LEARNING INDIVIDUAL SCIENTIFIC TASTE WITH PRIVILEGED HINDSIGHT DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Scientific taste shapes which research directions researchers pursue, and it varies across individuals. Existing work either evaluates directions along shared criteria or learns aggregate scientific judgment from community signals such as citations and reviews; neither captures how individuals translate the same assessments into different preferences. We model scientific taste as a researcher-specific mapping from criterion-level assessments in a shared, evidence-grounded evaluation space, together with proposal-visible evidence, to relative preferences. A population-level mapping learned from retrospective supervision provides a cold start when individual feedback is scarce. We introduce Privileged Hindsight Distillation (PHD), which uses later-emerging scientific evidence during training to improve proposal-time criterion assessment and the population mapping, while exposing only proposal-visible information at inference. We evaluate our approach primarily in quantum research, where structured public reviews are largely unavailable, and use pairwise proposal comparison as a direct interface for observing individual taste. With only 15 comparisons per researcher, our method captures sharply different tastes, including opposing preferences over the same research-direction pairs when they trade off across evaluation criteria. In an interactive study, personalized rankings also surface proposals associated with greater subsequent user engagement, connecting learned taste to downstream research exploration.

1 INTRODUCTION

Scientific taste broadly refers to the judgment that shapes what researchers choose to work on, how they approach scientific problems, and how they recognize the potential impact of research ideas and findings (Mitchison, 2013; Tjian, 2023; Tong et al., 2026). As large language models (LLMs) increasingly support scientific research from hypothesis and proposal generation to experimental design, execution, and paper drafting (Lu et al., 2024; Yamada et al., 2025; Schmidgall et al., 2025; Gottweis et al., 2026), generating research directions becomes more scalable, making it increasingly important to decide which are worth pursuing before their outcomes are known.

Scientific taste, however, varies across researchers. Researchers differ in which problems they consider important and which approaches they prefer (Mitchison, 2013), and substantial disagreement persists even when domain experts evaluate the same scientific work (Bornmann et al., 2010; Mutz et al., 2012). Learning individual taste is particularly challenging when direct feedback is scarce. Recent work evaluates research directions along criteria such as novelty, significance, and feasibility (Qiao et al., 2026; Pei et al., 2026), but criterion-level assessments alone do not specify how a particular researcher translates them into preferences. Other work learns scientific judgment at scale from ecosystem signals such as citations, public reviews, and publication outcomes (Koltun & Hafner, 2021; Tong et al., 2026; Gong et al., 2026; Rasool et al., 2026). When aggregated, these signals provide population-level supervision rather than the taste of any particular individual.

To model this variation, we separate shared evaluation from individual preference. At the proposal stage, each research direction is represented in a shared evaluation space over criteria such as novelty and generality, with assessments grounded in proposal claims and relevant literature (Zhang et al., 2026a). This common representation supports comparison across directions, while individual taste is modeled as a researcher-specific mapping from criterion-level assessments and proposal-visible evidence to relative preferences (Figure 1). The same shared evaluation can therefore support different

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.

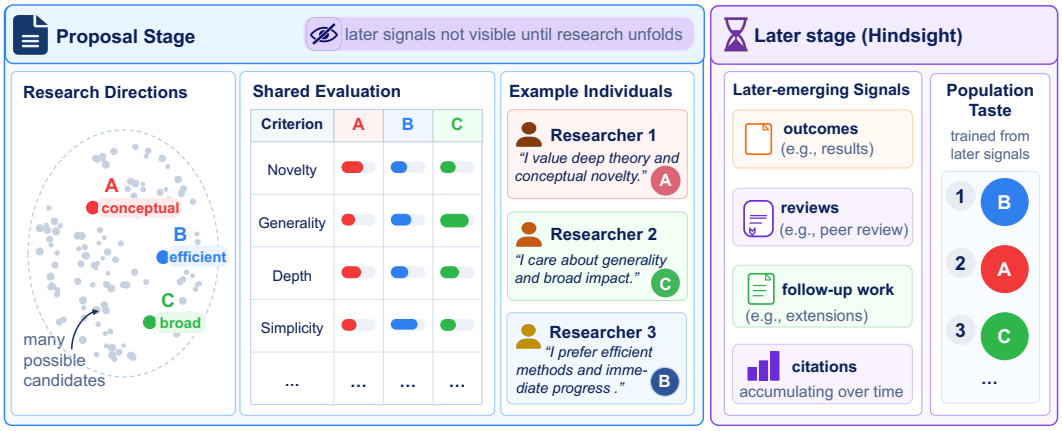

Figure 1: **Modeling scientific taste.** Proposal-stage evidence places research directions in a shared evaluation space, while researchers may map the same criterion profiles to different preferences. Later-emerging evidence can provide retrospective supervision for population-level learning but is unavailable at proposal-time inference.

choices: one researcher may favor conceptual novelty, while another prioritizes broad applicability. Although these criteria are neither fully objective nor exhaustive of scientific judgment, they provide a common basis for modeling variation across researchers.

A separate challenge is that most useful supervision signals emerge after research unfolds. Outcomes, community reviews, follow-up work and citations can reveal how a direction is received, but such later-emerging signals are unavailable when directions must be judged at the proposal stage. Their availability is also uneven across fields: Computer Science venues such as ICLR expose structured public reviews through OpenReview (Soergel et al., 2013), while comparable records are more limited in Quantum, where platforms such as SciRate (Collaboration) provide weaker community signals (scites). When aggregated, these later signals can still support learning a population-level taste mapping, which provides a useful cold start for personalization under sparse direct feedback.

To make such retrospective supervision usable at proposal time, we introduce **Privileged Hindsight Distillation (PHD)**. In PHD, a privileged teacher uses later-emerging evidence to refine the rubrics, conditional rules, and memory of a proposal-only student, improving both shared evaluation and population-level taste learning while preserving proposal-only inference. The resulting population mapping is then adapted to individual researchers using sparse direct feedback.

We observe individual scientific taste through pairwise proposal comparison, a lightweight interface for eliciting relative preferences without requiring absolute scores (Dras, 2015; Narimanzadeh et al., 2023; Baig et al., 2026). We evaluate our approach primarily in quantum research, with additional cross-domain experiments using real and simulated researchers. Our results show that PHD improves proposal-time evaluation and population-level prior learning, while sparse feedback enables adaptation to individual researchers. We further observe substantial disagreement across researchers, particularly when competing directions trade off across evaluation criteria, and deploy the complete framework in an interactive proposal exploration system, where personalized rankings more effectively surface proposals that users subsequently choose to inspect, refine, or pursue.

Contributions. **(1)** We formulate scientific taste by separating criterion-level assessment from how those assessments are valued. A shared, evidence-grounded evaluation space supports population and researcher-specific mappings to relative preferences, with the former serving as a transferable prior for individual adaptation. **(2)** We introduce Privileged Hindsight Distillation (PHD), which uses later-emerging scientific evidence as privileged supervision to learn criterion-level assessment and the population prior while preserving proposal-only inference. **(3)** We show that individual scientific taste can diverge substantially from population-level judgment: researchers make sharply different choices over the same research-direction pairs, and researcher-specific conditional mappings recover preference patterns that fixed criterion reweighting misses under sparse feedback.

2 RELATED WORK

Modeling Scientific Taste. Prior work measures research taste through citation-based impact (Koltun & Hafner, 2021), or learns scientific judgment from citation comparisons (Tong et al., 2026), institutional publication decisions (Gong et al., 2026), and future-impact supervision for peer review (Rasool et al., 2026). More recent work has moved scientific judgment toward earlier research stages: TASTE (Baig et al., 2026) benchmarks expert pairwise preferences of AI-Safety research proposals, while other studies compare human and LLM-generated ideas using a research-taste taxonomy (Chen et al., 2026b) or explicitly encode researcher-specific preferences (Zhang et al., 2026b). These studies examine aggregate judgment, proposal-level preferences, or differences in ideation patterns, whereas our focus is learning researcher-specific preferences from sparse feedback.

Privileged Information Learning. Learning using privileged information formalizes supervision that is available during training but absent at deployment (Vapnik & Vashist, 2009; Lopez-Paz et al., 2015). Recent language-model methods similarly distill privileged information into deployment-restricted policies (Yang et al., 2026; Zhao et al., 2026b); OPSD in particular lets a privileged teacher guide an unprivileged student on its own rollouts. Reflection and textual optimization methods improve agentic artifacts like prompts and LM programs (Yuksekgonul et al., 2024; Agrawal et al., 2026; Opsahl-Ong et al., 2024), reusable experience (Zhao et al., 2024), and episodic memory (Shinn et al., 2023). Our proposed PHD connects these two categories by exposing retrospective scientific evidence only to the teacher while distilling its benefit into a proposal-time agentic workflow.

Sparse Preference Learning. Personalized preference learning studies how user-specific judgments can be recovered from sparse supervision, including few-shot adaptation and population-informed preference learning (Poddar et al., 2024; Singh et al., 2025; Bose et al., 2026). Relevant work models heterogeneous preferences through structured preference representations (Shenfeld et al., 2025; Bose et al., 2025), while perspectivist learning preserves annotator-specific disagreement rather than collapsing judgments into a single consensus (Davani et al., 2022; Xu & Jurgens, 2026; Orlikowski et al., 2025; Sorensen et al., 2024). When feedback arrives sequentially, continual-learning methods commonly use replay to incorporate new observations while mitigating forgetting (Chaudhry et al., 2019; Aljundi et al., 2019). Our setting brings these two problems together by learning individual scientific taste from sparse, sequential researcher feedback.

3 METHOD

Our framework learns individual scientific taste from sparse feedback through three components, as illustrated in Figure 2(a). **Evaluation Space Learning** maps each research direction to a researcher-independent representation consisting of criterion-level assessments and associated proposal-visible evidence. **Population Taste Distillation** learns a population-level mapping from this shared representation to relative preferences. **Individual Taste Adaptation** then adapts the population mapping into a researcher-specific taste mapping using sparse sequential feedback. The first two components are improved through Privileged Hindsight Distillation (PHD), described in Section 3.2.

3.1 SHARED EVALUATION SPACE AND SCIENTIFIC TASTE

Let $\mathcal{C}$ be a finite set of candidate research directions, with $P \in \mathcal{C}$ representing one direction through its proposal. All proposal-stage information, including the proposal and literature available at that time, is denoted by $\mathcal{I}_{\text{vis}}(P)$, with $\mathcal{E}(P)$ representing the corresponding proposal-visible evidence. Both depend only on P, not on the researcher evaluating it. We define a shared K-dimensional evaluation space $\mathcal{S}$, with one dimension per evaluation criterion. A criterion assessor maps proposal-visible information to $\hat{s}(P) = f_{\text{crit}}(\mathcal{I}_{\text{vis}}(P)) = (\hat{s}_1(P), \ldots, \hat{s}_K(P)) \in \mathcal{S}$, where each $\hat{s}_k(P)$ is researcher-independent. Researchers therefore share the same criterion representation but may translate it and its supporting evidence into different preferences.

For researcher u, scientific taste is a researcher-specific mapping from criterion assessments and proposal-visible evidence to relative preferences, $T_u : \{(\hat{s}(P), \mathcal{E}(P))\}_{P \in \mathcal{C}} \longmapsto \succ_u$, where $A \succ_u B$ means that u prefers A to B. We observe scientific taste through pairwise comparisons. For $A, B \in$

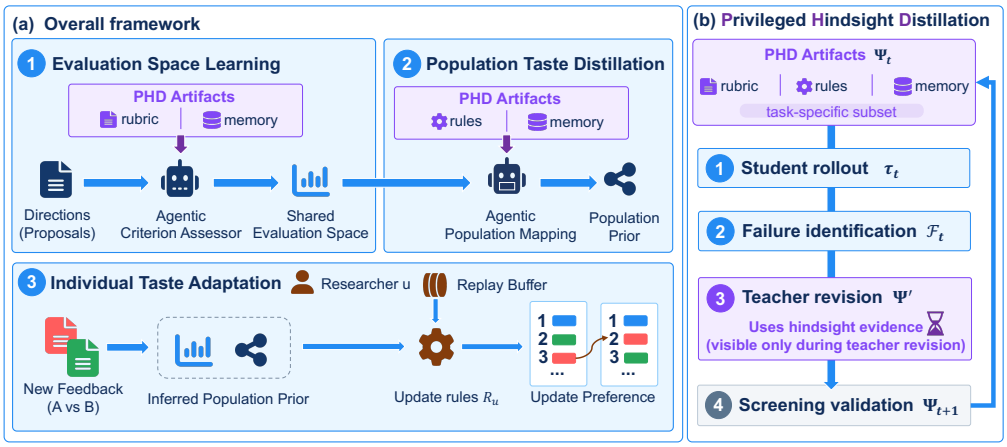

Figure 2: **Overview of our framework.** (a) The framework learns a shared evaluation space, distills a population-level taste mapping, and sequentially adapts it to individual researchers; PHD refines the task-specific artifacts used by the first two components. (b) PHD iteratively refines these artifacts using privileged hindsight during teacher revision while preserving proposal-only inference.

$\mathcal{C}$, let $x_{A,B} = \hat{s}(A) - \hat{s}(B)$ denote their criterion-difference vector and $\mathcal{E}_{A,B} = (\mathcal{E}(A), \mathcal{E}(B))$ their proposal-visible evidence. We instantiate T_u through a scalar preference function U_u, such that $A \succ_u B \iff U_u(x_{A,B}, \mathcal{E}_{A,B}) > 0$. Section 3.4 defines the shared functional form of U, and Section 3.5 personalizes it from researcher feedback.

3.2 Privileged Hindsight Distillation

Scientific judgments must be made at the proposal stage, before completed methods, experimental results, subsequent literature, citations, and community feedback become available. Once research unfolds, these later observations provide additional evidence for diagnosing errors in proposal-time judgments. We denote such later-emerging information by $\mathcal{I}_{\text{priv}}$ and introduce **Privileged Hindsight Distillation (PHD)** to exploit it during training while preserving proposal-only inference.

Inspired by On-Policy Self-Distillation (OPSD) (Zhao et al., 2026b), PHD keeps the base model frozen and instead refines the external artifacts governing the student. We denote the artifact state at refinement round t by Ψ_t, comprising a task-specific subset of rubric, conditional rules, and memory. Rubrics specify criteria, scoring anchors, and evidence requirements; conditional rules encode context-dependent adjustments; and memory stores reusable proposal-visible guidance.

As illustrated in Figure 2(b), PHD provides a shared four-stage refinement procedure for Evaluation Space Learning and Population Taste Distillation. For either task, let z denote a training input and $y^*(z)$ its training-only reference target; their specific forms are defined in the following sections.

1. Student rollout. Given proposal-visible information and the current artifacts Ψ_t, the student produces a rollout $\tau_t(z)$ containing the task prediction $\hat{y}_t(z)$, confidence, and proposal-visible supporting evidence. Neither $y^*(z)$ nor $\mathcal{I}_{\text{priv}}$ is available to the student.

2. Failure identification. The analyzer compares $\hat{y}_t(z)$ with the reference target $y^*(z)$ using the task-specific error criterion and collects failed cases into $\mathcal{F}_t$. Each failed case contains z, its rollout $\tau_t(z)$, and $y^*(z)$. At this stage, the analyzer only identifies prediction failures; it does not access the privileged hindsight evidence used to construct $y^*(z)$ or diagnose their causes.

3. Teacher revision. The training-only teacher receives Ψ_t, $\mathcal{F}_t$, and $\mathcal{I}_{\text{priv}}$. Privileged hindsight is used only at this stage to diagnose the identified failures and propose a candidate revision Ψ'_t.

4. Screening and validation. The candidate revision first undergoes leakage screening, which rejects artifacts containing privileged outcomes, reference answers, or conditions unavailable at the proposal stage. Revisions that pass screening are evaluated on held-out data using the task-specific validation metric and are accepted as Ψ_{t+1} when they satisfy the corresponding acceptance criterion.

Accepted artifacts are reused in the next refinement round. At inference, only the final accepted artifacts and proposal-visible information remain; reference targets, the analyzer, teacher, and privileged hindsight are removed. Sections 3.3 and 3.4 specify the task input, reference target, editable artifacts, failure criterion, and validation metric for the two shared components.

3.3 Evaluation Space Learning

Evaluation Space Learning instantiates the criterion assessor f_{crit} from Section 3.1. We use a shared $K = 6$-dimensional space covering generality, novelty, importance, parsimony, depth, and algorithmic complexity. Following rubric-based evaluation (Kim et al., 2024) and evidence-grounded scientific assessment (Zhang et al., 2026a), we construct two rubrics over the same criteria and score scales but with different information access.

Rubric A is the fixed reference rubric. It uses both proposal-visible information $\mathcal{I}_{\text{vis}}(P)$ and privileged hindsight $\mathcal{I}_{\text{priv}}(P)$ to produce the reference criterion vector $s^*(P) = (s_1^*(P), \ldots, s_K^*(P))$. **Rubric B** is the student rubric and a component of Ψ. It uses only $\mathcal{I}_{\text{vis}}(P)$ and governs the criterion assessor producing $\hat{s}(P)$. The details of two rubrics are provided in Appendices B.1 and B.2.

For PHD, the task instance is $z = P$, the reference target is $y^*(P) = s^*(P)$, and the rollout $\tau_t(P)$ contains $\hat{s}_t(P)$, confidence, and proposal-visible supporting evidence. The analyzer computes per-proposal MAE across the K criteria and uses high-error proposals to form the failure set $\mathcal{F}_t$. The editable artifacts are Rubric B and case memory. When the assessor repeatedly mis-scores a criterion by at least one score level, we revise the corresponding scoring anchor in Rubric B. Isolated failures are instead stored as proposal-visible case lessons in memory, each with a triggering condition and an absolute-level check. Held-out MAE determines whether a revision is retained. Complete prompts and an illustrative revision are provided in Appendices H.1 and B.4.

3.4 Population Taste Distillation

For this PHD instantiation, the task input is a proposal pair $z = (A, B)$. The student observes the criterion-difference vector $x_{A,B}$ and proposal-visible evidence $\mathcal{E}_{A,B}$ defined in Section 3.1, and predicts $\hat{y}_t(A, B) \in \{A, B\}$. The corresponding population reference target is $y^*(A, B)$, whose construction is detailed in Appendix C.

We model population and individual preferences using a common preference function. Each conditional rule r is evaluated independently on a research direction using only its proposal-visible representation. Let $a_r(P)$ denote its activation on P; for a pair (A, B), its contribution is $\phi_r(x_{A,B}, \mathcal{E}_{A,B}) = a_r(A) - a_r(B)$. The preference function is

$$U(x, \mathcal{E}; w, R) = w^\top x + \sum_{r \in R} \tilde{\alpha}_r \, \phi_r(x, \mathcal{E}),$$

where w captures a fixed weighting over criterion differences, R contains conditional rules, and $\tilde{\alpha}_r$ is the fitted rule coefficient adjusted by held-out reliability. The linear term captures preferences expressible by fixed criterion reweighting, while the rules capture context-dependent judgments beyond it. This functional form is shared by the population and individual mappings.

We first fit w_{pop} on population training pairs using $y^*(A, B)$ as supervision and no conditional rules, yielding $U_{\text{pop}}^{(0)}(x) = w_{\text{pop}}^\top x$. We then freeze w_{pop} and apply PHD to learn context-dependent residual judgments. The editable artifacts are population rules R_{pop} and memory: rules encode reusable context-dependent adjustments, while memory stores proposal-visible guidance for applying them. The resulting population mapping is $U_{\text{pop}}(x, \mathcal{E}) = U(x, \mathcal{E}; w_{\text{pop}}, R_{\text{pop}})$.

StrictPair, defined in Section 4.1, identifies failed pairs for $\mathcal{F}_t$ and serves as the held-out validation metric. The resulting U_{pop} provides the population prior for individual adaptation. Complete workflow and implementation details are provided in Appendix H.2.

3.5 Individual Taste Adaptation

Individual Taste Adaptation learns sparse sequential feedback of researcher u , stored in a time-ordered replay buffer $\mathcal{B}_u$. Each observation contains a proposal pair (A, B), its representation $x_{A,B}$

and $\mathcal{E}_{A,B}$ defined in Section 3.1, and the researcher preference $y_u(A, B) \in \{A, B\}$. Using the common preference function from Section 3.4, we personalize the population mapping through an individual residual Δw_u and researcher-specific rules R_u: $U_u(x, \mathcal{E}) = U(x, \mathcal{E}; w_u, R_{\text{pop}} \cup R_u)$, where $w_u = w_{\text{pop}} + \Delta w_u$. Adaptation starts from U_{pop} with $\Delta w_u = 0$ and $R_u = \emptyset$.

As feedback is appended to $\mathcal{B}_u$, Δw_u is refitted using the frozen population score as an offset, allowing criterion weights to deviate from the population prior. Recurring leave-one-out errors that remain after this linear adaptation motivate researcher-specific conditional rules. Each candidate rule is evaluated through cross-fitted replay over $\mathcal{B}_u$: in each leave-one-out fold, Δw_u and the rule coefficient are refitted on the remaining observations and evaluated on the held-out support comparison. A rule is retained only if replay yields a net predictive improvement without breaking previously correct preferences. Held-out query labels and privileged hindsight are never used during individual adaptation. Further fitting and rule-induction details are provided in Appendix H.3.

4 Experiments

4.1 Learning Individual Scientific Taste

Data. Quantum serves as our primary evaluation domain. We reconstruct research directions as proposal-stage inputs containing the research problem, methodology, evaluation plan, expected contributions, and risks, while excluding later-emerging information. All compared methods receive only proposal-visible inputs; as with other pretrained LLM evaluations, we do not assume control over parametric knowledge acquired during pretraining.

Individual-taste supervision must be collected directly from researchers, and is therefore naturally scarce. Each of seven quantum-algorithm researchers provides 15 sequential support comparisons for adaptation and 10 held-out query pairs for evaluation, with a binary preference and confidence annotation for each comparison. We complement this real-researcher cohort with 100 simulated researchers: 50 derived from publication histories of prominent quantum researchers (Sim.-Q) and 50 generated through controlled variation in criterion trade-offs (Sim.-Ctrl.). Further construction and supervision details are provided in Appendix E.

Evaluation metric. For pairwise preference prediction, we use **StrictPair (SP)**, which requires both presentation orientations of a proposal pair to be predicted correctly. For the i-th pair (A_i, B_i),

$$\text{SP}_i = \mathbb{I}[\hat{y}(A_i, B_i) = y^*(A_i, B_i) \ \wedge \ \hat{y}(B_i, A_i) = y^*(B_i, A_i)] \,,$$

where $\hat{y}$ and y^* denote the predicted and target preferences, respectively.

For individual taste, we report **Regret-Weighted StrictPair (RW-SP)** to account for researcher confidence. Following regret-based pairwise weighting (Halpern et al., 2025), comparison i receives weight $d_i = 2|p_i - 0.5|$, where p_i is derived from its confidence annotation (Appendix F.5). We compute $\text{RW-SP} = (\sum_i d_i \, \text{SP}_i)/(\sum_i d_i)$.

Baselines. We compare personalization approaches with increasingly explicit use of researcher feedback. Unless otherwise specified, all methods use `gpt-5.5` as the backbone, and no privileged hindsight is available during individual adaptation or inference. **Single Call (SC)** directly places all support comparisons in the prompt when predicting each held-out preference, while **Profile Prompting (PP)** first summarizes them into a researcher profile. SynthesizeMe (Ryan et al., 2025), FERMI (Kim & Yang, 2025), and CIPHER (Gao et al., 2024) represent agentic approaches that maintain or refine user-specific descriptions or preferences from interaction history. VPL (Poddar et al., 2024), LoRe (Bose et al., 2025), and FSPO (Singh et al., 2025) instead learn parameterized user-specific preference or reward representations. Detailed configurations are in Appendix F.

Results. Table 1 shows that representing the same sparse feedback through in-context or agentic personalization remains less effective than explicitly learning preference structure. Our method performs best on both real and simulated researchers while keeping the backbone frozen, indicating that explicit preference structure can be introduced through a transferable population mapping and lightweight researcher-specific state without per-user model-parameter updates.

Table 1: **Individual scientific taste prediction.** RW-SP across real and simulated researchers.

	In-context		Agentic				Parametric		
	SC	PP	SynthesizeMe	FERMI	CIPHER	**Ours**	VPL	LoRe	FSPO
Real ↑	0.479	0.481	0.477	0.419	0.524	**0.693**	0.629	0.595	<u>0.642</u>
Sim. ↑	0.423	0.519	0.473	0.366	0.399	**0.591**	0.556	0.547	<u>0.559</u>

Table 2: **Robustness across backbones.** E-MAE/E-OA denote MAE/Ordering Accuracy for Evaluation Space Learning, and P-SP denotes StrictPair for Population Taste Distillation. Shared-component entries report **Ours / w/o PHD**.

	Evaluation Space		Population	Individual RW-SP ↑		
Backbone	E-MAE ↓	E-OA ↑	P-SP ↑	Ours	PP	CIPHER
`gpt-5.5`	**0.183** / 0.216	**0.721** / 0.713	**0.833** / 0.667	**0.693**	0.481	0.524
`gpt-5.4-mini`	**0.233** / 0.292	**0.665** / 0.604	**0.667** / 0.567	**0.613**	0.407	0.332
`claude-opus-5`	**0.217** / 0.275	**0.675** / 0.612	**0.500** / 0.367	**0.664**	0.634	0.516
`gemini-3.6-flash`	**0.216** / 0.258	**0.699** / 0.662	**0.767** / 0.367	**0.647**	0.422	0.405

4.2 Robustness and Sensitivity Analysis

Shared-component data. We further evaluate whether PHD successfully distills privileged hindsight in the two shared components that support individual taste learning. Both Evaluation Space Learning and Population Taste Distillation use carefully curated, paper-disjoint splits. Constructing these evaluations requires controlled proposal-time reconstruction and privileged reference construction, while population comparisons additionally require temporally comparable and informative proposal pairs. A proposal may participate in multiple comparisons within the same split, but no underlying paper is shared across splits.

For Evaluation Space Learning, reference criterion vectors $s^*(P)$ are constructed with the fixed Rubric A using both proposal-visible information and privileged hindsight evidence. For Population Taste Distillation, reference preferences $y^*(A, B)$ are aggregated from a multi-family LLM-as-a-Judge panel given the same information. Since these are LLM-generated references rather than independent human labels, we use them to evaluate whether PHD can distill hindsight supervision into proposal-time judgments and for corresponding ablations, rather than for external benchmarking. Construction details are provided in Appendices E, B.1, and C.

Evaluation metrics. For Evaluation Space Learning, we report mean absolute error (MAE), computed as $\mathrm{MAE} = \sum_{P\in\mathcal{P}} \sum_{k=1}^{K} |\hat{s}_k(P) - s_k^*(P)|/(|\mathcal{P}|K)$, and Ordering Accuracy (OA) against the reference criterion vectors. Ordering Accuracy measures whether relative criterion-level orderings are preserved. For each criterion k, a proposal pair is correct when the predicted and reference differences have the same sign:

$$\mathrm{OA} = \frac{1}{|\mathcal{Q}|} \sum_{(A,B,k)\in\mathcal{Q}} \mathbb{I}[\mathrm{sgn}(\hat{s}_k(A) - \hat{s}_k(B)) = \mathrm{sgn}(s_k^*(A) - s_k^*(B))]\,,$$

where $\mathcal{Q}$ contains evaluated proposal-pair–criterion combinations. For Population Taste Distillation, since each reference preference is aggregated across multiple LLM judges, we report SP described in Section 4.1 rather than RW-SP, as judge-specific confidence no longer provides a meaningful pair-level measure of preference strength.

Robustness across backbones. We examine whether the gains from PHD persist across backbone models by substituting the backbone across all three learning components. For Evaluation Space Learning and Population Taste Distillation, we compare the complete method with its w/o PHD counterpart under each backbone. For Individual Taste Adaptation, we compare our complete method with PP and CIPHER, the strongest baselines within their respective categories, using the

same backbone. Table 2 shows that PHD improves criterion estimation and population preference prediction under all four backbones. At the individual level, our complete method also outperforms PP and CIPHER for every backbone, although absolute RW-SP varies across models. These results indicate that the observed gains are not specific to the default `gpt-5.5` configuration.

Sensitivity to population prior. We isolate the role of population knowledge by fixing the `gpt-5.5` backbone and individual adaptation procedure while replacing only the population prior, using the same researcher cohorts and RW-SP protocol as Section 4.1. We consider three existing judges as alternative priors: Auto-J (Li et al., 2024), a general-purpose pairwise judge; MLR-Judge (Chen et al., 2026a), which evaluates research proposals through rubric-based assessment; and SciJudge (Tong et al., 2026), a scientific judge for comparing scientific work. As shown in Table 3, differences between population priors persist after identical individual adaptation. Sparse individual feedback therefore does not fully overwrite the population initialization, indicating that the choice of population prior continues to affect downstream personalization.

Table 3: **Sensitivity to population prior.**

Population Prior	RW-SP ↑
Auto-J	0.583
SciJudge	0.644
MLR-Judge	0.667
Ours	**0.693**

4.3 Downstream Evaluation

Data and protocol. We deploy the complete framework in an interactive proposal-exploration system to test whether learned individual taste improves downstream research-direction discovery. We evaluate 20 real researchers from quantum, computer science, statistics, and medicine, together with simulated cohorts evaluated either within quantum or under cross-domain transfer, where taste learned from quantum is tested on proposals from other fields. Unlike the fixed offline evaluation in Section 4.1, accumulated interactions dynamically shape the candidate pool in subsequent rounds.

We evaluate both real and simulated users with different comparator sets. For real users, we compare against Single Call (SC) in a matched setting with the same backbone, candidate pool, proposal-visible information, and interaction history; SC directly conditions on the history, whereas ours updates an explicit individual taste mapping. Given the cost of real-user interaction, we restrict live comparison to this matched setting. Simulated-user evaluation additionally includes Profile Prompting (PP) and SynthesizeMe (SM), the strongest in-context and agentic baselines in Table 1.

For each comparison, we evaluate our method against one alternative method at a time using source-blind competitive-pair interleaving (Radlinski et al., 2008; Zhang et al., 2025), a widely used protocol for online evaluation in ranking and recommendation systems. Proposals selected by both methods are shown once, while ranking disagreements form adjacent competitive pairs with randomized order and hidden method identity. Each proposal receives an engagement score $e(P) \in \{0, 1, 2, 3\}$, denoting no substantive engagement, inspection, refinement, and adoption, respectively. Further system and evaluation details are provided in Appendix G.

Evaluation metrics. For the N_{CP} ranking disagreements, we define **Competitive-Pair Engagement Gain** as $\Delta_{\mathrm{CP}} = \sum_{c=1}^{N_{\mathrm{CP}}} (e_c^{\mathrm{ours}} - e_c^{\mathrm{base}})/N_{\mathrm{CP}}$, measuring the average engagement advantage within directly competing proposal pairs.

We also report **Win Rate**, the fraction of non-tied pairs with $e_c^{\mathrm{ours}} > e_c^{\mathrm{base}}$. To evaluate the full ranking beyond local disagreements, we treat engagement as graded relevance and report $\Delta_{\mathrm{NDCG}} = \mathrm{NDCG}_{\mathrm{ours}} - \mathrm{NDCG}_{\mathrm{base}}$ (Järvelin & Kekäläinen, 2002).

Table 4: **Downstream evaluation.**

Setting	Δ_{CP} ↑	Win Rate ↑	Δ_{NDCG} ↑
Simulated researchers			
Quantum / SC	0.169	0.667	0.064
Quantum / PP	0.130	0.571	0.020
Quantum / SM	0.146	0.641	0.037
Cross-domain / SC	0.120	0.583	0.006
Cross-domain / PP	0.067	0.577	0.018
Cross-domain / SM	0.151	0.635	0.044
Real researchers / SC	0.125	0.684	0.023

Results. Table 4 shows that learned individual taste improves proposal discovery over different baselines as the candidate pool evolves through interaction. The gains persist under cross-domain transfer, though with varying magnitude, highlighting differences in scientific taste across both individuals and research domains.

4.4 Role of Privileged Hindsight

To isolate privileged hindsight, we use the shared-component evaluation from Section 4.2 under `gpt-5.4-mini`, comparing w/o PHD (no artifact refinement), w/o $\mathcal{I}_{\text{priv}}$ (proposal-visible refinement only), and Ours (privileged hindsight available to the teacher).

Table 5 shows that proposal-visible refinement substantially improves criterion-level assessment over w/o PHD, but does not improve population preference prediction. Adding privileged hindsight yields a smaller further gain in E-MAE and E-OA but a much larger improvement in P-SP. This contrast suggests that hindsight is particularly important for learning how criterion-level differences translate into population preferences, rather than for criterion assessment itself.

Table 5: **Effect of hindsight.**

Variant	E-MAE ↓	E-OA ↑	P-SP ↑
Ours	**0.233**	**0.665**	**0.667**
w/o $\mathcal{I}_{\text{priv}}$	0.242	0.649	0.533
w/o PHD	0.292	0.604	0.567

4.5 Understanding Individual Taste Learning

Beyond Population Initialization. We first examine the two individual adaptation components under the same cohorts and RW-SP protocol as Section 4.1. As shown in Table 6, for real researchers, the population initialization is already strong, and residual reweighting alone does not improve the aggregate result. Conditional rules yield a modest gain on real researchers and Sim.-Q, whereas residual adaptation is sufficient on Sim.-Ctrl., whose preferences are explicitly constructed from criterion-level trade-offs. This suggests that some researcher-specific preferences require context-dependent adaptation beyond fixed criterion reweighting, motivating analysis of when individual taste departs from the population prior rather than assuming uniform gains. Appendix D.2 illustrates this with two real users (U04 and U05), whose opposing choices emerge only after individual adaptation. Removing PHD and adapting from the seed Rubric B and a randomly initialized prior further reduces RW-SP, suggesting that the PHD-derived shared structure provides a better starting point for personalization.

Table 6: **Individual adaptation ablation.** RW-SP across researcher cohorts.

Variant	Real ↑	Sim.-Q ↑	Sim.-Ctrl. ↑
Population	0.680	0.537	0.592
+ Residual	0.679	0.573	**0.606**
+ Rules	**0.693**	**0.588**	0.593
w/o PHD	0.670	0.540	0.581

Replay under sequential feedback. We ablate the cross-fitted replay gate from Section 3.5 under the same sequential support protocol as Section 4.1, while keeping the population initialization, feedback sequence, residual fitting, rule proposer, and coefficient estimation fixed. Since this analysis repeatedly evaluates adaptation over multiple support prefixes, we use gpt-5.4-mini to reduce inference cost. Table 7 shows that replay has little effect early but becomes beneficial as researcher-specific preference structure emerges. Held-out RW-SP remains non-monotonic, indicating that later feedback can revise or interfere with earlier preferences and motivating replay to preserve prior evidence.

Table 7: **Effect of replay-based rule acceptance.**

Support	0	5	10	12	13	14	15
Ours	0.538	0.535	0.580	0.603	0.648	0.599	0.613
w/o replay	0.538	0.535	0.580	0.575	0.615	0.578	0.602

5 Conclusion

We propose a framework for learning individual scientific taste from sparse researcher feedback, representing research directions in a shared, evidence-grounded evaluation space and adapting a population-level preference mapping to individual researchers. Privileged Hindsight Distillation uses later-emerging evidence during training to improve proposal-time assessment and population-level learning without exposing hindsight at inference. Experiments reveal substantial researcher-specific variation, including opposing preferences over the same proposal pairs, and show that conditional rules capture preference patterns beyond fixed criterion reweighting in real researchers. In interactive deployment, personalized rankings also surface proposals associated with greater subsequent engagement than the matched Single Call baseline. While our study is limited by sparse real-expert feedback and a finite set of evaluation criteria, the results suggest that population-level judgment provides a useful starting point but does not fully capture individual scientific preferences.

AI USE STATEMENT

Generative AI tools were used both as components of the proposed research methodology and as aids throughout the research process. Within the methodology, large language models are used to construct synthetic researcher representations and simulated interaction trajectories, provide proxy population-level preference supervision through a multi-family panel, extract structured research-philosophy signals from public scholarly corpora for simulated-user construction, and instantiate the student, teacher, analyzer, and auditing agents in Privileged Hindsight Distillation.

Additionally, generative AI tools were used for sourcing/searching for information, and language editing. All AI-assisted outputs were reviewed by the authors. LLM-generated population-preference judgments and researcher representations were treated as proxy supervision, evaluation references, or simulation artifacts rather than objective ground truth about scientific quality or individual researchers. Experimental claims were checked against executed results, and references and literature-related claims included in the manuscript were verified against the corresponding source material. The authors take responsibility for the final content of this work, including all text, claims, experimental results, citations, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies researcher-specific scientific preferences and includes both expert preference annotations and interaction data collected from researchers using the deployed system. These signals are used to study scientific taste and evaluate proposal ranking rather than to assess researchers themselves or make consequential decisions about individuals. All participants provided informed consent for the collection and research use of their preference and interaction data.

Results involving real researchers are reported in aggregate or using pseudonymous participant identifiers, and synthetic-user evaluations are clearly separated from evaluations involving human participants. Researcher representations constructed from public scholarly corpora are used to instantiate simulated interaction behavior and should not be interpreted as faithful psychological representations of the corresponding researchers. Such representations may incompletely or inaccurately characterize scientific preferences and are not intended for researcher assessment or other consequential uses.

The proposed system may also inherit biases from language models, scientific literature, citation patterns, and community-response signals on which parts of the workflow rely. We therefore treat LLM-panel judgments and retrospective community signals as sources of proxy supervision or privileged evidence rather than objective measures of scientific quality. The system is intended to support researchers in exploring and comparing research directions rather than to replace expert scientific judgment or autonomously determine which research should be pursued.

REPRODUCIBILITY STATEMENT

We document the implementation and evaluation details needed to reproduce the study across the main paper and Appendix. The Method section specifies the shared evaluation space, population-level and researcher-specific taste mappings, the Privileged Hindsight Distillation workflow, and the sequential individual adaptation procedure. The Experimental Setup describes the data construction and evaluation protocols, model families, baselines, researcher cohorts, and evaluation metrics. The Appendix provides further details on the rubrics for the six evaluation criteria, proposal reconstruction and retrospective evidence construction, data splits, agent prompts and textual instruments, artifact acceptance and leakage-control mechanisms, synthetic researcher construction, model and inference configurations, replay and adaptation procedures, and the deployment and source-blind interleaving protocols. Together, these materials specify the learning and evaluation pipeline used to obtain the reported results.

REFERENCES

Lakshya A Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziems, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J Ryan, Meng Jiang, et al. Gepa: Reflective prompt

evolution can outperform reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pp. 8479–8565, 2026.

Rahaf Aljundi, Eugene Belilovsky, Tinne Tuytelaars, Laurent Charlin, Massimo Caccia, Min Lin, and Lucas Page-Caccia. Online continual learning with maximal interfered retrieval. *Advances in neural information processing systems*, 32, 2019.

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.

Hasan Baig, Hailey Joren, and Joe Benton. Taste: Can ai models judge ai safety re-search proposals? *Anthropic Alignment Science Blog*, 2026.

Lutz Bornmann, Rüdiger Mutz, and Hans-Dieter Daniel. A reliability-generalization study of journal peer reviews: A multilevel meta-analysis of inter-rater reliability and its determinants. *PloS one*, 5(12):e14331, 2010.

Avinandan Bose, Zhihan Xiong, Yuejie Chi, Simon Shaolei Du, Lin Xiao, and Maryam Fazel. Lore: Personalizing llms via low-rank reward modeling. *arXiv preprint arXiv:2504.14439*, 2025.

Avinandan Bose, Shuyue Stella Li, Faeze Brahman, Pang Wei Koh, Simon Shaolei Du, Yulia Tsvetkov, Maryam Fazel, Lin Xiao, and Asli Celikyilmaz. Cold-start personalization via training-free priors from structured world models. *arXiv preprint arXiv:2602.15012*, 2026.

Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K Dokania, Philip HS Torr, and Marc'Aurelio Ranzato. On tiny episodic memories in continual learning. *arXiv preprint arXiv:1902.10486*, 2019.

Hui Chen, Miao Xiong, Yujie Lu, Wei Han, Ailin Deng, Yufei He, Jiaying Wu, Yibo Li, Yue Liu, and Bryan Hooi. Mlr-bench: Evaluating ai agents on open-ended machine learning research. *Advances in Neural Information Processing Systems*, 38, 2026a.

Ziyu Chen, Yilun Zhao, and Arman Cohan. Measuring the gap between human and llm research ideas. *arXiv preprint arXiv:2607.01233*, 2026b.

SciRate Collaboration. SciRate. URL `https://scirate.com/about`. Accessed September 18, 2026.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110, 2022.

Tianyu Ding, Aditya Nannapaneni, Bingfan Liu, and Ling Zhang. Autonomous research agents: A survey of ai scientists and the verification gap. *arXiv preprint arXiv:2608.05179*, 2026.

Mark Dras. Squibs: Evaluating human pairwise preference judgments. *Computational Linguistics*, 41(2):309–317, 2015.

Ge Gao, Alexey Taymanov, Eduardo Salinas, Paul Mineiro, and Dipendra Misra. Aligning llm agents by learning latent preference from user edits. *Advances in neural information processing systems*, 37:136873–136896, 2024.

Ziqin Gong, Ning Li, and Huaikang Zhou. Llms learn scientific taste from institutional traces across the social sciences. *arXiv preprint arXiv:2603.16659*, 2026.

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.

Daniel Halpern, Evi Micha, Ariel D. Procaccia, and Itai Shapira. Pairwise calibrated rewards for pluralistic alignment. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446, 2002.

Bo Jiang. Hindsight: Evaluating llm-generated research ideas via future impact. *arXiv preprint arXiv:2603.15164*, 2026.

Jaehyung Kim and Yiming Yang. Few-shot personalization of llms with mis-aligned responses. 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. 11943–11974, 2025.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. Prometheus 2: An open source language model specialized in evaluating other language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4334–4353, 2024.

Vladlen Koltun and David Hafner. A measure of research taste. *arXiv preprint arXiv:2105.08089*, 2021.

Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, Pengfei Liu, et al. Generative judge for evaluating alignment. In *International Conference on Learning Representations*, volume 2024, pp. 27547–27574, 2024.

David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. *arXiv preprint arXiv:1511.03643*, 2015.

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.

Timothy J. Mitchison. A question of taste. *Molecular Biology of the Cell*, 24(21):3278–3280, 2013. doi: 10.1091/mbc.E13-07-0410.

Rüdiger Mutz, Lutz Bornmann, and Hans-Dieter Daniel. Heterogeneity of inter-rater reliabilities of grant peer reviews and its determinants: A general estimating equations approach. *PLoS One*, 7 (10):e48509, 2012.

Hasti Narimanzadeh, Arash Badie-Modiri, Iuliia G Smirnova, and Ted Hsuan Yun Chen. Crowdsourcing subjective annotations using pairwise comparisons reduces bias and error compared to the majority-vote method. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2): 1–29, 2023.

Krista Opsahl-Ong, Michael J Ryan, Josh Purtell, David Broman, Christopher Potts, Matei Zaharia, and Omar Khattab. Optimizing instructions and demonstrations for multi-stage language model programs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 9340–9366, 2024.

Matthias Orlikowski, Jiaxin Pei, Paul Röttger, Philipp Cimiano, David Jurgens, and Dirk Hovy. Beyond demographics: Fine-tuning large language models to predict individuals' subjective text perceptions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2092–2111, 2025.

Rongcan Pei, Fang Guo, Qinglin Qi, Qi Zhu, Yun Luo, Jianhao Yan, Minjun Zhu, Qiujie Xie, Dehong Zheng, and Yue Zhang. Deepinstructor: An agentic ai instructor for experience-driven idea evaluation. *arXiv preprint arXiv:2609.22104*, 2026.

Sriyash Poddar, Yanming Wan, Hamish Ivison, Abhishek Gupta, and Natasha Jaques. Personalizing reinforcement learning from human feedback with variational preference learning. *Advances in Neural Information Processing Systems*, 37:52516–52544, 2024.

Shuofei Qiao, Yunxiang Wei, Xuehai Wang, Bin Wu, Boyang Xue, Ningyu Zhang, Hossein A Rahmani, Yanshan Wang, Qiang Zhang, Keyan Ding, et al. Innoeval: On research idea evaluation as a knowledge-grounded, multi-perspective reasoning problem. *arXiv preprint arXiv:2602.14367*, 2026.

Filip Radlinski, Madhu Kurup, and Thorsten Joachims. How does clickthrough data reflect retrieval quality? In *Proceedings of the 17th ACM conference on Information and knowledge management*, pp. 43–52, 2008.

Abdur Rasool, Xiaohui Huang, Yanqing Hu, and Linyi Yang. Reviewguard: Aligning llm-assisted peer review with long-term scientific impact. *arXiv preprint arXiv:2606.24892*, 2026.

Michael J Ryan, Omar Shaikh, Aditri Bhagirath, Daniel Frees, William Held, and Diyi Yang. Synthesizeme! inducing persona-guided prompts for personalized reward models in llms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8045–8078, 2025.

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.

Idan Shenfeld, Felix Faltings, Pulkit Agrawal, and Aldo Pacchiano. Language model personalization via reward factorization. *arXiv preprint arXiv:2503.06358*, 2025.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652, 2023.

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.

Chenglei Si, Tatsunori Hashimoto, and Diyi Yang. The ideation-execution gap: Execution outcomes of llm-generated versus human research ideas. In *International conference on learning representations*, volume 2026, pp. 74195–74240, 2026.

Anikait Singh, Sheryl Hsu, Kyle Hsu, Eric Mitchell, Stefano Ermon, Tatsunori Hashimoto, Archit Sharma, and Chelsea Finn. Fspo: Few-shot optimization of synthetic preferences personalizes to real users. *arXiv preprint arXiv:2502.19312*, 2025.

Soumitra Sinhahajari, Navonil Majumder, and Soujanya Poria. On the limits of llm-as-judge for scientific novelty assessment. *arXiv preprint arXiv:2606.12071*, 2026.

David Soergel, Adam Saunders, and Andrew McCallum. Open scholarship and peer review: a time for experimentation. 2013.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, et al. A roadmap to pluralistic alignment. *arXiv preprint arXiv:2402.05070*, 2024.

Robert Tjian. Terri and g&d: celebrating 50 years of enlightened scientific judgment. *Genes & Development*, 37(1-2):6–8, 2023.

Jingqi Tong, Mingzhe Li, Hangcheng Li, Yongzhuo Yang, Yurong Mou, Weijie Ma, Zhiheng Xi, Hongji Chen, Xiaoran Liu, Qinyuan Cheng, et al. Ai can learn scientific taste. *arXiv preprint arXiv:2603.14473*, 2026.

Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural networks*, 22(5-6):544–557, 2009.

Chenrun Wang, Mingxuan Zhu, Tiancheng Huang, Wenjie Li, Yujie Zhang, Zichen Zhu, Zhiying Zou, Kai Yu, and Lu Chen. Ligbench: A unified and human-aligned benchmark for llm-based research idea generation. *arXiv preprint arXiv:2608.13136*, 2026.

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.

Yinuo Xu and David Jurgens. Beyond consensus: Perspectivist modeling and evaluation of annotator disagreement in nlp. *arXiv preprint arXiv:2601.09065*, 2026.

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.

Chenxu Yang, Chuanyu Qin, Qingyi Si, Minghui Chen, Naibin Gu, Dingyu Yao, Zheng Lin, Weiping Wang, Jiaqi Wang, and Nan Duan. Self-distilled rlvr. *arXiv preprint arXiv:2604.03128*, 2026.

Runlong Ye, Matthew Varona, Oliver Huang, Patrick Yung Kang Lee, Michael Liut, and Carolina Nobre. The design space of recent ai-assisted research tools for ideation, sensemaking, and scientific creativity. *arXiv preprint arXiv:2502.16291*, 2025.

Mert Yuksekgonul, Federico Bianchi, Joseph Boen, Sheng Liu, Zhi Huang, Carlos Guestrin, and James Zou. Textgrad: Automatic" differentiation" via text. *arXiv preprint arXiv:2406.07496*, 2024.

Ming Zhang, Kexin Tan, Yueyuan Huang, Yujiong Shen, Chunchun Ma, Li Ju, Xinran Zhang, Yuhui Wang, Wenqing Jing, Jingyi Deng, et al. Opennovelty: An llm-powered agentic system for verifiable scholarly novelty assessment. *arXiv preprint arXiv:2601.01576*, 2026a.

Qing Zhang, Alex Deng, Michelle Du, Huiji Gao, Liwei He, and Sanjeev Katariya. Harnessing the power of interleaving and counterfactual evaluation for airbnb search ranking. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 5205–5214, 2025.

Yiwen Zhang, Eloise Zeng, Jaeha Lee, and Tony Yue Yu. An ai scientist that doesn't drift: Taste, structure, and falsifiable findings in a quadruped navigation research loop. *arXiv preprint arXiv:2608.07542*, 2026b.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. Expel: Llm agents are experiential learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19632–19642, 2024.

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

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 2026b.

A Privileged Hindsight Distillation Details

We provide additional implementation details for the four-step PHD procedure described in Section 3.2 and illustrated in Figure 2(b). The procedure identifies informative student errors, uses privileged hindsight evidence to derive reusable corrections, and accepts artifact revisions only after leakage auditing and validation.

A.1 Stage-Specific Failure Diagnosis

At each iteration, the analyzer compares the student rollout with the corresponding stage reference and organizes informative errors into a structured failure package. The analyzer identifies errors but does not use hindsight evidence to explain their underlying causes. Such explanations are produced by the privileged teacher in the subsequent step.

Rather than passing every disagreement to the teacher, the analyzer prioritizes failures that may yield reusable improvements. These include large prediction errors, inconsistencies between judgments and supporting evidence, explicit constraint violations, and errors recurring across multiple items. The resulting failure package is structured according to the learning task.

Evaluation-space assessment. Errors are identified at the proposal and criterion levels. For each selected case, the failure package records the student's predicted score, confidence, supporting comment, evidence-gap flag, and constraint-violation flag, together with the signed error against the reference criterion score. Recurring errors across proposals are also recorded to support the identification of systematic problems.

Population taste prediction. Errors are identified at the pairwise comparison level. Each selected case records the student's predicted preference, confidence, rationale, and reference preference, together with the relevant criterion coordinates and utility margin. When available, results from both comparison orientations are retained to expose orientation inconsistencies. Errors are grouped according to research type and decisive evaluation criteria to identify recurring preference failures.

The resulting package provides the teacher with structured evidence of which judgments failed and where the failures occurred. Determining why they occurred and how to correct them remains the responsibility of the privileged teacher.

A.2 Privileged Teacher Editing

The privileged teacher receives the current artifacts, student rollout, structured failure package, and privileged hindsight information. It examines the selected failures in light of completed methods, experimental outcomes, subsequent literature, citations, and community feedback to identify potential causes of the student's errors.

The teacher then proposes revisions to the agentic artifacts. These revisions may modify evaluation rubrics, introduce reusable memory lessons, or add and revise conditional rules. The teacher does not directly replace the student's judgments with privileged answers.

The key constraint is that hindsight may inform the discovery of a correction, but the resulting artifact must remain executable using proposal-visible information. For example, retrospective evidence may reveal that a scoring rubric systematically undervalues proposals extending an established research direction. The teacher may consequently revise the rubric to distinguish incremental extensions from substantive methodological advances. It may not encode the later experimental success of a particular proposal as an inference-time instruction.

The teacher's private diagnostic explanations are retained for analysis but are not incorporated into student-facing artifacts. Only the proposed artifact revisions are passed to the subsequent audit and validation steps.

A.3 Artifact Leakage Audit

A leakage auditor examines each proposed revision before it can be accepted. The audit is applied to student-facing artifacts, including rubric instructions, memory lessons, conditional predicates, and their associated explanations. It does not constrain the privileged information used internally by the teacher to derive a revision.

The auditor rejects revisions that directly expose retrospective outcomes, citation counts, community responses, reference criterion scores, or preference labels associated with individual proposals. It also rejects references to held-out items and conditional rules that require features unavailable at the proposal stage.

For conditional rules, the auditor additionally checks that their predicates operate on proposal-visible properties and conform to the predefined predicate grammar. Instance-specific rules that effectively identify a single proposal rather than expressing a reusable judgment principle are rejected.

This distinction permits hindsight-derived knowledge to be distilled into general evaluation instructions while preventing privileged instance-level information from entering the student's inference-time artifacts. The leakage constraints and artifact output schemas remain fixed throughout the refinement process.

A.4 Gated Validation and Artifact Acceptance

Passing the leakage audit is necessary but insufficient for accepting a revision. A candidate artifact must also demonstrate improved performance on validation data that did not motivate the edit.

Let Ψ_t denote the current artifacts and Ψ'_t the candidate revision. The student is evaluated under both configurations using the same held-out validation examples. The revision is accepted only when all three conditions are satisfied.

1. The leakage auditor approves all modified student-facing artifacts.
2. The candidate improves the task-specific validation objective.
3. The revision does not introduce material behavioral regressions.

For evaluation-space assessment, validation considers errors against reference criterion scores, with mean absolute error serving as the primary objective. Additional checks examine whether the revision introduces new errors in previously correct judgments.

For population taste prediction, validation uses StrictPair correctness on held-out comparisons. Regression checks examine correct-to-wrong preference flips, orientation inconsistencies, and failures concentrated in particular groups of comparisons.

The validation data exclude examples used to induce the candidate revision. A separate final evaluation set remains untouched by artifact selection and acceptance decisions.

If the candidate passes all acceptance conditions, its artifacts replace the incumbent configuration,

$$\Psi_{t+1} = \begin{cases} \Psi'_t, & \text{if the revision is accepted,} \\ \Psi_t, & \text{otherwise.} \end{cases}$$

Rejected revisions are rolled back, while accepted revisions guide subsequent student rollouts. This procedure aims to retain reusable improvements that generalize beyond the errors that originally motivated the edit.

Evaluation-space assessment. A revision is accepted only when the fixed leakage auditor returns a normalized risk below 0.15, its validation MAE improves by more than 0.01, its correct-to-wrong flip rate does not exceed 5%, and its net repair is positive.

Population taste prediction. Candidate revisions are induced on D_{induce}, fitted on $D_{\text{rule-select}}$, and evaluated for acceptance on $D_{\text{agent-val}}$. The final evaluation set D_{final} remains untouched.

Acceptance requires an improvement in StrictPair, a McNemar test with $p < 0.05$, and a paired bootstrap 95% confidence interval for the StrictPair difference that excludes zero.

Rule specificity. The leakage auditor rejects rules whose triggering behavior is concentrated on a single paper or cluster, according to the configured concentration threshold.

B Evaluation Space Learning

This section provides the complete criterion-annotation and proposal-time assessment instructions used to learn the shared evaluation space, together with a concrete example of how Privileged Hindsight Distillation (PHD) transforms teacher-side supervision into deployable artifacts. We first define the reference annotation rubric, then give the seed proposal-time rubric, and finally illustrate how PHD revises the rubric and student-safe memory on a held-out proposal.

B.1 Reference Criterion Annotation

Rubric A defines the reference annotation procedure used to construct

$$s^*(P) = \left(s^*_{\text{gen}}, s^*_{\text{nov}}, s^*_{\text{imp}}, s^*_{\text{par}}, s^*_{\text{dep}}, s^*_{\text{comp}}\right) \in \{0,1,2,3\}^6.$$

The six coordinates correspond to generality, novelty, importance, parsimony, depth, and algorithmic complexity. Each coordinate is assigned independently on an absolute 0–3 scale.

Unlike the deployable proposal-time assessor, Rubric A may use both proposal-visible evidence and privileged hindsight evidence. Proposal-visible evidence includes the proposal text, its explicit claims and assumptions, and literature available at proposal time. Privileged evidence includes the completed methodology and experimental findings in the corresponding paper, post-publication technical discussion, and subsequent community signals such as citations. The annotator must connect each assigned score to criterion-specific evidence rather than infer one criterion from writing quality, venue, author identity, or overall paper reputation.

Rubric A: Reference Criterion Annotation

Input. A research proposal, proposal-time prior literature, and teacher-side privileged hindsight evidence.

Evidence rule. Use all available evidence that is relevant to the target criterion. Every assigned score must be justified by criterion-specific evidence. If multiple sources disagree, record the disagreement and select the score best supported by the overall evidence. Reduce confidence when evidence is incomplete, assumptions are not directly comparable, or adjacent score levels remain similarly plausible.

Generality

Score 0. The method applies only to one special setting, and its scope or applicable regime is missing or merely implied.

Score 1. The method covers multiple instances or variants within one problem family, with limited variation in access, noise, or hardware assumptions.

Score 2. The method covers multiple problem, input, or noise settings within one broader class, with explicit applicability assumptions.

Score 3. The method covers multiple problem classes or distinct access, error, or hardware models, with explicit applicability conditions.

Novelty

Score 0. The contribution is substantially the same as identified prior work, with no meaningful structural difference.

Score 1. The contribution is a nearby transfer, implementation variation, or minor modification of a known method.

Score 2. The contribution forms a new workflow, formulation, or structured extension by combining known mechanisms and clearly distinguishes itself from prior work.

Score 3. The contribution introduces a new primitive, reduction, encoding, or formal object absent from the relevant prior-work lineage.

Importance

Score 0. There is no clear evidence of field demand; the motivation is absent, generic, or primarily self-asserted.

Score 1. The work addresses a useful but local subproblem, with limited evidence of broader impact.

Score 2. The addressed limitation is repeatedly identified as important across multiple works, settings, or application areas.

Score 3. An authoritative survey, roadmap, benchmark, or broad body of evidence identifies the problem as a central open problem or foundational bottleneck.

Parsimony

Score 0. The method requires many layers, assumptions, special cases, or ad hoc components for its stated functionality.

Score 1. The method is a moderately complex pipeline with several distinct construction or execution stages.

Score 2. A small number of coordinated core subroutines, oracles, or assumptions achieves the stated objective.

Score 3. A unified primitive, representation, or objective replaces multiple otherwise separate or ad hoc components.

Depth

Score 0. The work addresses a single setting and provides little explanation of the underlying mechanism.

Score 1. The work remains within one task family and provides mainly intuition or partial technical detail, with central analysis deferred.

Score 2. The work unifies multiple tasks or primitives, or provides concrete analytical, experimental, or diagnostic grounding of the mechanism.

Score 3. The work develops a transferable formalism supported by theorem-level, lower-bound, causal, or comparably strong mechanistic evidence.

Algorithmic Complexity

Score 0. The work makes no meaningful complexity or resource claim relative to an appropriate baseline.

Score 1. The work makes an informal efficiency claim without directly comparable quantitative evidence.

Score 2. The work presents an explicit resource improvement relative to a domain baseline under stated assumptions and resource metrics.

Score 3. The work establishes a theoretical bound or strong Pareto improvement while reasonably accounting for state preparation, oracle access, error dependence, measurement, and readout costs.

Criterion-specific evidence requirements. Generality requires evidence about covered problem classes, access models, hardware regimes, and the assumptions under which the claimed coverage holds. Novelty is assessed relative to proposal-time prior work across problem formulation, quantum primitive, reduction or encoding, complexity model, measurement or verification strategy, and assumption set. Importance requires external evidence that the target problem is recognized beyond the proposal's own motivation. Parsimony is evaluated from the complete method structure, including necessary subroutines, oracles, ancillas, measurements, post-selection steps, and special-case handling. Depth jointly measures the breadth of unification and the strength of mechanistic grounding. Algorithmic complexity compares the claimed and realized resource behavior with appropriate baselines while accounting for hidden computational costs.

B.2 Seed Proposal-Time Rubric

The deployable criterion assessor $\hat{\Pi}$ uses the same six-dimensional evaluation space and the same absolute 0–3 scale as the reference annotation, producing

$$\hat{s}(P) = (\hat{s}_{\mathrm{gen}}, \hat{s}_{\mathrm{nov}}, \hat{s}_{\mathrm{imp}}, \hat{s}_{\mathrm{par}}, \hat{s}_{\mathrm{dep}}, \hat{s}_{\mathrm{comp}}) .$$

Its information boundary, however, is different. At deployment, the assessor can access only the proposal and literature available at evaluation time. Completed outcomes, post-publication discussion, subsequent citations, and other hindsight information are unavailable.

We therefore initialize the student with the proposal-visible instruction shown below, which we refer to as *seed Rubric B*. The seed rubric is an initial operationalization of the six evaluation criteria rather than a fixed definition of how the criteria must be inferred. PHD subsequently refines its score boundaries and evidence-use rules while preserving the same proposal-time information boundary.

Seed Rubric B: Proposal-Time Criterion Assessment

Input. A research proposal and literature available at proposal-evaluation time.

Evidence constraint. Use only information available when the proposal is evaluated. Permitted evidence includes proposal claims, equations, assumptions, methodological descriptions, planned comparisons, named prior work, and proposal-time literature. Do not use completed

experimental outcomes, future reception, post-publication discussion, subsequent citations, or other hindsight information.

Scoring rule. Score each criterion independently on an absolute 0–3 scale. Do not infer one criterion from another, and do not reward writing style or perceived author reputation. If the available evidence is insufficient to distinguish adjacent levels, mark the evidence gap rather than silently assuming the stronger claim.

Generality

Score 0. The proposal addresses one special setting; its scope or applicable regimes are missing or merely suggested.

Score 1. The proposal explicitly covers multiple instances within one problem family, but provides little cross-setting variation.

Score 2. The proposal covers multiple problem, input, or noise settings within one class and states the relevant assumptions.

Score 3. The proposal covers multiple problem classes or distinct access, error, or hardware models and states their applicability conditions.

Novelty

Score 0. The proposed method is substantially the same as named prior work, with no structural difference stated.

Score 1. The proposal describes a nearby transfer or minor modification of a known technique or formulation.

Score 2. The proposal defines a new workflow by combining known mechanisms and states a concrete difference from named prior work.

Score 3. The proposal introduces a new primitive, reduction, encoding, or formal object that is absent from the prior-work lineage described in the proposal.

Importance

Score 0. The proposal provides no clear evidence of field demand; its motivation is weak, generic, or absent.

Score 1. The proposal addresses a useful local subproblem but gives limited support for broader impact.

Score 2. The proposal presents the limitation as repeatedly important across multiple settings, applications, or prior works.

Score 3. The proposal supports the claim that the problem is a central open problem or foundational bottleneck for a broad research area.

Parsimony

Score 0. The proposed method is a multi-layered pipeline with many assumptions, special cases, or ad hoc components.

Score 1. The proposed method is moderately complex and contains several distinct construction or execution stages.

Score 2. A small number of core subroutines or oracles achieves the stated objective with limited additional machinery.

Score 3. A unified primitive, representation, or objective replaces multiple otherwise separate components.

Depth

Score 0. The proposal addresses a single setting and gives little or no explanation of the underlying mechanism.

Score 1. The proposal stays within one task family and provides intuition or partial detail, while its main proofs or diagnostics remain deferred.

Score 2. The proposal covers multiple tasks or primitives, or gives concrete lemma, error-analysis, diagnostic, or ablation plans that ground the mechanism.

Score 3. The proposal develops a transferable formalism together with explicit plans for theorem-level, lower-bound, causal, or comparably strong evidence.

Algorithmic Complexity

Score 0. The proposal makes no complexity or resource claim relative to any baseline.

Score 1. The proposal makes an informal efficiency claim without comparable quantitative evidence.

Score 2. The proposal states an explicit improvement relative to a domain baseline, with resource metrics and assumptions.

Score 3. The proposal states a theoretical bound or strong Pareto improvement and reasonably accounts for hidden state-preparation, oracle, error, measurement, and readout costs.

Required structured output. For every evaluation criterion d, return

$$\hat{s}_d(P) = (\text{score}, \text{confidence}, \text{supporting comment}, \text{evidence-gap indicator}, \text{constraint-violation flag}) .$$

The supporting comment must identify proposal-visible evidence for the score. The evidence-gap indicator is activated when the proposal does not provide enough information to distinguish adjacent score levels. The constraint-violation flag records a direct conflict between an assigned score and its anchor, such as assigning maximal algorithmic complexity when no resource claim is present.

B.3 Learning Rubric B and Student-Safe Case Memory

During training, the privileged teacher observes the student's assessment $\hat{s}(P)$, the Rubric-A reference $s^*(P)$, and the supporting teacher-side evidence. Rather than transferring privileged facts into the deployable workflow, the teacher diagnoses recurring proposal-time reasoning errors and converts them into reusable student-safe supervision.

We update two complementary student-side artifacts. Systematic errors that expose ambiguity in score boundaries are translated into revisions of Rubric B. Errors that are more useful as concrete precedents are stored in a separate student-safe case memory. In both cases, hindsight-dependent facts are removed before the artifact becomes accessible to the student.

Rubric revisions. Accepted rubric updates refine evidence criteria and adjacent-score boundaries. Representative learned principles include the following:

- several benchmark models, parameter regimes, or subcases within one task family do not automatically establish broad generality;
- a new workflow, robustness regime, or theorem-level extension built from known primitives generally supports intermediate novelty rather than maximal novelty;
- multiple subroutines do not necessarily imply low parsimony when the subroutines directly support one coherent mechanism;
- explicit query, gate, depth, qubit, or error analysis can support an intermediate algorithmic-complexity score even when the strongest form of resource advantage has not been established.

Case memory. Each memory entry records a proposal-visible trigger, the affected evaluation criterion, a relevant score region, and a reusable scoring lesson. Teacher-only evidence, reference outcomes, and information unavailable at proposal time are removed. At inference time, retrieved entries therefore serve as calibration precedents rather than as additional facts about the target proposal.

Candidate rubric and memory states are evaluated on the validation split using criterion-level mean absolute error,

$$\mathcal{L}_{\text{eval}} = \frac{1}{6|\mathcal{D}_{\text{val}}|} \sum_{P \in \mathcal{D}_{\text{val}}} \sum_{d=1}^{6} |\hat{s}_d(P) - s_d^*(P)| .$$

A candidate state is retained only when it satisfies the validation gate. Otherwise, the student is rolled back to the incumbent rubric and memory. Student-facing artifacts are additionally screened by a pattern-based leakage check before being admitted into the deployable workflow.

Example Learned Rubric-B Revision

Affected criterion Novelty

Observed failure pattern The student repeatedly treats a new arrangement of established mechanisms as evidence of a fundamentally new primitive.

Accepted calibration rule A new workflow, formal threshold, robustness regime, or structured extension of an existing method can justify novelty 2. Novelty 3 should be reserved for a new primitive, reduction, encoding, or formal object not already present in the relevant prior lineage.

Example Student-Safe Memory Entry

Dimension Parsimony

Trigger A proposal contains several proof or implementation components that all support the same central mechanism.

Calibration lesson Parsimony is not only about having few steps; it is about whether those steps perform separate jobs or jointly support one central mechanism. A proposal with several proof or implementation components can still receive an intermediate parsimony score when the components are coordinated around the same objective.

Excluded information Completed experimental outcomes, teacher-side hindsight evidence, the Rubric-A reference outcome, and any information unavailable at proposal time.

B.4 Evaluation-Space Learning Case Study

We illustrate the learning process on the held-out proposal *Quantum Hamiltonian Descent* (arXiv 2303.01471). The held-out proposal is used only for analysis. Rubric revisions and memory entries are learned from analogous failure patterns in the training split rather than from this example itself. Scores are reported in the order generality, novelty, importance, parsimony, depth, and algorithmic complexity. The progress is shown in Table 8

Reference Annotation

Rubric A assigns the proposal the reference criterion vector

$$s^*(P) = (3, 2, 2, 2, 2, 2).$$

The high generality score reflects a common Hamiltonian-descent framework across smooth and non-convex optimization, digital fault-tolerant simulation, and analog Ising-machine implementations under distinct assumptions. Novelty is intermediate because the proposal introduces a distinctive workflow based on quantizing continuous-time optimization dynamics while relying on known components such as Hamiltonian simulation, product formulas, quantum Fourier transforms, and Ising embeddings. Parsimony is also intermediate because these components support one central mechanism despite a nontrivial implementation pipeline. Algorithmic complexity receives an intermediate score because the proposal states explicit resource quantities and baseline comparisons without establishing the strongest level of resource advantage with every hidden cost resolved.

Assessment Before PHD Learning

Using the seed Rubric B without learned case memory, the proposal-time student predicts

$$\hat{s}_{\text{seed}}(P) = (3, 3, 2, 0, 2, 1).$$

Three characteristic errors are visible. First, the student interprets the Hamiltonian formulation as a wholly new primitive and assigns novelty 3 instead of 2. Second, it counts implementation components without recognizing their shared conceptual role and assigns parsimony 0 instead

of 2. Third, it treats the absence of a fully established quantum advantage as evidence of little algorithmic-complexity contribution and assigns 1 instead of 2.

How the teacher turns training failures into student artifacts. On training proposals, the privileged teacher identifies analogous recurring errors. For example, when the student treats multiple prime families, decoding regimes, or circuit variants within one task as evidence of broad generality, the teacher converts the diagnosis into a proposal-visible boundary rule: multiple instances or noise regimes within one task family do not by themselves justify a high generality score. Similarly, repeated confusion between a new combination of known mechanisms and a genuinely new primitive produces the novelty calibration shown above. For parsimony, the teacher learns to distinguish raw component count from functional coherence and stores the more concrete precedent in case memory.

Assessment After PHD Learning

With the validation-selected Rubric B and case memory, the student predicts

$$\hat{s}_{\text{learned}}(P) = (2, 2, 2, 1, 2, 1).$$

The learned novelty boundary corrects the novelty score from 3 to 2. The parsimony prediction moves from 0 to 1, reducing but not eliminating the error. Algorithmic complexity remains underestimated. The generality score moves from the correct seed value of 3 to 2, illustrating that a transferable calibration rule can still be over-applied to an individual proposal.
The proposal-level MAE decreases from 0.667 with the seed Rubric B to 0.500 after learning. Across the held-out test split, absolute-score MAE decreases from 0.216 for the w/o PHD student to 0.183 for the validation-selected PHD student state. This aggregate comparison evaluates the revised rubric and memory jointly. It therefore measures improvement of the learned student state rather than isolating the causal contribution of the rubric revision or the case memory individually.

C Population Taste Distillation

C.1 Hindsight Reference Panel

We construct population-level reference supervision using an independent multi-family LLM panel. The panel is used only to construct offline reference judgments for training and evaluation. Unlike the deployable student, panel judges may access privileged hindsight evidence in addition to proposal-time information. These privileged inputs are never exposed to the student during deployment.

Panel prompt. For every proposal pair (P_A, P_B), each judge independently receives the same normalized evidence bundle together with the following instruction.

Hindsight Reference Panel Prompt

You are an offline reference judge comparing two research proposals to construct a population-level preference reference. You receive (i) the complete text of proposals A and B, (ii) reference assessments along the six evaluation criteria, (iii) source-paper and proposal-time evidence supporting those assessments, and (iv) privileged evidence about subsequent community response, including citation function and polarity, methodological adoption, and reuse.
Using the complete evidence bundle, determine which proposal is preferable as a research direction at the population level. Consider the criterion-level assessments and community-response evidence jointly, and explain any disagreement between them. Your judgment must not depend on presentation order or proposal length. You must select exactly one winner, either A or B. Ties and uncertain outcomes are not permitted.
Return a JSON object containing the selected winner, the estimated probability that A is preferred to B at the population level, confidence in the judgment, a short rationale, the decisive factors, an explanation of how the criterion-level and community-response evidence agree or conflict, and whether a proposal-only reader could plausibly reach the same conclusion.

Dimension	Reference	Seed	Learned
Generality	3	3	2
Novelty	2	3	2
Importance	2	2	2
Parsimony	2	0	1
Depth	2	2	2
Algorithmic complexity	2	1	1
MAE	–	0.667	0.500

Table 8: **Evaluation-space learning on a held-out proposal.** Rubric A provides the reference coordinates. Seed and Learned show the proposal-time predictions before and after PHD learning.

Evidence source	Content	Student-visible
Proposal text	Complete proposal claims, assumptions, equations, and methodology	Yes
Proposal-time literature	Prior work and external evidence available when the proposal is evaluated	Yes
Criterion annotation	Six-dimensional Rubric-A assessment and its supporting evidence	No
Completed work	Source-paper methodology, experimental findings, and realized analyses	No
Community response	Follow-up citations, scites in SciRate, citation function and polarity, adoption distance, methodological integration, and primitive reuse	No

Table 9: **Evidence available to the hindsight reference panel.** Proposal-visible sources may also be accessed by the deployable workflow, whereas privileged sources remain teacher-side information.

Evidence bundle. All judges receive the same canonical A/B orientation and the same normalized evidence bundle. The bundle combines proposal-visible information with teacher-side hindsight evidence as summarized below.

Proposal texts are truncated only when they exceed the fixed context budget. No model-specific evidence preprocessing is applied, so all panel members judge the same logical pair from the same evidence state.

Independent annotation. Three model families independently annotate every proposal pair at temperature zero, namely GPT-5.5, Claude Opus 5, and Gemini 3.6 Flash. Judges cannot observe one another's outputs. Each judge returns a hard winner

$$y_j \in \{A, B\},$$

together with an estimated probability

$$q_j = P_j(A \succ B),$$

confidence, rationale, decisive factors, and an explanation of how the criterion-level and hindsight evidence support the decision.

Single-judge normalization. Each response is parsed against the required output schema. The reported probability is clipped to $[0, 1]$ and normalized to remain directionally consistent with the hard winner,

$$y_j = A \Longrightarrow q_j \geq 0.51, \qquad y_j = B \Longrightarrow q_j \leq 0.49.$$

A valid normalized annotation therefore cannot assign $q_j = 0.5$. Malformed responses are treated as annotation failures and regenerated or excluded rather than converted into tie labels.

Cross-family aggregation. The population reference label is determined only by the three hard winner labels,

$$y_{\text{panel}}(A, B) = \text{MajorityVote}(y_1, y_2, y_3), \qquad y_j \in \{A, B\}.$$

Because the panel contains three judges and every valid judge selects exactly one proposal, majority voting always produces a unique binary preference. Per-model probabilities, confidence values, rationales, and decisive factors are retained as privileged teacher-side metadata but do not alter the final reference label.

Reference Construction Summary

For each proposal pair, all three model families independently inspect the same hindsight evidence bundle. Each model produces one binary preference. The majority vote forms $y_{\text{panel}}(A, B)$, which provides the population-level reference supervision used during Population Taste Distillation. Rich judge outputs are retained only as teacher-side diagnostic information.

C.2 Pairwise Instantiation of Population Taste

Section 3.4 defines population taste generally as a mapping from features in the shared evaluation space to research preference. In our experiments, scientific taste is observed through pairwise proposal comparison. We therefore instantiate the general mapping using the relative evaluation-space features of the two research directions.

For a proposal pair (A, B), let $\hat{s}(A), \hat{s}(B) \in \mathbb{R}^6$ denote their estimated positions in the shared evaluation space and define the pairwise criterion difference as

$$x_{A,B} = \hat{s}(A) - \hat{s}(B).$$

The linear population prior uses $w_{\text{pop}}^\top x_{A,B}$, where w_{pop} captures population-level weighting of the six evaluation criteria. PHD augments this prior with proposal-visible conditional rules, giving the pairwise instantiation of the population mapping from Section 3.4,

$$u_{\text{pop}}(x_{A,B}, \mathcal{E}_{A,B}) = w_{\text{pop}}^\top x_{A,B} + \sum_{r \in R_{\text{pop}}} \tilde{\alpha}_r \, \phi_r(x_{A,B}, \mathcal{E}_{A,B}),$$

where $\mathcal{E}_{A,B}$ contains proposal-visible evidence for the pair, ϕ_r denotes the activation of rule r, and $\tilde{\alpha}_r$ is its reliability-adjusted contribution. The induced preference probability is $p_{\text{pop}}(A \succ B) = \sigma(u_{\text{pop}}(x_{A,B}, \mathcal{E}_{A,B}))$.

From General Taste Mapping to Pairwise Supervision

The underlying object being learned is a population mapping over evaluation-space features. Pairwise comparison is the observation interface used in our task rather than the definition of population taste itself. Accordingly, the general feature representation $\hat{s}(P)$ becomes the relative representation $x_{A,B}$ when the mapping is trained and evaluated against the panel preference $y_{\text{panel}}(A, B)$.

D Individual Taste Adaptation Details

Individual Taste Adaptation freezes the learned evaluation-space assessor and population state and updates only the current researcher's personal state from revealed pairwise feedback. No privileged hindsight evidence, hidden researcher profile, or held-out query label is available during adaptation. The implementation first fits a population-centered residual Δw_u by MAP logistic regression, then considers conditional personal rules only when recurring errors remain after continuous criterion reweighting. Accepted rule updates are checked by replay over the accumulated support buffer before the personal state is retained for evaluation.

Ablation variants. Cold Start uses the frozen population mapping without researcher-specific updates. Residual-only fits Δw_u while keeping the researcher-specific rule set empty. Complete uses both Δw_u and accepted researcher-specific conditional rules. The w/o PHD variant removes PHD from the shared initialization and runs the same downstream individual-adaptation procedure starting from the seed Rubric B and a randomly initialized population prior, while keeping the backbone, support budget, replay procedure, and adaptation mechanism unchanged.

D.1 Replay-Based Rule Acceptance

Candidate personal rules are evaluated over the accumulated support buffer using leave-one-out cross-fitted replay. For each support comparison i, we refit the individual residual Δw_u and the candidate rule coefficient α_r on the remaining $n-1$ comparisons and evaluate the held-out comparison. The resulting out-of-fold predictions are used to compute three rule effects: ***fixed***, the number of previously incorrect StrictPair predictions corrected by the rule; ***broken***, the number of previously correct predictions made incorrect; and

$$\text{net} = \text{fixed} - \text{broken}.$$

We additionally compute the change in negative log-likelihood, $\Delta_{\text{NLL}} = \text{NLL}_{\text{before}} - \text{NLL}_{\text{after}}$.

Under the default replay gate, a candidate rule is accepted only if (i) its fitted coefficient is nonzero, (ii) it activates on at least $\max(4, \text{round}(k/3))$ observed comparisons, (iii) $\text{net} \geq 1$, (iv) $\text{broken} = 0$, and (v) $\Delta_{\text{NLL}} > 0$. The deployed rule coefficient is obtained by aggregating the out-of-fold fitted coefficients.

The replay ablation keeps the same rule proposer, coefficient fitting, support sequence, and downstream evaluation, but bypasses conditions (ii)–(v); only the nonzero-coefficient check is retained. Thus, the ablation isolates the effect of replay-based rule acceptance rather than removing individual residual fitting or altering RW-SP.

D.2 A Case Study of Divergent Individual Taste

We examine two researchers, U04 and U05, who evaluate the same ten proposal pairs using identical frozen criterion features. Five pairs elicit opposing preferences, providing a direct test of whether personalized mappings can recover conflicting choices rather than reproducing a shared population prediction.

Table 10 shows that the cold-start mapping fails to recover both preferences on any of the five conflicting pairs. Residual-only adaptation recovers one, while the complete mapping recovers three. The improvement is particularly visible for U04, whose accuracy increases from 4/10 to 8/10. U05's accuracy, however, decreases from 7/10 at cold start to 6/10 after adaptation. This example illustrates how conditional individual rules can recover opposing preferences that residual weighting misses, while also showing that improved differentiation does not guarantee higher accuracy for every researcher.

E Data Construction and Evaluation Protocols

Source pool. We construct the shared proposal corpus from SciRate under the search query `quantum algorithm`. After removing noise clusters and requiring clusters sufficiently rich for controlled comparison, we stratify the retained papers by topical cluster, release time, and scites. From this filtered pool, we curate two complementary sources: a broad SciRate subset for topical coverage and a classic lookup set containing historically salient quantum-algorithm papers used as a diagnostic core.

Proposal reconstruction. Each retained paper is converted into a proposal-time document. We preserve the problem statement, proposed method, evaluation plan, expected contributions, and risks, while rewriting results and other post-hoc outcomes as objectives or hypotheses. Reconstruction uses full-text extraction followed by an LLM rewrite under a fixed paper-to-proposal prompt, so that downstream judges receive only information that could plausibly have been stated before the work unfolded. Controlled strengthen, weaken, and revision variants used elsewhere as comparator training signals are excluded from the Evaluation Space: absolute criterion assessment is defined only on canonical paper-derived proposals, rather than synthetic variants of the same underlying work.

Eligibility and paper-disjoint splits. Both Evaluation Space Learning and Population Taste Distillation use carefully curated, paper-disjoint splits. Constructing these evaluations requires controlled proposal-time reconstruction and privileged reference construction. A retained proposal

Table 10: **An illustrative case of divergent scientific taste.** Two researchers evaluate the same ten proposal pairs using identical criterion features. Predictions are shown as U04/U05. Bold indicates a correct individual prediction. A star indicates that both opposing preferences are correctly recovered. Proposal titles are abbreviated for readability.

Pair	Research directions (A / B)	Gold	Cold	Residual	Full
Opposing preferences between U04 and U05					
005/008	Eigenvalue processing / Non-unitary dynamics	B/A	A/**A**	**B**/B	**B**/B
009/040	Optimal quantum simulation / Fourier-based LCHS	A/B	**A**/A	⋆**A/B**	⋆**A/B**
009/045	Optimal quantum simulation / Circuit-efficient randomized methods	A/B	B/**B**	B/**B**	⋆**A/B**
013/029	Linear matrix differential equations / Gibbs expectation	B/A	A/**A**	A/**A**	⋆**B/A**
027/034	Quantum Hamiltonian Descent / Provably efficient algorithm	B/A	A/**A**	A/**A**	A/**A**
Aligned preferences between U04 and U05					
013/023	Linear matrix differential equations / Sign-embedding algorithms	B/B	A/A	A/**B**	**B/B**
028/045	Quantum Hermite transform / Circuit-efficient randomized methods	A/A	**A/A**	B/B	**A**/B
028/049	Quantum Hermite transform / End-to-end quantum simulation	A/A	**A/A**	**A**/B	**A**/B
029/036	Gibbs expectation / Efficient quantum simulation	A/A	B/B	B/B	B/B
039/040	BELT block encoding / Fourier-based LCHS	A/A	**A/A**	**A/A**	**A/A**
Correct predictions (U04 / U05)		–	4/10 / 7/10	4/10 / 6/10	**8/10** / 6/10
Opposing-choice recovery		–	0/5	1/5	**3/5**

must therefore be a canonical paper-derived proposal rather than a rewrite variant, admit a single six-dimensional absolute reference under Rubric A, support a frozen privileged-evidence package, and have its underlying paper assigned to exactly one of train, validation, or test. These constraints prevent alternative reconstructions of the same work from being treated as independent examples and ensure that privileged evidence cannot leak across splits. The validation split is used only for artifact-update gating, while the test split remains fixed throughout development.

Population Taste pairs. Population Taste Distillation constructs comparisons within the same frozen proposal universe, with additional constraints on temporal comparability and informativeness. We retain temporally matched comparisons with sufficiently informative differences, while controlling proposal degree and winner imbalance when forming the broader annotation queue. The splits remain paper-disjoint across train, validation, and test: an underlying paper may participate in multiple comparisons within the same split, but never appears in another split, and no proposal pair is duplicated across splits. Privileged population references are obtained from a three-family LLM-as-a-Judge panel that observes the same proposal-visible text together with hindsight evidence. The resulting benchmark is therefore determined jointly by proposal-time reconstructability, privileged-reference availability, cross-split paper isolation, temporal comparability, and pair quality rather than by unconstrained pair-level sampling.

Individual taste sessions. Individual Taste Adaptation operates under a different scarcity regime because its supervision must come directly from researchers. Offline, each of seven researchers annotates 25 proposal pairs in one sitting: 15 sequential support comparisons used for adaptation and 10 held-out query pairs reserved for evaluation, each accompanied by a binary preference and confidence score. The support/query split is defined at the pair level: no query pair is used during adaptation, although the same proposal may appear in both support and query comparisons.

The evaluation therefore measures generalization to unseen preference comparisons rather than to entirely unseen proposals.

The support budget is chosen to reflect the scale of supervision naturally available during interactive use. Across logged deployment sessions, a single sitting yields at most three pool rounds, corresponding to at most six reconstructed competitive preference signals, while even the most active user accumulates at most twelve such signals across sessions. The 15-support setting therefore evaluates adaptation at approximately the same supervision scale encountered in deployment rather than assuming dense preference annotation.

The seven-researcher study is intended to test whether sparse researcher-specific feedback can improve held-out pairwise preference prediction within individuals. It is not used to estimate the population distribution of scientific taste across researchers; population-level conclusions are evaluated separately through the Population Taste benchmark above.

Simulated researcher cohorts. We complement the seven real researchers with 100 simulated researchers under the same 15-support and 10-query protocol. Sim.-Q contains 50 quantum-researcher profiles derived from the publication histories of prominent researchers, while Sim.-Ctrl. contains 50 controlled profiles whose preferences are generated by varying criterion-level trade-offs. Sim.-Ctrl. is designed to isolate preference variation that can be represented by fixed criterion reweighting, whereas Sim.-Q provides broader publication-history-derived preference patterns. Simulated cohorts are reported separately from the real-researcher results and are not treated as human evidence.

Researcher identifiers and privacy. Throughout the evaluation and appendix, real researchers are represented only by pseudonymous identifiers such as U04 and U05. Researcher-specific adaptation uses observed preference feedback rather than identity-derived attributes or hidden profile information.

F Baseline Configurations and Metric Details

Table 11 summarizes the backbone, input representation, and feedback configuration of the evaluated baselines. Unless otherwise specified, API-based baselines use `gpt-5.5` at temperature 0. Published methods retain their released backbones or use the adaptations specified below.

For individual taste evaluation, all methods use the same seven real-researcher sessions, with 15 support comparisons and 10 held-out query pairs per researcher. StrictPair requires correct predictions under both proposal orientations, and individual performance is reported using Regret-Weighted StrictPair (RW-SP).

F.1 Evaluation-Space Baselines

The *SC* (Single Call) baseline directly assesses each proposal using `gpt-5.5` without the learned student workflow.

Prometheus 2 (Kim et al., 2024) uses the released absolute grader `prometheus-eval/prometheus-7b-v2.0`. Rubric B anchors are supplied through the official score-rubric slot. Proposal text is truncated to 12,000 characters. Inference uses greedy decoding with a maximum of 512 new tokens. The original scores from 1 to 5 are mapped to the Rubric B scale as $\{0, 1, 2, 3, 3\}$.

Prometheus 2 + SFT applies LoRA fine-tuning to the same backbone using the criterion reference labels from the training set. We use rank $r = 16$, scaling factor $\alpha = 32$, and target modules q, k, v, and o. Training runs for 5 epochs with learning rate 10^{-4}, effective batch size 16, and maximum sequence length 6144. The checkpoint is selected using validation L_{hard}.

InnoEval (Qiao et al., 2026) uses the released `SingleIdeaPipeline`. Proposal text is supplied as input, and Rubric B is provided through `user_metric`. We use 3 personas and a retrieval window from 2018 to 2026. The backbone is `gpt-5.5`, with agent temperature 0.7.

Table 11: **Baseline configurations.** Backbones, input representations, and target-researcher feedback for the evaluated methods. External population judges are evaluated as alternative priors under the downstream individual-adaptation protocol. Detailed implementation settings are provided below.

Component	Method	Backbone	Input	Support
Evaluation	SC	`gpt-5.5`	Proposal text	N/A
Evaluation	Prometheus 2	`prometheus-7b-v2.0`	Proposal, Rubric B	N/A
Evaluation	Prometheus 2 + SFT	`prometheus-7b-v2.0`	Proposal, Rubric B	N/A
Evaluation	InnoEval	`gpt-5.5`	Proposal, Rubric B	N/A
Population	SC	`gpt-5.5`	Raw proposal pair or criterion features	N/A
Population	Linear Mapping	Learned linear weights	Criterion features	N/A
Alternative prior	Auto-J	`autoj-13b`	Proposal pair	0
Alternative prior	MLR-Judge	`claude-opus-5`, `gemini-3.6-flash`	Independent proposal scores	0
Alternative prior	SciJudge	`SciJudge-4B`	Title, abstract, date	0
Individual	SC	`gpt-5.5`	Proposal pair, support comparisons	15
Individual	PP	`gpt-5.5`	Proposal pair, researcher profile	15
Individual	SynthesizeMe	`gpt-5.5`	Proposal pair, persona, demonstrations	15
Individual	VPL	GPT-2 encoder, released VAE	Frozen embeddings	15
Individual	FERMI	`gpt-5.5`	Personalized judge prompts	15
Individual	CIPHER	`gpt-5.5`	Proposal pair, retrieved preference notes	15
Individual	LoRe	GPT-2 encoder, rank-8 basis	Preference-difference embeddings	15
Individual	FSPO	`Qwen2.5-1.5B-Instruct`	Proposal pair, in-context examples	15

For each evaluation criterion, we first average the persona scores to obtain s_{10} on the original 0–10 scale and map it to Rubric B as

$$\hat{s} = \mathrm{clip}_{[0,3]}\left(\mathrm{round}\left(\frac{3s_{10}}{10}\right)\right).$$

We use the default `round` mode, which follows Python's round-to-even convention for half-integer values. The optional `linear` mode omits rounding and applies only $\mathrm{clip}_{[0,3]}(3s_{10}/10)$.

F.2 Population-Taste Baselines

The *SC* baseline directly predicts pairwise preferences using `gpt-5.5`. We evaluate two input configurations, one using raw proposal text and the other using the frozen criterion features $\hat{s}(P)$.

The *Linear Mapping* baseline learns a population-level weight vector over the shared criterion features without conditional taste rules. The student without PHD uses the student workflow without privileged teacher refinement, while the complete method additionally incorporates PHD-refined agentic artifacts. These methods are evaluated against the same selected LLM-panel reference.

F.3 Alternative Population Priors

We evaluate external scientific judges as alternative population priors for downstream adaptation. They are not used to construct the LLM-panel reference for Population Taste Distillation. Their outputs are converted into preference priors and evaluated under the same adaptation protocol.

Auto-J (Li et al., 2024) uses the released `GAIR/autoj-13b` model and its pairwise-tie protocol. We evaluate both proposal orientations using response-order swapping. The model receives full proposals, subject to a tokenizer maximum length of 3500 tokens, and runs in float16 precision.

MLR-Judge (Chen et al., 2026a) scores each proposal independently using an adapted MLR-Bench proposal rubric. The official two-model ensemble is remapped to `claude-opus-5` and `gemini-3.6-flash`. We average the models' `OverallAssessment` scores for each proposal and convert the resulting pairwise difference into a soft preference prior,

$$q_{\mathrm{ref}} = \mathrm{clip}_{[0,1]}\left(0.5 + \frac{\mathrm{score}_A - \mathrm{score}_B}{18}\right),$$

where score_A and score_B denote the ensemble-averaged scores of proposals A and B, respectively.

SciJudge (Tong et al., 2026) uses the released `OpenMOSS-Team/SciJudge-4B`. Each proposal pair is represented using its title, abstract, and date. We evaluate both (A, B) and (B, A) orientations. After reversing the labels from the latter orientation, we obtain two votes v_{ab} and v_{ba} in the original (A, B) orientation, each taking a value in $\{A, B, \varnothing\}$.

We construct the soft preference prior for A as

$$q_{\text{ref}} = \begin{cases} 0.75, & v_{ab} = v_{ba} = A, \\ 0.25, & v_{ab} = v_{ba} = B, \\ 0.65, & \text{only one valid vote, favoring } A, \\ 0.35, & \text{only one valid vote, favoring } B, \\ 0.50, & \text{otherwise.} \end{cases}$$

The final case includes conflicting votes and cases where both votes are missing. These values are assigned by the voting protocol rather than obtained through probability calibration.

Generation uses sampling with `do_sample=True`, temperature 0.7, `top_p=0.8`, and `top_k=20`.

F.4 Individual-Taste Baselines

All personalization methods receive the same 15 support comparisons and are evaluated on the same 10 held-out query pairs. Support memory uses arXiv abstracts, while queries use full proposal text unless otherwise specified.

The *SC* baseline directly predicts pairwise preferences using the support comparisons in the input context. The *PP* (Profile Prompting) baseline first summarizes the researcher using the 15 support annotations and then uses this summary for preference prediction, rather than including all comparisons directly in the query prompt. Both variants use `gpt-5.5` at temperature 0.

SynthesizeMe (Ryan et al., 2025) uses `gpt-5.5` to fit a researcher persona and select up to 4 informative demonstrations from the 15 text-only support preferences. We use a training-to-validation ratio of 0.7 and evaluate 10 search candidates.

VPL (Poddar et al., 2024) trains the released variational autoencoder using frozen GPT-2 last-token embeddings. Training runs for 40 epochs with batch size 8, learning rate 3×10^{-4}, latent dimension 128, and KL weight 10^{-3}.

FERMI (Kim & Yang, 2025) iteratively refines personalized judge prompts using `gpt-5.5`. We run $T = 4$ optimization rounds, generating $K = 3$ new prompts per round. The memory size is 4, and the retrieval procedure selects the top 3 entries. Optimizer sampling uses temperature 0.8, while other API-based inference follows the default temperature of 0.

CIPHER (Gao et al., 2024) induces a short preference note from each support judgment. At query time, it retrieves the top 5 relevant notes to condition pairwise preference prediction.

LoRe (Bose et al., 2025) learns a rank-8 shared preference basis from GPT-2 preference-difference embeddings. We use Adam with learning rate 0.05, 2000 joint-training steps, and 500 few-shot adaptation steps. Researcher-specific weights w_u are fitted using the 15 support comparisons.

FSPO (Singh et al., 2025) adapts `Qwen2.5-1.5B-Instruct` using LoRA and IPO on leave-one-support-out contexts. We use LoRA rank $r = 8$, 2 training epochs, learning rate 10^{-4}, IPO coefficient $\beta = 0.1$, and 4 in-context demonstrations. Candidate preferences are scored using length-normalized continuation log-probabilities.

F.5 Confidence Mapping for RW-SP

RW-SP uses an orientation-specific preference probability rather than the raw choice-relative confidence annotation. Let $c_i \in [0, 1]$ denote the researcher's confidence in the selected direction for comparison (A_i, B_i). We convert it to

$$p_i = P(A_i \succ B_i) = \begin{cases} c_i, & \text{if the researcher selects } A_i, \\ 1 - c_i, & \text{if the researcher selects } B_i. \end{cases}$$

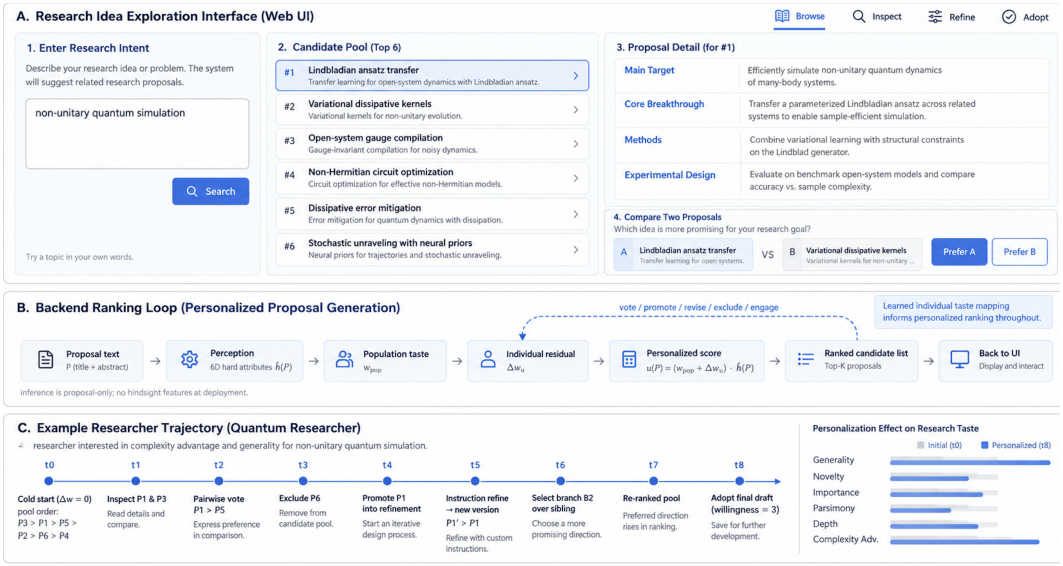

Figure 3: **Interactive proposal exploration system.** The illustration presents the web interface, its main functionalities, and an example of how researchers explore candidate directions and provide feedback. The learned individual taste mapping informs personalized proposal ranking throughout the interaction process.

The corresponding regret weight is

$$d_i = 2|p_i - 0.5|.$$

Thus, $p_i = 0.5$ receives zero weight, while a deterministic preference $p_i \in \{0, 1\}$ receives unit weight.

When multiple persona judgments are aggregated, each choice-relative confidence is first converted to an orientation-specific probability using the same rule, and the probabilities are then averaged:

$$q_{\text{user}} = \frac{1}{m} \sum_{j=1}^{m} p_{ij}.$$

The aggregate confidence stored by the data pipeline, $|2q_{\text{user}} - 1|$, is a preference-strength statistic and is not reused as p_i in RW-SP.

If an orientation-specific probability is unavailable, we fall back to the hard preference label, corresponding to $p_i \in \{0, 1\}$. If all comparisons have zero regret weight, RW-SP falls back to the unweighted StrictPair mean. The probabilities p_i are used only for evaluation weighting. Individual residual and rule fitting use the observed binary preference labels rather than p_i as soft training targets.

G Interactive Proposal Exploration and Deployment Evaluation

System interface and interaction workflow. We integrate the learned individual taste mapping into an interactive web system for exploring research directions. The system supports proposal exploration and preference feedback, allowing researchers to interact with candidate directions and refine their interests through successive interactions. The learned taste mapping is used to personalize proposal rankings based on the shared evaluation-space representation and the researcher's accumulated feedback.

Figure 3 illustrates the system interface, its main functionalities, and an example of the user interaction process. The interface provides a practical setting in which preference learning affects which research directions are presented to researchers, rather than being evaluated solely through offline pairwise predictions.

Researcher cohorts and evaluation protocol. We evaluate the system with 20 real researchers from multiple disciplines, including quantum research, computer science, statistics, and medicine. Their interaction trajectories provide behavioral evidence of which displayed proposals they choose to inspect, refine, or pursue.

To complement the real-user evaluation, we conduct offline exploration simulations with the same 100 simulated researcher profiles described in Appendix E: 50 Sim.-Q profiles derived from quantum-researcher publication histories and 50 Sim.-Ctrl. profiles generated from controlled criterion trade-offs. The simulated cohorts broaden the range of preference configurations, and we report them separately rather than treating simulated engagement as equivalent to observed human behavior.

At each interaction round, both ranking policies receive the same candidate proposal pool, proposal-visible information, and accumulated user feedback. Our personalized policy incorporates this feedback through the learned individual taste mapping, whereas the SC baseline uses the same `gpt-5.5` backbone and directly conditions on the complete interaction history in its prompt and infers researcher-specific preferences in context. Both policies use the same base model. This controlled setup holds the candidate pool, backbone, and available interaction evidence fixed, allowing the comparison to test whether explicitly learning an individual taste mapping provides benefits beyond direct in-context personalization.

Interaction history and supervision scale. Across logged deployment sessions, a single sitting yields at most three pool rounds, corresponding to at most six reconstructed competitive preference signals, while even the most active user accumulates at most twelve such signals across sessions. The complete interaction history therefore remains compact enough to be included in every `SC` prompt without truncation. These trajectories also motivate the 15-support budget used in the offline individual-taste evaluation in Appendix E, which operates at approximately the same scale of researcher supervision.

Source-blind competitive-pair interleaving. We adapt competitive-pair team-draft interleaving (Radlinski et al., 2008; Zhang et al., 2025) to compare the two ranking policies within a shared display.

Proposals selected by both policies are displayed only once and receive no policy-specific credit. When the policies disagree, their selections are placed together as adjacent competitive pairs. Presentation order within each pair is randomized to reduce systematic position effects. Researchers are not informed which ranking policy contributed each proposal.

This design focuses the evaluation on disagreements between the two ranking policies. Rather than comparing engagement across separately presented ranked lists, it observes user responses to competing proposals under a shared candidate pool and display.

Engagement signals and evaluation metrics. We assign each displayed proposal an engagement score $e(P) \in \{0, 1, 2, 3\}$ based on the observed interaction: 0 indicates no substantive engagement, 1 inspection, 2 editing or refinement, and 3 adoption.

Our primary metric is Competitive-Pair Engagement Gain, which measures the average engagement difference between proposals contributed by the two policies within competitive pairs,

$$\Delta_{\mathrm{CP}} = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \left(e_c^{\mathrm{ours}} - e_c^{\mathrm{base}} \right),$$

where $\mathcal{C}$ denotes the set of competitive pairs. A positive value indicates greater observed engagement with proposals contributed by the personalized policy.

We additionally report Win Rate over non-tied competitive pairs. A policy wins a pair when its contributed proposal receives a higher engagement score than the competing proposal.

To complement these pairwise comparisons, we evaluate the reconstructed rankings using NDCG and MRR. NDCG treats engagement scores as graded relevance, while MRR considers proposals with $e(P) \geq 2$ as substantive follow-up. These metrics examine whether proposals associated with greater engagement appear earlier in each policy's reconstructed ranking.

Since engagement is observed under the shared interleaved display, the reconstructed ranking metrics are interpreted descriptively. Competitive-pair outcomes remain the primary comparison between ranking policies.

Finally, engagement reflects researchers' immediate interaction with candidate directions rather than the eventual scientific outcomes of pursuing them. The deployment evaluation therefore measures usefulness for proposal exploration, not long-term scientific impact.

H Agentic Workflow and Prompt Specifications

This appendix specifies the language-model agents and deterministic operators used to learn the three components of our framework. We follow the terminology of the main text throughout. In particular, $\hat{s}(P)$ denotes the criterion vector of research direction P, equivalently its position in the shared evaluation space. The six coordinates correspond to generality, novelty, importance, parsimony, depth, and algorithmic complexity.

Rubric A provides hindsight-informed reference criterion assessments, whereas Rubric B is the proposal-visible instrument available to the deployable criterion assessor. The construction of the multi-family hindsight reference panel is described separately in Appendix C.1, and its pairwise instantiation is given in Appendix C.2.

All language-model calls use temperature zero unless otherwise stated. Dynamic inputs are shown in square brackets. Retrieval, regression, orientation swapping, replay, clustering, predicate parsing, reliability estimation, validation, and acceptance decisions are deterministic operations rather than language-model agents.

H.1 Evaluation Space Learning

Evaluation Space Learning estimates

$$\widehat{\Pi}(P) = \hat{s}(P) \in [0,3]^6$$

from proposal-visible information. Its learned student-side artifacts are the deployable Rubric B and a student-safe case memory. The complete specifications of Rubric A and Rubric B are provided in Appendix B.1 and Appendix B.2.

H.1.1 Proposal-Time Criterion Assessor

The deployable criterion assessor is the only language-model call in the forward perception path. Before each assessment, the system deterministically retrieves the k nearest student-safe memory entries and places them in the prompt together with the current Rubric B.

Proposal-Time Criterion Assessor

System
You evaluate one research proposal independently on an absolute 0–3 scale along six evaluation criteria: generality, novelty, importance, parsimony, depth, and algorithmic complexity.
There is no competing proposal and no pairwise preference task. Use only proposal-visible information, the supplied evaluation rubric, and the retrieved student-safe memory. Do not use publication outcomes, citations, adoption, community response, or any information that would only become available after the proposal stage.
Return exactly one JSON object.

User
Proposal identifier: [PROPOSAL ID]
[RETRIEVED STUDENT-SAFE CASE MEMORY]
[DEPLOYMENT RUBRIC B]
[PROPOSAL TEXT]
For each evaluation criterion, return: `score` in $[0,3]$; `confidence` in $[0,1]$; a proposal-grounded `comment`; `evidence_gap`; and `constraint_violation`.

H.1.2 PRIVILEGED CRITERION TEACHER

A deterministic error selector compares the student's evaluation-space prediction with the Rubric-A reference and prioritizes training proposals with large criterion-level error. The selected proposal, student assessment, reference annotation, and teacher-side evidence are then passed to the privileged teacher.

Privileged Criterion Teacher

System
You are the privileged teacher for Evaluation Space Learning. You receive one research proposal, the student's six-dimensional criterion assessment, the corresponding Rubric-A reference assessment, and privileged teacher-side evidence.
Diagnose why the proposal-time student under- or over-estimated one or more evaluation criteria. Convert each diagnosis into reusable guidance for future proposal-time assessments.
Privileged evidence may motivate the diagnosis, but every emitted lesson must be executable using proposal-visible information alone. Do not copy reference scores, source-paper outcomes, evidence identifiers, citations, adoption, community-response facts, or other hindsight-dependent information into the student-facing lesson.
Teach absolute proposal-level assessment rather than pairwise preference. Return exactly one JSON object.

User
[EVALUATION-CRITERION DEFINITIONS]
[PROPOSAL TEXT]
[STUDENT SCORES, CONFIDENCE, COMMENTS, AND FLAGS]
[RUBRIC-A REFERENCE ASSESSMENT — TEACHER ONLY]
[PRIVILEGED CRITERION EVIDENCE — TEACHER ONLY]
Return `lessons[]`. Each lesson contains: `lesson_type`; `evaluation_criterion`; `proposal_visible_trigger`; `suggested_check`; an optional `target_level`; `confidence_effect`; `leakage_audit`; and `student_safe_text`.

Audit and validation. Teacher-generated lessons pass through schema validation, leakage screening, deduplication, and conflict resolution before being attached to a candidate Rubric B or case memory. The candidate student state is then rerun on held-out proposals. An update is retained only when it improves the validation criterion; otherwise the rubric and memory are rolled back to the incumbent checkpoint. Leakage screening and predictive validation are performed separately: the former enforces the proposal-visible information boundary, whereas the latter prevents student-side updates from degrading held-out criterion assessment.

H.2 POPULATION TASTE DISTILLATION

Population Taste Distillation freezes the evaluation-space assessor and learns how evaluation-space features are translated into population-level research preferences. Population supervision is provided by the hindsight reference panel described in Appendix C.1. The deployed student can access the evaluation-space features and proposal-grounded evidence but cannot access panel labels, panel rationales, source-paper outcomes, community response, or other privileged signals.

H.2.1 POPULATION TASTE COMPARATOR

The deployment comparator applies the current population mapping to a proposal pair. Conditional-rule activations are computed outside the language model and supplied as fixed inputs. Given the criterion-difference representation $x_{A,B}$, proposal-visible evidence $\mathcal{E}_{A,B}$, and the current population artifacts, the comparator instantiates the preference function defined in Section 3.4,

$$u_{\text{pop}}(x_{A,B}, \mathcal{E}_{A,B}) = w_{\text{pop}}^{\top} x_{A,B} + \sum_{r \in R_{\text{pop}}} \tilde{\alpha}_r \phi_r(x_{A,B}, \mathcal{E}_{A,B}).$$

The binary preference is determined solely by the sign of this utility:

$$\hat{y}(A,B) = \begin{cases} A, & u_{\text{pop}}(x_{A,B}, \mathcal{E}_{A,B}) > 0, \\ B, & u_{\text{pop}}(x_{A,B}, \mathcal{E}_{A,B}) < 0. \end{cases}$$

Thus, auxiliary outputs such as confidence, decisive criteria, and rationale do not override the sign-based preference decision.

Population Taste Comparator

System
You are the deployment-constrained Population Taste Comparator. Your task is to apply the learned population-level research preference policy to two research directions.
Use only the supplied evaluation-space feature vectors, proposal-grounded evidence, externally computed conditional-rule activations, and learned student-safe guidance.
Never infer or request reference-panel labels, panel rationales, source-paper outcomes, citations, adoption, community response, or any other retrospective signal. Treat all externally computed numerical quantities as fixed inputs.
Apply the supplied population mapping to obtain the scalar preference utility u_{pop}. The preferred research direction is determined solely by the sign of this utility: positive utility selects A and negative utility selects B. Confidence, decisive criteria, and rationale are auxiliary outputs and must not alter this sign-based decision.
When the two research directions are close under the evaluation-space features, use the proposal-grounded comments, confidence values, evidence-gap flags, and constraint-violation flags only as permitted by the learned student-safe policy and conditional rules.
Return exactly one JSON object.

User
Apply the learned population mapping to research directions A and B.
[CRITERION-LEVEL FEATURE VECTOR AND EVIDENCE FOR A]
[CRITERION-LEVEL FEATURE VECTOR AND EVIDENCE FOR B]
[EXTERNALLY COMPUTED CONDITIONAL-RULE ACTIVATIONS]
[LEARNED STUDENT-SAFE POPULATION POLICY AND CASE MEMORY]
Return: `preference_utility`; `preference_probability`; `winner` (`A` or `B`); `confidence`; `decisive_criteria`; `used_rule_ids`; and a proposal-grounded `rationale`.
The returned `winner` must be consistent with `preference_utility`: return `A` if `preference_utility` > 0 and `B` if `preference_utility` < 0.

Pair orientation swapping is used only as a deterministic robustness and validation operation. It does not participate in construction of the reference-panel label. StrictPair and related orientation-consistency statistics are likewise computed outside the language-model workflow.

H.2.2 Privileged Population Teacher

The privileged teacher analyzes systematic student disagreements with the population reference. It may inspect panel-derived teacher-side metadata but must convert those diagnoses into instructions that remain executable under the deployment information boundary.

Privileged Population Teacher

System
You are the privileged teacher for Population Taste Distillation. Your task is to improve the student-safe population preference policy used by a deployment-constrained comparator.
You may inspect the population reference, teacher-side panel reasoning, decisive factors, and the reported relationship between criterion-level evidence and privileged hindsight evidence. The deployment comparator cannot inspect these signals.
Translate recurring failures into reusable instructions that can be evaluated from evaluation-space features and proposal-grounded evidence alone. Never mention citations, adoption, popularity, community response, the reference panel, or the reference label in a deployed instruction.
If a reference decision depends only on retrospective information for which no proposal-visible proxy exists, abstain rather than inventing a proxy.
Propose several small and distinct policy edits. Each candidate should target one identifiable failure mode so that it can be validated independently. Return exactly one JSON object.

User
Round: [ROUND INDEX]
Current population policy: [POLICY INSTRUCTIONS AND CUED MEMORY]

Student failures: [STUDENT INPUT, A/B AND B/A DECISIONS, AND PRIVILEGED REFERENCE RECORD]
Student successes that should not regress: [ANTI-REGRESSION CASES]
Return `candidates[]`. Each candidate contains: `candidate_id`; `diagnosis`; the complete replacement `instructions`; optional `memory_additions`; optional `memory_deletions`; and `abstained_pair_ids`.

Candidate packages are atomized into individual instruction or memory changes, screened for privileged-information leakage, and evaluated independently. The acceptance decision is made by a deterministic validation gate on the agent-validation split.

H.2.3 Conditional-Rule Proposer

A complementary teacher path proposes executable conditional rules. Privileged information may motivate a hypothesis, but every accepted rule must reduce to a predicate over one proposal's deployment-visible representation.

Privileged Conditional-Rule Proposer

System
You are a privileged hypothesis proposer for Population Taste Distillation. You do not issue the final preference judgment. Instead, you propose falsifiable conditional-taste hypotheses that a deterministic compiler and validation procedure will accept or reject.
Privileged information may inspire a hypothesis, but every deployed predicate must be evaluable independently on one research direction using its evaluation-space feature vector or a registered proposal-text extractor.
Pair-specific predicates, source-paper outcomes, citations, adoption, community response, publication venue, and reference-panel opinions are forbidden in deployed rules.

User
[CURRENT FAILURE TAXONOMY]
[WORKED EXAMPLES OF PREVIOUSLY ACCEPTED RULES]
[REFINEMENT HEURISTICS]
[ERROR CLUSTER]
For each case, you receive:

- evaluation-space features and proposal-grounded evidence;
- the current model prediction and margin;
- the population-reference preference and teacher-side rationale; and
- privileged retrospective signals.

Propose a small set of nonredundant hypotheses. Return `hypotheses[]`, each containing: `predicate`; `tier`; `direction` in $\{-1, +1\}$; `failure_mode`; a deployment-safe `explanation`; and a teacher-only `privileged_cue`.

Tier-1 predicates are arithmetic or Boolean functions of the six evaluation-space features and their deployment-visible evidence flags. Tier-2 predicates invoke previously registered semantic extractors over proposal-visible text. In both cases, a rule r defines a single-direction activation $\phi_r(P)$. For a proposal pair (A, B), the rule enters the comparison only through the activation difference

$$\phi_r(A) - \phi_r(B).$$

This preserves the general single-direction population mapping while instantiating it under the pairwise observation interface.

H.2.4 Proposal-Text Property Evaluator

Tier-2 predicates use a frozen proposal-text evaluator.

Proposal-Text Property Evaluator

System
Judge one research proposal against one precisely stated, present-tense property. Do not compare it with another proposal and do not speculate about future impact, adoption, reception, or other hindsight outcomes.
Return exactly one JSON object containing: `holds`; a proposal-grounded evidence span; and `confidence`.

User
Property: [REGISTERED PROPERTY DESCRIPTION]
Proposal-grounded text: [PROPOSAL OR PROPOSAL-GROUNDED EVIDENCE RECORD]
If the supplied text does not clearly establish the property, return `holds=false`.

The property description, evaluator model, and prompt are frozen when the extractor is registered and become part of its cache key.

H.2.5 Workflow Reflection

The executable-rule path additionally uses a reflection agent that receives only aggregate workflow statistics. It cannot observe evaluation examples or approve its own revisions.

Population-Rule Workflow Reflection

System
You improve the conditional-rule generation workflow using aggregate round-level statistics.
You may revise the proposer preamble, worked examples, failure taxonomy, refinement heuristics, or case-selection strategy. You may not revise the predicate grammar, information-boundary contract, validation criterion, or output schema.
You never observe the evaluation examples themselves.

User
[NUMBER OF PROPOSED AND TYPE-CHECKED HYPOTHESES]
[VALIDATION SURVIVAL RATE AND HELD-OUT GAIN]
[CLUSTER COVERAGE AND REDUNDANCY]
[COMMON REJECTION REASONS]
[CURRENT PROPOSER ARTIFACTS]
Diagnose the largest current weakness and propose one bounded revision. Return: `diagnosis`; `change_kind`; the corresponding replacement or addition; and `change_note`.

The incumbent and revised proposer configurations are subsequently evaluated under identical budgets by deterministic operators. The reflection agent cannot approve its own revision.

H.3 Individual Taste Adaptation

The numerical adaptation and replay gate are described in Appendix D. This section records the language-model prompts used only to propose conditional personal rules and proposal-text properties after residual fitting leaves recurring support errors.

H.3.1 Evaluation-Space Personal Rule Proposer

The first agentic rule path searches for conditional preferences expressible within the learned evaluation space.

Evaluation-Space Personal Rule Proposer

System
You diagnose why a personalized scientific-taste model still mispredicts a specific researcher's choices after a six-dimensional weight residual has already been fitted.

Continuous criterion reweighting has therefore already been attempted. Do not propose that an evaluation criterion should simply receive more or less global weight. Instead, propose a conditional rule that activates only under a repeated pattern shared by the remaining errors.
Each proposed trigger must be evaluable independently on one research direction from its evaluation-space feature vector. State only the preferred direction of the effect; rule strength will be estimated statistically outside the language model.
A rule that activates almost everywhere is a disguised global weight update. A rule that activates only on one example is an anecdote. Return only predicates under the frozen executable grammar.

User
Researcher: [USER ID]
Fitted individual weighting: [POPULATION WEIGHTS PLUS PERSONAL RESIDUAL]
Cross-fitted support errors that remain after residual fitting: [PER-SIDE EVALUATION-SPACE FEATURES, USER CHOICE, MODEL CHOICE, AND MARGIN]
Shared trigger pattern: [PATTERN]
Propose at most [BUDGET] conditional rules.

This proposer is invoked only after sufficient individual support has accumulated and a repeated residual-error pattern can be identified.

H.3.2 Proposal-Text Personal Property Proposer

If no evaluation-space conditional rule survives replay, the second agentic path may propose a proposal-visible property outside the representational span of the six predefined evaluation criteria.

Proposal-Text Personal Property Proposer

System
A personalized criterion-weight residual and evaluation-space conditional rules have already failed to explain the remaining choices. Propose a reusable property of one research proposal's text that captures the unresolved taste pattern.
The property must describe a dimension of research preference rather than a research topic, proposal identity, or pair identifier. It must describe what the proposal presently states, not its future impact, adoption, reception, or other hindsight outcome.

User
Researcher: [USER ID]
Fitted individual weighting: [WEIGHTS]
Remaining support errors: [EVALUATION-SPACE FEATURES, PROPOSAL-GROUNDED TEXT, USER CHOICES, AND OPTIONAL USER-PROVIDED RATIONALES]
Propose at most [BUDGET] properties. Each property must:

- be answerable for one research direction independently;
- plausibly generalize across research topics;
- separate a nontrivial subset of the support interactions rather than activating universally or on one proposal only; and
- avoid citations, adoption, reception, hidden user information, and other hindsight-dependent signals.

Return `extractors[]`, each containing: `name`; `description`; `direction`; `explanation`; and `why_these_cases`.

H.3.3 Property Rewrite Agent

If a proposed proposal-text property fails the statistical gate, a rewrite agent may inspect only the rejection diagnostics and the same observed support history. No additional labels are revealed.

Personal Property Rewrite Agent

System
A previously proposed personal-property extractor failed statistical validation. Propose a substantively different taste property rather than renaming the failed property.

User
[REMAINING CROSS-FITTED ERRORS]
[FAILED PROPERTY DESCRIPTIONS]
[ACTIVATION COUNTS, FIXED/BROKEN COUNTS, NLL CHANGE, AND REJECTION REASONS]
Propose new present-tense, per-proposal properties that separate the unresolved cases and activate on enough of the complete support buffer to estimate their effect.
Topic labels, pair memorization, hindsight outcomes, and hidden researcher information are forbidden.
Return the same extractor schema as the initial proposal-text property proposer.

H.4 Deterministic Components

For clarity, the following operations contain no language-model prompt: nearest-neighbor case retrieval, error ranking, logistic fitting of w_{pop} and Δw_u, proposal-order swapping, StrictPair computation, predicate parsing, rule-activation computation, coefficient fitting, reliability updates, cross-fitted replay, policy atomization, leakage screening, validation-gate decisions, rollback, and held-out evaluation.

This separation ensures that language-model agents propose or diagnose candidate textual artifacts, whereas acceptance and numerical parameter estimation remain externally controlled.

I Additional Related Work

Research Artifact Evaluation. The rapid development of scientific ideation systems (Wang et al., 2024; Baek et al., 2025; Si et al., 2025; Lu et al., 2024; Gottweis et al., 2026) has made the evaluation of ideas, proposals, and papers increasingly important (Ding et al., 2026; Ye et al., 2025). Existing benchmarks assess research artifacts along criteria such as novelty, feasibility, and potential impact (Wang et al., 2026; Jiang, 2026; Zhao et al., 2026a; Qiao et al., 2026), yet holistic LLM judgments do not always predict the quality of executed research (Si et al., 2026) and show particular limitations in novelty assessment (Sinhahajari et al., 2026). Evidence-grounded methods such as OpenNovelty (Zhang et al., 2026a) instead support assessment with retrieved literature and explicit evidence. Our work builds on this direction by grounding proposal-time criterion assessment in proposal claims and relevant literature, and organizing these assessments within a shared evaluation space.