slot values. For every slot, the parser returns one of three states: (i) an exact base-vocabulary label; (ii) an explicit open-set value when the claim states a scientifically meaningful value absent from the base vocabulary; or (iii) `UNSTATED` when the claim does not specify the attribute. Open-set values then receive category ids using the canonicalization procedure below. `UNSTATED` is mapped to $\perp$ and is never treated as a scientific value.

For each controlled E2 topic, the base vocabulary is constructed from that topic's visible benchmark inputs using the same vocabulary-construction procedure as in the system evaluation and is frozen before any recovery-set claims are scored. A claim contributes to a component axis only when all slots in that component's direction signature are specified; incomplete claims reduce specification coverage rather than pairwise disagreement. The retained categorical records are compared with

the positive-semidefinite kernel in §4.2. Neither the author-assigned direction labels nor the author annotations used to define the recovery target are provided to vocabulary construction, component extraction, or open-set canonicalization. They are used only after eligibility is fixed to compute the gold effective count on the same retained items and the resulting error. The system evaluation applies the same parser, canonicalization, eligibility rule, and kernel to generated hypotheses.

**Open-set slot canonicalization.** Canonicalization runs once per topic and slot, after parsing and before the kernel. It does not use Novelty mode ids. Base-vocabulary labels are frozen anchors and are not merged with one another. `UNSTATED` values do not enter. An open-set string may join one anchor or form a new category with other open-set strings.

Strings are first normalized with Unicode compatibility decomposition, case folding, and collapsed whitespace. Exact matches after normalization collapse immediately. An open-set string that matches an anchor joins that anchor. Each remaining string is sent, together with the anchors, to one `deepseek-chat` call at temperature zero. Inputs are sorted lexicographically. The call assigns every remaining string to exactly one anchor or to a new category id. Strings that share a new id denote the same value of that slot. Two anchors are never placed in one category.

Equality is defined by the slot. Cause matches the same causal factor, locus the same site of action, and effect sign the same qualitative direction. Method family matches the same family of methods, and target the same entity. The prediction observable matches the same measured quantity, and its direction the same qualitative change. Experiment design matches the same study design, and comparison the same contrast. Abbreviations, aliases, and paraphrases merge. Values that differ in the slot's defining attribute stay apart, including factors that merely co-occur and narrower values that change the category. Information that belongs to another slot is ignored.

Invalid JSON, a missing string, or a merge of two anchors is retried with the same prompt. If the call is still invalid, each remaining string keeps its own category and the failure is logged. A string attached to an anchor uses that anchor as its category id. A new category uses the lexicographically smallest raw string in its group. The kernel compares these ids and nothing else. E2 and the system evaluation call this procedure. Each topic and slot stores the input strings, the model output, and the final ids. Two annotators with Ph.D.s in computer science reviewed the resulting decisions on 200 AI items; agreement is 0.86 (Appendix F).

For spectral entropy we use $0 \log 0 = 0$ and omit eigenvalue terms at or below $10^{-12}$ for numerical stability. When a single four-axis aggregate is needed, we use

$$\widetilde{D} = \frac{1}{4} \sum_{f \in \mathcal{F}} \widetilde{D}_f,$$

provided that all four axes are available; the primary report remains the component-wise triple $(D_f, \widetilde{D}_f, C_f)$.

**E2 direction-judge coverage.** Cluster-$e^H$ scores are available for 5,116 sets for DeepSeek, 5,087 for GLM, 5,092 for Qwen and 5,120 for Intern, out of 5,120 planned sets per model. A returned direction count without a usable ten-item cluster assignment does not supply a cluster-entropy score. Table 15 reports these available-case metrics; their denominators differ slightly. GPT-4.1 and Gemini-2.5-Flash cover all 5,120 sets. Table 2 selects GPT-4.1 and Gemini-2.5-Flash for consistency with the Value comparison.

**GPT-4.1 and Gemini-2.5-Flash direction judges.** GPT-4.1 and Gemini-2.5-Flash use the same direction-counting prompt and the same 5,120 seeded recovery sets as the other LLM judges (seed 0, 32 repetitions, $K = 10$). Each request contains only the ten component claims; gold direction labels, slot annotations and planned cluster sizes are withheld. The judge returns a cluster assignment for each claim. We compute cluster $e^H$ and compare it with the same $G_{\text{eff}}$, reporting MAE and $\text{Acc}_{\pm 1}$ over 1,280 sets per axis. Temperature remains zero for all attempts. Output caps start at 8192 tokens and increase to 16384 and 32768 only after truncation; output budgets are model dependent across the full baseline suite. Both judges use the provider's default reasoning settings. Incomplete or malformed cluster assignments are missing. These two judges' results require valid assignments for all 5,120 sets.

**Table 15:** Full E2 direction-judge comparison, with Ours as a reference. $n$ gives valid recovery sets out of 5120 planned sets. Entries are MAE and $\text{Acc}_{\pm 1}$ against $G_{\text{eff}}$ on each axis. DeepSeek, GLM, Qwen and Intern use available valid assignments; GPT-4.1 and Gemini-2.5-Flash cover the full cohort. Ours uses the open-set structured component representation and complete-signature kernel used by the primary Diversity pipeline, while direct judges receive only the component claims.

| | | MAE↓ | | | | $\text{Acc}_{\pm 1}$ ↑ | | | |
|---|---|---|---|---|---|---|---|---|---|
| Method | $n$ | $M$ | $I$ | $P$ | $E$ | $M$ | $I$ | $P$ | $E$ |
| DeepSeek | 5116 | 0.82 | 3.00 | 1.01 | 2.62 | 0.74 | 0.15 | 0.72 | 0.27 |
| GLM | 5087 | 0.93 | 2.29 | 0.91 | 1.79 | 0.66 | 0.25 | 0.73 | 0.45 |
| Qwen | 5092 | 0.85 | 2.39 | 1.20 | 2.45 | 0.69 | 0.15 | 0.62 | 0.20 |
| Intern | 5120 | 1.34 | 3.04 | 1.05 | 2.91 | 0.49 | 0.12 | 0.66 | 0.24 |
| GPT-4.1 | 5120 | 1.21 | 3.42 | 1.56 | 3.22 | 0.56 | 0.05 | 0.47 | 0.15 |
| Gemini-2.5-Flash | 5120 | 1.74 | 3.46 | 2.25 | 3.30 | 0.49 | 0.11 | 0.38 | 0.21 |
| **Ours** | 5120 | **0.49** | **0.55** | **0.32** | **0.73** | **0.82** | **0.83** | **0.92** | **0.71** |

**Table 16:** E2 component-availability diagnostic across the 320 source claims. An input claim is counted as unavailable on an axis here when no valid slot is extracted for that component. These rates describe parser availability before construction of the controlled recovery sets; they are distinct from the complete-signature coverage used by the final Diversity score. Unavailable parser responses are reported separately.

| Domain | Claims | Unavailable | $M$ | $I$ | $P$ | $E$ |
|---|---|---|---|---|---|---|
| All | 320 | 0 | 0.022 | 0.016 | 0.025 | 0.034 |
| AI/ML | 80 | 0 | 0.025 | 0.025 | 0.075 | 0.100 |
| Earth/environment | 120 | 0 | 0.033 | 0.000 | 0.000 | 0.017 |
| Physics/materials | 80 | 0 | 0.013 | 0.025 | 0.025 | 0.013 |
| Life sciences | 40 | 0 | 0.000 | 0.025 | 0.000 | 0.000 |

Under the final complete-signature rule, all controlled recovery sets remain scorable on every axis: $M$: 1280/1280; $I$: 1280/1280; $P$: 1280/1280; $E$: 1280/1280. Component-level specification coverage is determined by $m_f/K$ as defined in §4.2.

### CONDITIONED VALUE REASONER

```
SYSTEM = "You are a careful scientific reasoner. Respond with a single JSON object and
    nothing else."

PROMPT:
Research question: {question}
Known observations: {visible}
{hypothesis_block}
Answer the following question.
{qtext}
{answer_spec}
Reason briefly and internally, then return only JSON: {"answer": "..."}

HYPOTHESIS_BLOCK (with-h conditions; omitted on the baseline):
Working hypothesis -- Treat the following hypothesis as a candidate explanation.
Use it when it determines an answer choice. When it does not determine the
answer, answer from the known observations. Do not change an answer only to
make use of the hypothesis.
{hypothesis}
```

# F INDEPENDENT ANNOTATION

Independent annotation checks Novelty mechanism equivalence, Diversity open-set canonicalization, and the interpretation of the authored Value treatments. These audits are separate from the benchmark scoring procedures in Sections E and D.

**Novelty mechanism equivalence.** Two annotators with Ph.D.s in computer science reviewed 100 pairs of Novelty mode assignments on AI topics. Agreement with the matcher is $0.82$.

**Question validity and Value treatments.** One annotator with a Ph.D. in computer science reviewed the hidden-evidence questions and the authored A/B/C/D treatments. The annotator had access to $D^{\text{hid}}$ and the cited findings and did not see model scores. The review found the questions to concern those findings and the four treatments to match the E3 definitions: Supported entails the keyed answers, Partial supplies an incomplete account, Irrelevant does not determine them, and Misleading contradicts them.

**Diversity slot canonicalization.** Two annotators with Ph.D.s in computer science reviewed the open-set canonicalization in Appendix E on 200 AI items. Agreement is $0.86$. They saw the slot strings and the assigned category ids.

# G VALIDATION PROTOCOLS

This appendix gives the controlled constructions underlying E1 and E2.

## G.1 EXPLORATION NOVELTY: REFERENCE-COVERAGE CALIBRATION

*Synthetic.* 10 Zipf mode populations (exponents 0.7–1.6), 5 seeds; $n_{\text{build}} \in \{80, 160, 320\}$, held-out 160, big sample 1600. With a clean matcher, in-pool MAE for $r{=}0$ is 0.037 at $n{=}320$. Injecting 10% splits or 5% merges produces similar calibration when the same noise process affects both build and held-out samples. *Real E1.* In-pool on 4 AI + 12 science topics plus one chemistry aggregate (five mechanism cards). The build sample is the five-family scoring reference, $n_{\text{build}}{=}150$ (5 families $\times 30$), predicting the unseen share of a large independent sample from the same reference ($n_{\text{big}}{=}600$ AI, $\approx 300$ science); unseen mass $r{=}0$; learning-curve subsamples $n \in \{50, 100, 150\}$ (30 repeats; empirical share measured on the fixed large sample); and leave-one-family-out estimates of $\hat{P}_{\text{unseen}}$. Matcher annotation is reported in Appendix F.

## G.2 DIVERSITY: CONTROLLED DIRECTION RECOVERY

320 primary papers on eight formal topics (lost in the middle and in-context learning; coral heat adaptation and ICI primary resistance; cuprate pseudogap and magic-angle graphene; AMOC threshold and cloud-feedback sign), 40 per topic, with author-assigned $M/I/P/E$ labels; planned size distributions (ground truth $G^{\text{eff}}$); $K{=}10$, 5120 sets, without replacement ($\geq 10$ papers per direction). Gold is the author direction id on a short claim restatement. For the Prediction axis, the restatement names one measured variable without exposing the gold slot surface form.

For each topic, we construct the topic-local base vocabulary from the visible benchmark inputs using the same procedure as in the system evaluation and freeze it before any recovery-set claims are scored. We then use `deepseek-chat` for open-set component extraction. Explicit values outside the base vocabulary are retained and assigned category ids by Appendix E; unstated values are marked $\perp$. On each component axis, only claims with complete direction signatures enter the positive-semidefinite categorical kernel. Gold direction labels are withheld during vocabulary construction, extraction, and open-set canonicalization. After eligibility is fixed, they are used to compute $G^{\text{eff}}$ and the recovery error on the retained items.

The score uses the component kernels and effective direction counts in §4.2, compared with topic-conditional TF-IDF cluster $e^H$ ($\tau{=}0.06$, best bag-of-words cut), cosine-threshold cluster $e^H$ on MiniLM, mean bidirectional $P(\text{entail}) \geq 0.02$ NLI cluster $e^H$ (DeBERTa-v3-small; Kuhn both-argmax over-fragments), and LLM direction judges with the same prompt (official `deepseek-chat`, `glm-5.3-flash`, `Qwen3.6-35B-A3B-FP8`, `intern-s1`, GPT-4.1 and Gemini-2.5-Flash; coverage is given above).

# H REFERENCE-COVERAGE SENSITIVITY

This appendix reports reference-pool calibration and reference-pool-size sensitivity analyses.

**E1 learning curve and leave-one-family-out.** Figure 3 has four macro-domains. The build sample is the five-family scoring reference, 30 hypotheses per family ($n_{\text{build}}$=150). The AI panel is four topics ($n_{\text{big}}$=600): mean pred. 0.013 vs. emp. 0.014, MAE 0.008. The science panels are 12 science topics plus one chemistry aggregate ($n_{\text{big}}\approx$300): MAE 0.017 (mean pred. 0.034 vs. emp. 0.043). Panel (e) pools 17 units: unseen-mass MAE 0.087/0.042/0.013 at $n$=50/100/150 ($n$=150 is 15 units with a full build). Near-duplicate modes on two science topics were merged with the corresponding reference modes before the large-sample comparison. These analyses calibrate the finite scoring reference against larger independent samples from the same five families. Leaving out one family changes $\hat{P}_{\text{unseen}}$ by a median 0.003 on the AI topics (range −0.017–0.023) and by a median 0.013 on the science topics.

## H.1 REFERENCE-POOL SIZE SENSITIVITY ON SYSTEM OUTPUTS

We test whether reference-pool size changes Exploration Novelty comparisons on actual AI-scientist outputs. The analysis uses the 25 held-out topics for which all 11 GPT-4.1 systems have ten successfully matched hypotheses. Generated hypotheses and the full-pool mode assignments are fixed throughout.

For each topic, we draw nested, family-balanced reference subsets of size $n \in \{30, 60, 90, 120\}$ from the full $n = 150$ pool. Each size is evaluated over 30 repeats. Within a topic and repeat, all 11 systems share the same reference subset. We recompute the Good–Turing coverage term and reference-mode counts, rescore the fixed hypotheses with

$$N(h) = \left(1 - \hat{P}_{\text{unseen}}\right) \frac{1}{r(h) + 1},$$

and average $N_{\text{mean}}$ equally across the 25 topics.

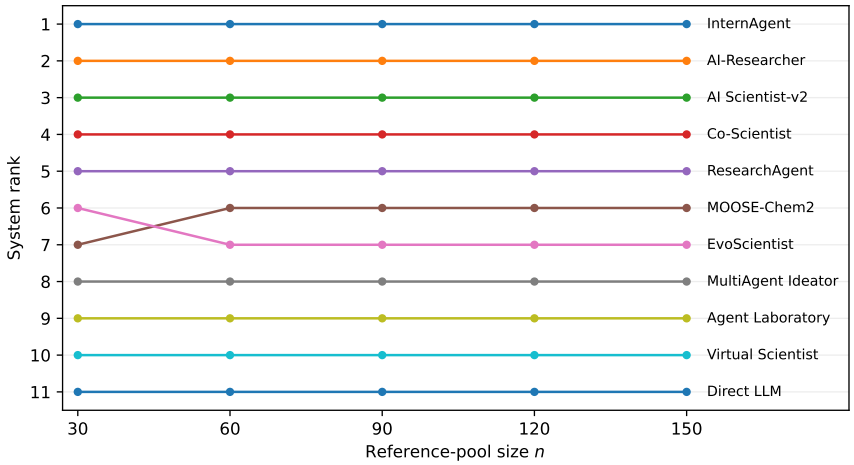


**Figure 8:** Exploration Novelty ranking stability across reference-pool sizes. Each line shows one GPT-4.1 system on the 25-topic complete-case subset. For $n < 150$, ranks are computed from mean scores over 30 nested, family-balanced subsamples; $n = 150$ uses the full reference pool. Full-pool mode assignments are held fixed.

Figure 8 shows that the mean system ranking changes little with reference-pool size: EvoScientist and MOOSE-Chem2 exchange positions between $n = 30$ and $n \geq 60$, while the remaining ordering is stable. Across the 30 repeats, mean Spearman correlation with the full $n = 150$ ranking is 0.963, 0.985, 0.987, and 0.994 for $n = 30, 60, 90, 120$, respectively. The minimum correlation is 0.909 at $n = 30$ and at least 0.936 for $n \geq 60$.

# I ETHICS AND SCOPE

This appendix records the evaluation boundary, annotation provenance, and dual-use considerations. The benchmark evaluates hypotheses and proposed experiments without executing them. Topic evidence is drawn from published scientific literature. Visible inputs and hidden question keys are separated in the evaluation interface. Proposed experiments may still contain actionable domain content; any later execution requires its own scientific and ethical review. One annotator with a Ph.D. in computer science reviewed Value question validity and treatment interpretation (Appendix F). Two annotators with Ph.D.s in computer science reviewed Diversity slot canonicalization on 200 AI items; agreement is 0.86 (Appendix F). Two annotators with Ph.D.s in computer science reviewed Novelty mode assignments on 100 pairs from AI topics; agreement with the matcher is 0.82 (Appendix F).