GUIDING LLM AGENTS WITH MODEL-INTERNAL UNCERTAINTY FOR SCIENTIFIC DECISION-MAKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Scientific experiments can be expensive and slow to return feedback, making uncertainty important for deciding which actions to take under a limited evaluation budget. For large language model agents, however, it remains unclear whether internal representations provide useful uncertainty signals when labeled scientific interactions are scarce. We investigate this question using a frozen Qwen3.5-4B agent in Materials Discovery Environments (MADE) and DiscoveryWorld. Our approach fits a lightweight logistic predictor offline to estimate action-failure risk from public-state and output features augmented with internal activation summaries. These summaries are extracted at the prediction positions of generated action tokens and used to score candidate actions before execution. Four experimental conditions separate single-proposal planning, two-proposal generation, public-feature risk selection, and selection augmented with internal features. Exploratory development comparisons that preserve the public-feature transformation show improved failure ranking: AUROC increases from 0.5467 to 0.5622 in MADE and from 0.9667 to 0.9867 in DiscoveryWorld Proteomics. Online evaluations identify task-specific improvements in DiscoveryWorld, while analyses of matched actions and shared candidate pools distinguish predictive information from changes in action selection. The findings support internal activations as an additional source of failure-risk information in the tested settings, while showing that improved prediction alone does not ensure better scientific outcomes.

1 INTRODUCTION

Scientific agents must select experiments before their consequences are known. In materials discovery, high-fidelity simulations and physical evaluations can be costly, while synthesis and characterization can delay feedback (Malik et al., 2026; Szymanski et al., 2023). A poor proposal can consume part of a limited evaluation budget before feedback arrives. Estimating failure risk before execution could help the agent choose what to evaluate.

Large language model (LLM) agents can plan scientific workflows and invoke external tools (Boiko et al., 2023; M. Bran et al., 2024). A plausible proposal may fail during execution or produce no useful outcome. We estimate uncertainty as the probability of a specified action-failure event, learned from observed execution data. This estimate differs from uncertainty about a material's physical behavior or a scientific hypothesis. Its usefulness depends on the predicted event and the scientific objective.

Internal representations can help predict model errors (Azaria & Mitchell, 2023; Kossen et al., 2024). Their value for scientific agents with limited labeled interactions requires a separate test. Distinguishing failures across examples need not rank two actions correctly in the same state. Extra proposals can affect performance independently of risk estimates, and avoiding an immediate failure need not advance the scientific task. We therefore compare predictors under the same proposal procedure and assess prediction, action choice and scientific outcomes separately.

To investigate this, we augment a frozen LLM with a fitted predictor designed to select between two generated proposals. We define our baseline predictor, denoted as P, which relies exclusively on public-state and output features. We compare this against an augmented predictor, denoted as P+C, which additionally incorporates Captured model-internal signals—specifically, MLP input/output statistics extracted just before action-value tokens, alongside two auxiliary measurements. Here,

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.

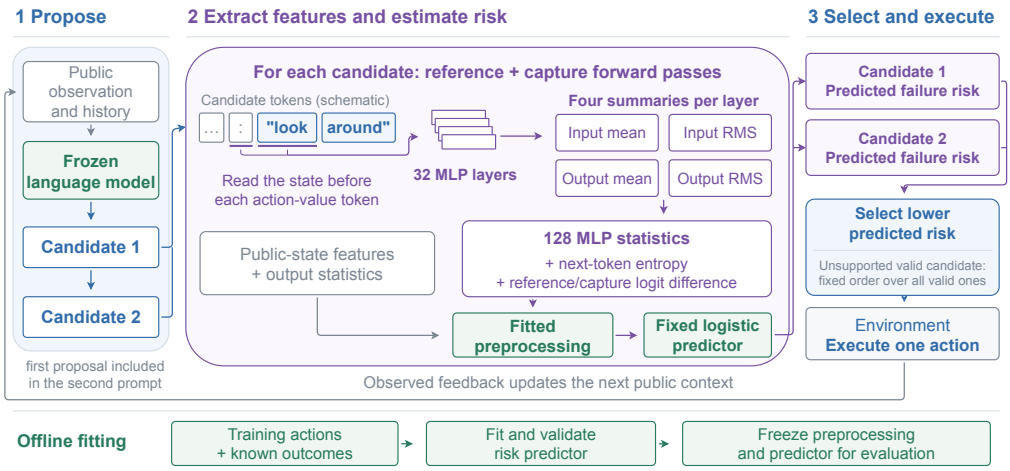

Figure 1: P+C for two valid, supported proposals. The frozen LLM generates them sequentially. Each is scored from public/output features, 128 MLP mean and root-mean-square (RMS) statistics at action-prediction positions, and two auxiliary measurements. Offline fitting (lower band; Section 3.4) fixes preprocessing and the logistic predictor before evaluation. One action is executed; its feedback updates the public context. Ties and missing support follow Algorithm 1.

uncertainty means the predicted probability of a specified failure event. Execution feedback changes prompts, but the LLM and predictor remain fixed.

We test whether model internals provide useful action-failure uncertainty estimates for scientific agents and whether these estimates improve their decisions. Three research questions guide the study:

- **RQ1:** Do model-internal features improve action-failure prediction beyond public-state and generated-output features?
- **RQ2:** When does that signal translate into better sequential scientific decisions?
- **RQ3:** On the same candidate pools, how often do the public and internally augmented predictors select different executable actions?

We compare predictions on identical examples for RQ1, matched four-arm trajectories and task-level outcomes for RQ2, and action selections on shared candidate pools for RQ3. Controls for public features and proposal count distinguish the augmented selector from extra generation alone. All themes, budgets and comparison directions are retained.

2 Related Work

Predicting errors and uncertainty. Self-evaluation and verbal confidence ask a model to assess the reliability of its answers (Kadavath et al., 2022; Lin et al., 2022). Semantic uncertainty groups answers by meaning before measuring their variability (Kuhn et al., 2023); semantic entropy can detect confabulations, a subset of hallucinations (Farquhar et al., 2024). Calibration measures agreement between predicted probabilities and observed frequencies (Guo et al., 2017). Our target is a specified action failure, distinct from answer uncertainty and scientific return.

Scientific agents. ReAct combines reasoning with external actions (Yao et al., 2022), while Reflexion uses language feedback and episodic memory without updating model weights (Shinn et al., 2023). Coscientist and ChemCrow connect language models to experimental planning and chemistry tools (Boiko et al., 2023; M. Bran et al., 2024). MADE evaluates materials discovery under an oracle budget (Malik et al., 2026); DiscoveryWorld evaluates scientific reasoning through interaction (Jansen et al., 2024). They distinguish valid tool use from scientific progress.

Predictors based on internal representations. Hidden-state classifiers can distinguish true from false statements (Azaria & Mitchell, 2023), and semantic-entropy probes estimate semantic uncertainty from the representations of one generation (Kossen et al., 2024). Transcoders approximate MLP computations for circuit analysis (Dunefsky et al., 2024). Circuit-based Reasoning Verification uses structural features of attribution graphs to diagnose reasoning errors (Zhao et al., 2025). In this work, We instead add pooled MLP statistics and two auxiliary measurements to public-state and output features to predict action failure, without constructing a sparse dictionary or attribution graph. The original comparison changes the representation, transformation and fitted predictor together; it does not isolate an MLP effect.

3 METHODS

Appendix A defines the notation used in the text, Algorithm 1 and Figure 1. We use c for proposal indices and j for token positions.

3.1 DECISION PROCESS AND FAILURE TARGETS

Let h_t contain the public observation and retained interaction history at decision t. The frozen policy π_{θ_0} generates a response, giving a prompt–completion token sequence $q_{t,c}$ and a parsed action $a_{t,c}$. If the response cannot be parsed, $a_{t,c} = \perp$. Different responses, including different rationales, can describe the same executable action. Selection uses only information available before the decision, including prior tool results and visible buffer scores in MADE. Hidden evaluator state and the proposed action's unobserved outcome are excluded; private task-progress scores are not inputs to online action selection. They are used only for outcome evaluation, not online selection.

A labeled record is $(h_i, q_i, a_i, y_i, e_i, g_i)$, where e_i is its episode and g_i its declared fitting group. Labels are obtained from observed outcomes of the executed action. In MADE, the eligible records are `select_for_evaluation` actions, with

$$y_i = 1 - \mathbb{1}\{\text{the evaluated candidate is stable and newly discovered}\}. \tag{1}$$

All 350 MADE training and development records are selection actions with an observed evaluation. Preparation and query actions are not assigned labels from later outcomes. In DiscoveryWorld, $y_i = 1$ means that the official response reports unsuccessful execution. An action that executes but makes no task progress is not a failure under this label. Unknown outcomes are excluded from fitting, and unexecuted proposals receive no counterfactual labels.

The predictor $\widehat{r}(h, q)$ estimates a specified failure event under the data-collection policy, not future task return. Calibration requires separate evaluation, especially after a policy or task change. We test whether selecting actions with lower predicted risk improves scientific outcomes.

3.2 TWO-PROPOSAL WORKFLOW AND CONTROLS

Four arms distinguish proposal generation from risk-based selection. *Single* generates one proposal ($K = 1$); *Two* generates two and uses a fixed, risk-independent ordering; *P* selects using public and generated-response features; and *P+C* adds the activation summaries and two auxiliary measurements. Two–Single measures the combined effect of an extra proposal and feedback, and P–Two tests public-feature risk selection. The primary P+C–P comparison tests the added features together with their transformation and separately fitted predictor, not an isolated MLP effect.

All two-proposal arms generate exactly two proposals ($K = 2$) in sequence with the same sampling-seed rule, generation settings and prescribed public feedback. The second is requested even when the first is valid. Both remain selectable; duplicate actions are recorded rather than resampled. In MADE, the second prompt includes the first proposal and up to three recent public tool failures. In DiscoveryWorld, it includes the first parsed action, any parsing error and a fixed instruction to propose an alternative. Predicted risks are never presented as observed environment feedback.

The two-proposal arms share eligibility and fallback rules, but their trajectories can diverge after a selected action changes the next observation. We therefore measure prediction, within-state choices and scientific return separately. Action changes refer to different executable inputs, not merely different candidate indices or necessarily different environmental effects.

3.3 PUBLIC AND INTERNAL FEATURES

For feature comparisons, A contains generated-response and action features, including generation-probability summaries and action descriptors; P adds a fixed public-state summary. B and C add full-sequence and action-position features to A, respectively; P+B and P+C add them to P without duplicating A. The public features are a specified baseline, not a complete prompt encoding. Appendix B gives dimensions and implementation details.

We identify the generated-token positions $\mathcal{V}(q)$ that contain executable action values: the tool name and scalar argument values in MADE, and the executable JavaScript Object Notation (JSON) fields in DiscoveryWorld. Field names, rationale text and tokens containing only syntax are excluded. A token may include both an action value and adjacent punctuation if the value can be aligned unambiguously; tokens that inseparably mix action values with excluded fields are unsupported. The positions always refer to the original generated token IDs, without retokenizing the response.

For each target token at position j, we extract the state at its *prediction position* $j-1$:

$$\mathcal{I}(q) = \{j-1 : j \in \mathcal{V}(q)\}. \tag{2}$$

This uses the causal state before the target token is emitted. Let $u^s_{\ell pk}$ be MLP input or output coordinate k at layer ℓ and position p, with $s \in \{\text{in}, \text{out}\}$ and width d^s_ℓ. We compute

$$\mu^s_\ell = \frac{\sum_{p\in\mathcal{I}} \sum_{k=1}^{d^s_\ell} u^s_{\ell pk}}{|\mathcal{I}| d^s_\ell}, \qquad \rho^s_\ell = \left(\frac{\sum_{p\in\mathcal{I}} \sum_{k=1}^{d^s_\ell} (u^s_{\ell pk})^2}{|\mathcal{I}| d^s_\ell} \right)^{1/2}. \tag{3}$$

Pooling produces 128 activation statistics from the frozen Qwen3.5-4B model (Qwen Team, 2026). Full-sequence pooling replaces $\mathcal{I}$ with all unmasked positions of the same prompt–completion sequence. Both summaries discard vector directions, individual feature identities and interactions between features.

Feature extraction uses a reference forward pass and a pass with observation hooks on the same token sequence (Appendix B). Two additional measurements accompany the MLP statistics: the raw next-token entropy at the final sequence position and the maximum absolute difference between reference and captured logits. These measurements are shared by the full-sequence and action-position views. Thus the added representation contains 130 fields, and its effect cannot be attributed to the 128 MLP statistics alone.

3.4 RISK FITTING AND CALIBRATION

For each feature set, let $x(h,q) \in \mathbb{R}^d$ contain its d raw values. The fitted transformation $\phi(h,q)$ imputes nonfinite values with training medians, standardizes using training statistics, and appends missingness indicators and three failure-prototype fields. Thus $\phi_i = \phi(h_i, q_i) \in \mathbb{R}^{2d+3}$. The prototype is the mean standardized vector among training failures; the appended fields are root-mean-square (RMS) distance, cosine similarity and an indicator for its absence. Vocabulary, preprocessing and prototypes are fitted within each cross-validation training fold.

The original online transformations are not nested: when C is added, distance and cosine are recomputed in the larger feature space rather than retaining their P-only values. The comparison therefore changes the added features, two nonlinear summaries and the separately fitted predictor together. Section 4.2 also reports a later development-set analysis that keeps the complete public transformation unchanged.

The predictor is regularized logistic regression. For n_{train} eligible training records, with episode-balanced weights $w_i = n_{\text{train}}/(E n_{e_i})$, we fit

$$(\widehat{\beta}, \widehat{b}) = \arg\min_{\beta,b} \left[\sum_{i=1}^{n_{\text{train}}} w_i \, \mathcal{L}_{\text{BCE}}\big(y_i, \sigma(\beta^\top \phi_i + b)\big) + \frac{\|\beta\|_2^2}{2C_{\text{reg}}} \right], \tag{4}$$

where E is the number of training episodes, n_e their eligible record counts, $\mathcal{L}_{\text{BCE}}$ is binary cross-entropy (BCE), σ is the logistic sigmoid, and the intercept is unpenalized. The four values $C_{\text{reg}} \in \{0.001, 0.01, 0.1, 1\}$ are compared by held-out negative log-likelihood (NLL) under the declared group/episode weighting; exact ties prefer smaller C_{reg}. MADE groups folds by training chemistry;

DiscoveryWorld groups them by training episode. Each view selects its own regularization using the same procedure. Episode balancing equalizes total episode weight, but does not make correlated actions independent.

The reported failure probability is

$$\widehat{r}(h,q) = \sigma\Big((\widehat{\beta}^\top \phi(h,q) + \widehat{b})/T\Big), \quad T > 0. \tag{5}$$

MADE fixes $T = 1$. DiscoveryWorld fits T on a separate calibration world by minimizing weighted NLL, with $\log T \in [-4, 4]$ (Guo et al., 2017). Development labels are not used to fit T. A fixed positive temperature preserves within-predictor ranking; calibration is evaluated separately from action selection. Neither θ_0 nor $\phi, \widehat{\beta}, \widehat{b}, T$ is updated during online evaluation.

3.5 Support, Selection and Fallback

Let S_t contain parser-valid proposal indices and let $\eta_{t,c} \in \{0, 1\}$ indicate whether the active predictor can score a proposal. Both risk arms use the same task-eligibility and action-token alignment rules: an eligible evaluation tool in MADE and the declared action schema in DiscoveryWorld. Each predictor also needs the features required by its own view; P+C additionally requires successful activation capture. The public-feature selector checks this alignment without capturing activations. Support certifies feature availability, not in-distribution status or calibrated risk. If $S_t \neq \emptyset$ and $\eta_{t,c} = 1$ for *every* $c \in S_t$, a risk arm uses

$$c_t^* = \arg\min_{c \in S_t} (z_{t,c}, \kappa_{t,c}), \tag{6}$$

with lexicographic ordering and a fixed risk-independent key $\kappa_{t,c}$. The executed action is $a_t^* = a_{t,c_t^*}$; the star denotes selection, not global optimality. MADE uses $z = \widehat{r}$; DiscoveryWorld uses the temperature-scaled logit to avoid ranking through a saturated sigmoid.

If any valid candidate is unsupported, selection falls back to $\arg\min_{c \in S_t} \kappa_{t,c}$ over *all* valid candidates. Unsupported scores are never replaced by zero. With no valid candidate, the benchmark adapter records rejection or applies its prescribed recovery; no fabricated action outcome is introduced.

Native recovery can also precede a policy decision. In MADE, the fixed handler is invoked after ten no-action generation decisions or tool attempts since the last evaluation-step response (or initialization). A two-proposal decision with no executable action increments this counter once; the counter is distinct from the candidate-evaluation budget. Such actions bypass proposal generation and risk selection and are counted separately.

4 Experiments

4.1 Evaluation Protocol

Tasks and controls. All experiments use the frozen Qwen3.5-4B checkpoint and the four arms in Section 3: Single, Two, P and P+C. MADE uses Mg–Sn–Sr and Au–K–Tb with policy seeds 501–510 at budgets B10/B20/B30. The earlier DiscoveryWorld Proteomics/Normal cohort, DW-P, uses worlds 3/4, seeds 501–510 at B10/B20/B30 and 501–503 at B100. These instances were previously examined during development. The separate standard-DW evaluation uses Plant Nutrients, Space Sick, It's (not) Rocket Science!, and Lost in Translation, with five official API seeds per theme, policy seed 501, and both Easy and Normal. DW-P is not a fifth theme of this evaluation.

Budgets and outcomes. MADE's budget counts evaluations by the pretrained Orb-v3 interatomic potential. SUN is the stable, unique and novel discovery count $D(B)$; mSUN is $D(B)/B$ under the configured stability tolerance. The normalized area under the discovery curve (AUDC) is

$$\mathrm{AUDC}_B = \frac{2}{B^2} \int_0^B D(\tau)\, \mathrm{d}\tau, \tag{7}$$

using trapezoidal integration on the recorded evaluation-attempt coordinate. Stability can change with the phase diagram; monotonicity is not imposed. DW-P limits requests and explicit ticks.

Table 1: RQ1 development-only controls preserving the complete public transformation. P+MLP appends 128 MLP summaries and their missingness indicators, without the two auxiliary measurements. MADE has 100 records from two chemistries; DW-P has 40 records from two episodes in one world. NLL and Brier average equally over chemistry/episode groups; AUROC is pooled. These heads were not deployed online.

Environment	Features	NLL ↓	Brier ↓	AUROC ↑
MADE	P	0.6858	0.2457	0.5467
MADE	P+MLP	0.6829	0.2450	0.5622
DW-P	P	0.2034	0.0620	0.9667
DW-P	P+MLP	0.2940	0.0771	0.9867

Standard DW allows 100 post-load rounds on Easy and 1000 on Normal. Its task score is 100 times official normalized progress. Each budget defines a separate trajectory, not a prefix or a common cost unit across benchmarks. Execution success, progress and task completion are reported separately.

Matching and technical recovery. The original MADE runs provide 239 of 240 validated endpoints. The failed P run leaves 19/20/20 complete four-arm cases, while Single–P+C comparisons have twenty pairs at every budget. A separately registered recovery adds a sixty-case four-arm sensitivity analysis. DW-P has all 264 endpoints. Standard Easy has 78 original endpoints: P is missing Plant Nutrients/API seed 2 and Space Sick/API seed 2. All-arm comparisons therefore have eighteen original cases, or twenty after the two registered recoveries. Normal has eighty endpoints and twenty complete cases. Original and amended results are distinct views, never independent replications. Missing outcomes are not replaced by zero; Appendix E documents recovery and uncertainty estimates.

4.2 RQ1: Incremental Information about Action Failure

We compare predictors on identical labeled examples, with the same fitting groups and search procedure. MADE supplies 250 training and 100 development records; DW-P supplies 60 training, 60 calibration and 40 development records. The original P+C predictor increases AUROC from 0.5467 to 0.5610 in MADE and from 0.9667 to 0.9867 in DW-P. Table 7 gives all six original feature views.

The original P+C–P comparison also changes two auxiliary measurements, feature-space prototype summaries and the separately fitted head. To examine the MLP contribution more closely, we fit controls on the training data and evaluate them on the development sets. These controls preserve every coordinate of the complete public transformation and append only standardized MLP statistics and their missingness flags. Table 1 reports these nested comparisons; the auxiliary-only and combined controls remain in Appendix C.1. Adding MLP statistics improves ranking on these development examples in both environments. MADE also shows a small decrease in probability error; DW's Brier and NLL move in the opposite direction. Thus these results support incremental ranking information in the tested representation, not uniformly better probabilities. The development groups are small and correlated, and the newly fitted heads were not used online. Their performance cannot explain the online selector's outcomes by itself.

4.3 RQ2: From Failure Signals to Scientific Decisions

Matched scientific outcomes. Table 2 retains every MADE and DW-P budget. Differences between P+C and P vary in direction, and the reported discovery/progress intervals include zero. This motivates examining individual tasks and the proposal opportunities on which a risk selector can act. Improvements over Single assess the complete procedure; comparisons against Two and P are needed to distinguish risk selection from extra proposals and public information.

DiscoveryWorld across themes and difficulties. The standard-DW evaluation tests transfer from Proteomics-trained risk predictors to four official themes. Figure 2 expands the comparison to every theme and all three controls. Points are matched API-seed differences, and diamonds are their five-case means. These seeds vary environment instances, not policy retraining; the native simulator's object-level RNG behavior is unchanged, so seed matching does not guarantee identical hidden

Table 2: Original four-arm comparison on identical cases. MADE reports mSUN and DW-P reports official Proteomics/Normal progress. The last column is P+C minus P. The missing original MADE public-feature case is excluded from all four means; the separate amended result is reported in the appendix. Arm names are defined in Section 3.

Cohort	B	n	Single	Two	P	P+C	Δ
MADE	10	19	0.4579	0.4158	0.4632	0.4632	+0.0000
MADE	20	20	0.3850	0.4475	0.3975	0.4025	+0.0050
MADE	30	20	0.4567	0.4533	0.4017	0.3767	-0.0250
DW-P	10	20	0.0312	0.0500	0.0625	0.0688	+0.0063
DW-P	20	20	0.0312	0.0813	0.0688	0.0688	+0.0000
DW-P	30	20	0.0875	0.0750	0.1000	0.0750	-0.0250
DW-P	100	6	0.1250	0.2917	0.2708	0.2708	+0.0000

Table 3: Standard DiscoveryWorld mean normalized task score multiplied by 100; differences use unrounded values. Each row matches all four arms on n cases. Easy uses 100 post-load rounds and Normal 1000. Easy original excludes two incomplete P cases; amended includes their separately registered recoveries. The Easy rows overlap.

Cohort	n	Single	Two	P	P+C	$\Delta_{\mathrm{P+C-P}}$
Easy original	18	15.97	27.08	23.38	26.62	+3.24
Easy amended	20	17.50	26.25	25.42	28.96	+3.54
Normal	20	22.42	27.58	24.16	22.57	-1.60

states across arms. In the amended Easy view, Plant Nutrients improves by 40.0 points over Single and 20.0 over P; Space Sick improves by 12.5 and 7.5 points, respectively. Both also exceed Two. On Normal, P+C has higher mean scores than P on Plant Nutrients, Rocket Science and Space Sick, but Two matches or exceeds P+C on those themes. All comparators are therefore needed to interpret the gains. The full figure retains Rocket Science's Easy decrease, Translation's Normal decrease and tied outcomes rather than averaging them away.

Across all themes, amended Easy P+C–P is $+3.5417$ points and Normal is -1.5986; both exploratory four-theme intervals include zero. These results locate gains in partial task progress within specific settings, rather than establishing a general advantage of internal features. Appendix E.3 provides all theme means, original and amended results, and the completion details.

4.4 RQ3: Changes in Action Selection

Figure 3 partitions the same 4128 recorded candidate pools from 126 P+C trajectories. A jointly eligible pool contains two parser-valid candidates with distinct executable inputs that both original predictors can score. This gives three mutually exclusive categories: no such pair, an eligible pair with the same selected action, and an eligible pair with different selected actions. All 456 observed P/P+C action differences fall within the 1559 eligible pools; the other 1103 eligible pools yield the same selection. Appendix D gives the validation procedure and per-budget counts. The seven rows retain their own denominators because the benchmarks and budgets induce different numbers of decisions. The displayed fractions are not pooled performance estimates or independent replicates. The large non-eligible portions show how often the prerequisites for this paired ranking comparison are absent; eligible agreement shows that a different predictor need not select a different action. Neither category alone explains an outcome gain. MADE uses saved P+C scores in this replay, rather than reconstructed internal features, and the original support, fallback and tie rules remain unchanged.

Support and candidate diversity determine the opportunities for risk selection. MADE scores 60.0–66.8% of recorded P+C candidates. DW-P scores every candidate, but 53.8–64.0% of valid pairs are duplicate actions. In standard DW, duplicates account for 1157/2000 valid pairs on amended Easy and 11043/19999 on Normal. Against the fixed Two ordering at each recorded pool, P+C changes 418/2000 Easy decisions and 4557/20000 Normal decisions. These counts establish that the selector

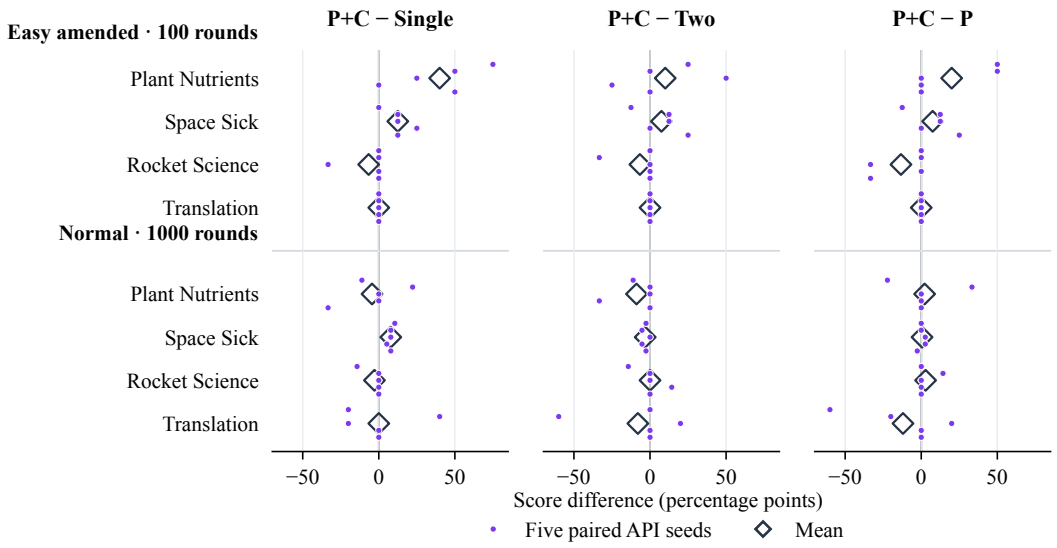

Figure 2: P+C minus each comparator on every standard-DiscoveryWorld theme, in task-score percentage points. Each row shows all five paired API-seed differences and their mean; these are observations, not confidence intervals. Easy uses the separately amended twenty-case view, including two registered P recoveries; Normal uses the twenty original cases. The original eighteen-case Easy comparison remains in Table 3 and Appendix E.3. Comparators share the frozen policy; Two and P also share the two-proposal procedure.

can intervene; they do not assign outcomes to unexecuted alternatives or identify the cause of a score gain.

Cost and interpretation. Two proposals generally require more generated tokens than Single; P+C additionally performs reference and capture passes. Equal environment budgets are not equal computation. Appendix F reports observed costs and coverage. Taken together, the three RQs distinguish available predictive information, opportunities to choose a different action, and realized scientific return. The current data do not isolate whether a given unhelpful choice arose from an inadequate candidate set, within-state ranking or a mismatch between failure and progress.

5 Limitations

The study uses one language-model checkpoint and limited tasks. Repeated policy seeds on previously examined MADE and DW-P instances do not establish new-task transfer; standard DW covers four themes and two difficulties. The nested feature controls are development-only, whereas the deployed P+C selector includes auxiliaries and a view-specific transformation. A failure probability is not a task-value estimate, and action disagreement does not show which choice is better. The five API seeds per standard-DW theme do not measure model-retraining variability. Intervals are exploratory, hardware and load were uncontrolled, and some interrupted-operation costs remain unknown. Computational and simulated evaluations do not establish experimentally validated discoveries. Laboratory costs and feedback delays motivate the study; they are not reproduced by these benchmarks, and we do not infer reductions in physical experimental cost or duration.

6 Conclusion

We evaluate the use of model-internal features to estimate action-level uncertainty and guide action selection in scientific agents. The development controls address RQ1 with evidence of additional ranking information; the online comparisons address RQ2 with task-specific partial-progress differences. For RQ3, recorded candidate pools distinguish scoring support from changes in the selected

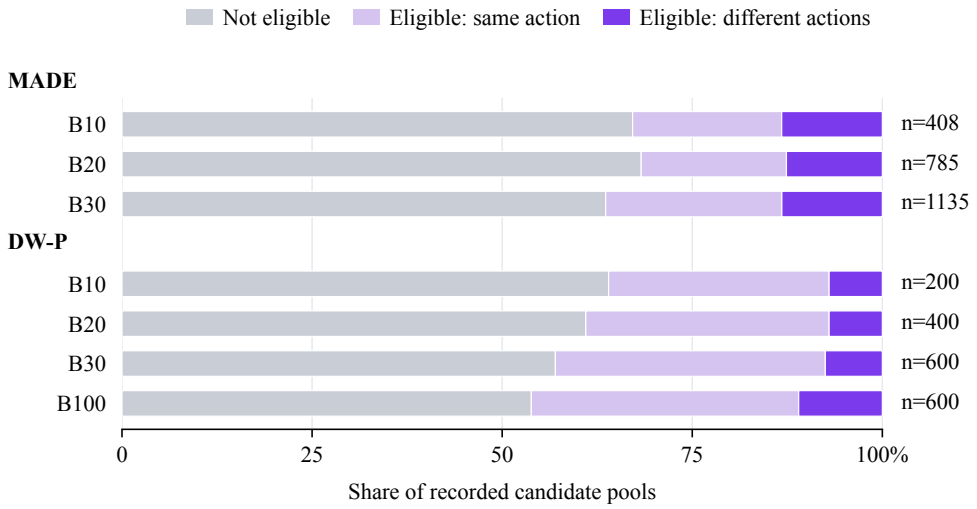

Figure 3: Opportunities for different action selections on the recorded MADE/DW-P candidate pools. Each bar includes every pool at that budget; n counts pools, not independent trials. Eligible means two parser-valid, distinct executable inputs supported by both original P and P+C predictors. Light and dark purple distinguish agreement from disagreement between their selected actions. Gray pools lack this joint eligibility. All observed disagreements occur in the eligible subset. The records come from P+C-visited states; DW-P is the earlier Proteomics cohort. This is a choice diagnostic, not a comparison of outcomes for unexecuted alternatives.

executable action. These findings concern the tested risk predictors and settings; they do not establish a general improvement in scientific discovery or an optimal allocation of experimental resources.

Reproducibility

The package provides feature schemas, prompts, label rules, model identifiers, source hashes and numerical reproduction scripts. It reproduces analyses from supplied data, but does not distribute every raw trajectory, model checkpoint or simulator needed to rerun all experiments.

Artificial Intelligence Assistance

Generative artificial intelligence (AI) assisted research design, code, tests, experiment orchestration, analysis, literature checking, figures and writing. Results come from recorded executions; unexecuted alternatives have no assigned outcomes. The authors are responsible for verifying all methods, results and claims.

References

Amos Azaria and Tom Mitchell. The internal state of an LLM knows when it's lying. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 967–976, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.68. URL https://aclanthology.org/2023.findings-emnlp.68/.

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

Jacob Dunefsky, Philippe Chlenski, and Neel Nanda. Transcoders Find Interpretable LLM Feature Circuits, 2024. URL https://arxiv.org/abs/2406.11944.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, June 2024. doi: 10.1038/s41586-024-07421-0. URL http://dx.doi.org/10.1038/s41586-024-07421-0.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On Calibration of Modern Neural Networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Peter Jansen, Marc-Alexandre Côté, Tushar Khot, Erin Bransom, Bhavana Dalvi Mishra, Bodhisattwa Prasad Majumder, Oyvind Tafjord, and Peter Clark. DiscoveryWorld: A virtual environment for developing and evaluating automated scientific discovery agents, 2024. URL https://arxiv.org/abs/2406.06769.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language Models (Mostly) Know What They Know, 2022. URL https://arxiv.org/abs/2207.05221.

Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth Malik, and Yarin Gal. Semantic Entropy Probes: Robust and Cheap Hallucination Detection in LLMs, 2024. URL https://arxiv.org/abs/2406.15927.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation, 2023. URL https://arxiv.org/abs/2302.09664.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching Models to Express Their Uncertainty in Words, 2022. URL https://arxiv.org/abs/2205.14334.

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D. White, and Philippe Schwaller. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6(5):525–535, May 2024. doi: 10.1038/s42256-024-00832-8. URL http://dx.doi.org/10.1038/s42256-024-00832-8.

Shreshth A Malik, Tiarnan Doherty, Panagiotis Tigas, Muhammed Razzak, Stephen J. Roberts, Aron Walsh, and Yarin Gal. MADE: Benchmark Environments for Closed-Loop Materials Discovery, 2026. URL https://arxiv.org/abs/2601.20996.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language Agents with Verbal Reinforcement Learning, 2023. URL https://arxiv.org/abs/2303.11366.

Nathan J. Szymanski, Bernardus Rendy, Yuxing Fei, Rishi E. Kumar, Tanjin He, David Milsted, Matthew J. McDermott, Max Gallant, Ekin Dogus Cubuk, Amil Merchant, Haegyeom Kim, Anubhav Jain, Christopher J. Bartel, Kristin Persson, Yan Zeng, and Gerbrand Ceder. An autonomous laboratory for the accelerated synthesis of inorganic materials. *Nature*, 624 (7990):86–91, November 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06734-w. URL http://dx.doi.org/10.1038/s41586-023-06734-w.

Ruoyao Wang, Peter Jansen, Marc-Alexandre Côté, and Prithviraj Ammanabrolu. ScienceWorld: Is your agent smarter than a 5th grader? In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11279–11298, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.775. URL https://aclanthology.org/2022.emnlp-main.775/.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing Reasoning and Acting in Language Models. *arXiv preprint arXiv:2210.03629*, 2022.

Zheng Zhao, Yeskendir Koishekenov, Xianjun Yang, Naila Murray, and Nicola Cancedda. Verifying Chain-of-Thought Reasoning via Its Computational Graph, 2025. URL https://arxiv.org/abs/2510.09312.

A NOTATION

Table 4 summarizes the mathematical symbols and reporting conventions. Proposal index c differs from token index j. The fitting and evaluation sample counts are defined separately below.

Table 4: Shared notation and reporting conventions.

Symbol / code	Definition and scope
Interaction	
t, c, i	Decision, proposal and labeled-record indices, respectively; c is not a token position.
$h_t;\ \theta_0, \pi_{\theta_0}$	Public observation and retained interaction history; fixed language-model parameters and proposal policy.
$q_{t,c};\ a_{t,c}, \perp$	Original prompt-completion token sequence; parsed executable action, or an invalid proposal. Different sequences can encode the same action.
K	Number of proposals: one for Single, two for Two, P and P+C. Duplicate actions remain selectable.
$y_i \in \{0, 1\}$	Observed failure: not both stable and newly discovered (MADE), a known unsuccessful execution response (DW), or a specified rejection-message proxy (ScienceWorld). Unknown outcomes are unlabeled.
e_i, g_i	Episode and fitting-group identities of record i; the declared groups determine training/validation separation.
Internal readout	
$j, p;\ \mathcal{V}(q), \mathcal{I}(q)$	Token position j, activation position p, action-value positions $\mathcal{V}(q)$ and prediction positions $\mathcal{I}(q) = \{j - 1 : j \in \mathcal{V}(q)\}$.
$\ell, L;\ s, k, d_\ell^s$	Layer index and count; boundary $s \in \{\text{in}, \text{out}\}$, coordinate index and width. Here $L = 32$ and $d_\ell^s = 2560$.
$u_{\ell pk}^s$	Activation at boundary s of layer ℓ, sequence position p and coordinate k.
μ_ℓ^s, ρ_ℓ^s	Candidate-specific mean and root mean square over aligned prediction positions and coordinates. Four statistics per layer give 128 values.
$H_{\text{next}}(q), \Delta_{\text{cap}}(q)$	Raw next-token entropy at the final sequence position and maximum reference/capture logit discrepancy; the two auxiliary fields, distinct from the 128 MLP statistics.
Features and fitting	
$x(h, q) \in \mathbb{R}^d$	The d-dimensional raw feature vector for one candidate and its public context; distinct from the scalar cost coordinate x.
$\widetilde{x}, v$	Imputed, standardized raw features and the mean standardized training-failure prototype.
$\phi(h, q), \phi_i$	Fitted map, with $\phi_i = \phi(h_i, q_i) \in \mathbb{R}^{2d+3}$: standardized values, missingness indicators and three prototype summaries.
$\beta, b; \widehat{\beta}, \widehat{b}$	Logistic coefficient vector and unpenalized intercept; hats denote fitted values.
$n_{\text{train}}, E, n_e, w_i$	Training-record count, episode count, records in episode e, and episode-balanced weight $w_i = n_{\text{train}}/(E n_{e_i})$.
$\mathcal{L}_{\text{BCE}}, C_{\text{reg}}$	Binary cross-entropy and inverse regularization parameter, with penalty $\|\beta\|_2^2/(2C_{\text{reg}})$. This C_{reg} is distinct from feature code C.
$\sigma, T, \widehat{r}(h, q)$	Logistic sigmoid, positive risk-logit temperature and failure probability $\widehat{r}(h, q) = \sigma((\widehat{\beta}^\top \phi(h, q) + \widehat{b})/T)$. T is fixed during evaluation, not a sampling temperature.
Decision rule	
$S_t, \eta_{t,c}$	Parser-valid proposal indices and binary score-support indicator. Risk ranking requires a nonempty valid set with every member supported.
$z_{t,c}, \kappa_{t,c}$	Ordering score and fixed tie/fallback key. z is failure probability in MADE and temperature-scaled logit in DW/ScienceWorld; lower is preferred.
c_t^*, a_t^*	Selected proposal index and its executed action. The star denotes selection, not optimality.

Continued on next page

Table 4 (continued)

Symbol / code	Definition and scope
Evaluation	
$B, \tau, D(\tau)$	Configured budget; MADE evaluation-attempt coordinate and SUN count. Units are benchmark-specific. ScienceWorld uses native step limit B, request guard $2B$ and the original post-action $\mathrm{moves} > B$ check.
$\mathrm{SUN}, \mathrm{mSUN}, \mathrm{AUDC}_B$	MADE discovery count $D(B)$, fraction $D(B)/B$, and normalized area $\mathrm{AUDC}_B = (2/B^2)\int_0^B D(\tau)\,\mathrm{d}\tau$, using recorded-axis trapezoidal interpolation.
$n_{\mathrm{eval}}, n_{\mathrm{pair}}$	Number of evaluated units specified by the table and number of complete matched pairs. A table-local n has that table's declared unit, not automatically a training-record count.
$x,\ n(x)$	Scalar cumulative cost x and number $n(x)$ of matched cases jointly observed there. This count can change along a curve; scalar x differs from the feature vector $x(h, q)$.
SD; W/T/L	Sample standard deviation and wins/ties/losses in each table's declared comparison direction. Paired SD uses $n_{\mathrm{pair}} - 1$. In W/T/L columns, T and L mean ties and losses only.
Feature and method names	
A,P,B,C; P+B,P+C	A: output/action fields; P: A plus public state; B/C: full-sequence/action-position bundles added to A. P+B/P+C add the corresponding bundle to P; each bundle has 128 MLP statistics and two auxiliaries.
Single / Two / P / P+C	Single uses one proposal; Two uses two with fixed ordering. P and P+C rank two proposals using public and augmented features, respectively. Data identifiers appear in Appendix B.

A.1 Abbreviations and Benchmark Names

Table 5: Abbreviations and names used in the paper. Mathematical symbols and feature-view codes are defined in Table 4.

Term	Meaning and use
AI; LLM	Artificial intelligence; large language model. The policy is the fixed Qwen3.5-4B checkpoint.
MLP; RMS	Multilayer perceptron; root mean square. We summarize MLP inputs and outputs by their mean and RMS.
MADE	*MAterials Discovery Environments*, the formal expansion in the benchmark paper's abstract (Malik et al., 2026). It evaluates sequential materials discovery.
DW; DW-P	DiscoveryWorld; our earlier DiscoveryWorld Proteomics/Normal cohort. DW-P is distinct from the later four-theme Easy/Normal study.
SUN	Stable, unique and novel discovery count, $D(B)$, under the configured benchmark stability criterion.
mSUN	Fraction of (meta)stable, unique and novel candidates: $D(B)/B$. The prefix refers to (meta)stability. MADE uses a stability tolerance of 0.1 electronvolts (eV) per atom here.
AUDC	Area under the discovery curve, normalized as $(2/B^2)\int_0^B D(\tau)\,\mathrm{d}\tau$.
Orb / ORB	The pretrained Orb-v3 interatomic-potential model used for energy evaluation; checkpoint `orb-v3-conservative-inf-omat`. ORB is the capitalization used in our cost tables, not a separate algorithm acronym.
AUROC	Area under the receiver operating characteristic curve; a measure of ranking discrimination.
NLL; BCE	Negative log-likelihood; binary cross-entropy. Both assess probability predictions under the binary failure target.
Brier score	Mean squared error of predicted failure probabilities; a named metric, not an acronym.
L-BFGS	Limited-memory Broyden–Fletcher–Goldfarb–Shanno optimizer (`lbfgs` in the implementation).
LR; CV	Logistic regression; cross-validation. Each predictor's regularization is selected using training folds.
SD; CI	Standard deviation; confidence interval. The reported bootstrap intervals have the task/seed scope stated with each analysis.
W/T/L	Wins, ties and losses of the stated method relative to its paired comparator.
JSON	JavaScript Object Notation, the format used for generated action objects.
API; RPC	Application programming interface; remote procedure call. RPC records identify actual environment or tool requests and responses.
CPU; FP32	Central processing unit; 32-bit floating-point arithmetic.
FLOP	Floating-point operation. Capture passes and cached generation steps are not equivalent FLOP counts.
RNG; ID	Random number generator; identifier. Matching scenario IDs and seeds need not restore identical hidden environment state.
T5	Text-to-Text Transfer Transformer. We use the cited ScienceWorld T5 agent's termination rule, not its training procedure.
K; B	Proposal count and environment budget. $K = 1$ for Single and $K = 2$ for Two, P and P+C. The units of B depend on the benchmark.

B Protocol, Implementation and Fitting

Arm identifiers and feature dimensions. Single, Two, P and P+C correspond to `Native1` (`Native` in DiscoveryWorld), `Common2`, `Public2` and `ActionPool2`, respectively. MADE A, P and P+C contain 15, 31 and 161 raw fields; DiscoveryWorld uses 10, 40 and 170. The $2d + 3$ transformation gives 325 fields for MADE P+C and 343 for DiscoveryWorld P+C. The frozen model has $L = 32$ captured MLP layers and input/output width 2560, yielding $32 \times 4 = 128$ activation statistics.

Table 6: Selected MADE inverse regularization and temperature, and fitted DiscoveryWorld temperature. The feature-view name C and the regularization parameter C_{reg} are different quantities. DiscoveryWorld temperature uses only world 1 calibration data.

View	MADE C_{reg}	MADE T	DW T
A	0.01	1.0000	1.0027
B	0.001	1.0000	1.3300
C	0.001	1.0000	1.5626
P	0.01	1.0000	1.4884
P+B	0.001	1.0000	1.3455
P+C	0.001	1.0000	1.5485

Capture validation and transfer. Feature extraction processes the same token sequence twice without gradients: a reference forward pass and a pass with observation hooks. Their final-position logits must agree within absolute tolerance 10^{-6} and relative tolerance 10^{-5}. This checks numerical agreement at the final output position; it does not verify equality of every intermediate activation. The implementation copies the full MLP input and output tensors before computing the summaries; both forward passes and the transfer time are included in the reported costs.

Preprocessing and fitting. Imputation uses each fitting column's finite training median, or zero if the column has no finite training value. Means and population standard deviations are computed after imputation; standard deviations below 10^{-8} are replaced by one. For standardized vector $\widetilde{x} \in \mathbb{R}^d$ and training-failure prototype v, the additional fields are $\sqrt{\|\widetilde{x} - v\|_2^2/d}$, $(\widetilde{x}^\top v)/\max(\|\widetilde{x}\|_2\|v\|_2, 10^{-8})$, and an indicator that no failure prototype was available. Missingness indicators refer to the original input before imputation. The transformation output uses 32-bit floating point (FP32). Fitting uses the limited-memory quasi-Newton optimizer `lbfgs`, an intercept, no class weighting, tolerance 10^{-8}, and at most 2000 iterations; convergence failures are not silently accepted. The original MADE and DiscoveryWorld six-view analyses each use 72 cross-validation fits and six final fits. MADE has no temperature fit. DiscoveryWorld has six calibration fits, with bounded log temperature, tolerance 10^{-6}, and at most 500 scalar-optimizer iterations.

The public vocabulary and failure prototype are fitted on the training portion of each fold. Held-out training groups select regularization; the final fit uses all declared training records. DiscoveryWorld calibration uses only world 1. Development records supply the reported diagnostics. Model checkpoint, feature order, transform, coefficients, temperature, and numerical-runtime identities are bound before evaluation. Saved calibration/development predictions are checked against the online predictor implementation. Such reproduction checks do not establish statistical calibration or generalization.

MADE's deterministic common key depends on task, budget, seed, decision index, executable tool/arguments, and candidate index. DiscoveryWorld orders by a hash of world seed, policy seed, budget, attempt index, and executable action, followed by prefix hash and candidate index. The same ordering resolves risk ties and implements the risk-independent control within a benchmark. A proposal's rationale can affect hidden features even when its executable action is unchanged. No outcomes are assigned to rejected proposals, and no environment-state counterfactual is inferred from their scores.

B.1 Benchmark Scope and Budgets

DiscoveryWorld cohorts. The short-budget Proteomics study and the later official-task evaluation are separate cohorts. The latter uses the four multitask test themes in Appendix C.2 of Jansen et al. (2024): Plant Nutrients, Space Sick, It's (not) Rocket Science!, and Lost in Translation. Easy and Normal each use the five official API seeds 0–4, policy seed 501, and four arms. Easy allows 100 post-load decision rounds and Normal allows 1000. The frozen risk heads were trained only on Proteomics, a subset of the proposed training themes. This is cross-theme transfer of the specified agent, not reproduction of the paper's GPT-4o agents or complete eight-theme, three-difficulty evaluation.

Algorithm 1 Frozen risk-guided scientific interaction

Require: Frozen policy π_{θ_0}; frozen feature transform and predictor; proposal count K; selector mode; native budget and feedback adapters

Initialize public history h and native budget counters
while the environment is nonterminal and the native budget permits a decision **do**
  if the native adapter prescribes a fixed non-policy recovery **then**
    Execute and log that recovery through the native adapter
    Update native counters and public history; **continue**
  end if
  for $c = 1, \ldots, K$ **do**
    Generate q_c from π_{θ_0} using h and prescribed feedback on earlier proposals
    Parse a_c; record generation, validity, and exact original tokens
  end for
  $S \leftarrow \{c : a_c \neq \perp\}$
  if $S = \emptyset$ **then**
    Apply the benchmark's existing rejection/recovery handler
  else
    if risk selection is enabled **then**
      For each $c \in S$, compute η_c and, when supported, the frozen score z_c
    end if
    if risk selection is enabled and every $c \in S$ has $\eta_c = 1$ **then**
      $c^* \leftarrow \arg\min_{c \in S}(z_c, \kappa_c)$
    else
      $c^* \leftarrow \arg\min_{c \in S} \kappa_c$
    end if
    Execute only a_{c^*} through the native environment adapter
  end if
  Update native counters and h using actual public feedback
  Log outcomes of executed actions; leave unexecuted outcomes unknown
end while

The official-task adapter follows the pinned Hypothesizer loop's post-load accounting: a known parse rejection or invalid action still consumes a round and advances one world tick. Initialization and scenario warm-up ticks are recorded separately. The original simulator random number generator (RNG) behavior is retained; matching API seeds does not guarantee identical object-level random states across arms. Only agent 0 is controlled, even when the official Space Sick Normal scenario exposes additional user-agent interfaces. Unknown execution outcomes remain incomplete rather than being replayed. Administrative time limits censor unfinished episodes; they do not create completed 100- or 1000-round results. Official task completion and normalized progress are reported separately from action-execution success. Explanatory knowledge scoring was not evaluated and is not assigned a zero.

C Prediction on Identical Examples

C.1 Development Comparisons and Nested Sensitivity

The original six-view development results compare predictors on the same rows within each benchmark. They are distinct from the later analysis that preserves the entire public transformation. Both remain development-only evidence.

Preserving the public feature transformation. The original P and P+C transformations agree on shared raw columns, standardized features and missingness indicators, as well as their medians, means, scales and failure-prototype coordinates. Adding C nevertheless recomputes distance and cosine in the expanded space. Both values change on all 350 MADE and 160 DW records, so the complete transformed P representation is not a subset of the original P+C representation. The no-prototype indicator is unchanged.

Table 7: Failure prediction on the same development rows within each benchmark: MADE has 100 rows from two chemistries; DiscoveryWorld has 40 records from two episodes on one world. These are development diagnostics. Higher AUROC and lower Brier/NLL are preferred. No row-level independence or significance claim is made.

View	MADE AUROC	Brier	NLL	DW AUROC	Brier	NLL
A	0.5275	0.2455	0.6841	0.9567	0.0768	0.2460
B	0.4508	0.2502	0.6938	0.9433	0.0753	0.2579
C	0.5590	0.2451	0.6832	0.9833	0.0844	0.3252
P	0.5467	0.2457	0.6858	0.9667	0.0620	0.2034
P+B	0.4814	0.2493	0.6921	0.9533	0.0718	0.2427
P+C	0.5610	0.2450	0.6830	0.9867	0.0772	0.2945

Table 8: Post-hoc development sensitivity with the complete public transform retained. Appended fields receive training-only standardization and missingness indicators. C_{reg} is selected within the original three training-group folds; DW temperature uses only its separate 60-row calibration split. MADE has 100 development rows across two chemistries; DW has 40 records from two episodes on one world. These newly fitted heads have not been evaluated online. NLL and Brier are averaged equally across MADE chemistries or DW episodes; lower is better. AUROC is pooled.

Benchmark	View	Dim.	Macro NLL	Macro Brier	AUROC
MADE	P	65	0.6858	0.2457	0.5467
MADE	P + aux	69	0.6853	0.2455	0.5508
MADE	P + MLP	321	0.6829	0.2450	0.5622
MADE	P + both	325	0.6829	0.2450	0.5610
DW	P	83	0.2034	0.0620	0.9667
DW	P + aux	87	0.2011	0.0609	0.9700
DW	P + MLP	339	0.2940	0.0771	0.9867
DW	P + both	343	0.2936	0.0771	0.9867

The post-hoc, development-only analysis fits new offline predictors while preserving all $2d_P + 3$ public coordinates. Appended fields contribute only standardized values and missingness indicators. For P, P plus auxiliaries, P plus MLP statistics and P plus both, respectively, the resulting dimensions are 65/69/321/325 for MADE and 83/87/339/343 for DW. None of the new predictors was evaluated online.

Re-evaluating the original float32 predictors in a different numerical runtime produces logit differences up to 1.91×10^{-6}, despite identical feature columns, example order and labels. No classification at probability 0.5 or ordering of the evaluated offline pairs changes. This is not bitwise reproduction of the original online runtime; its recorded predictions and validation checks remain unchanged. Each of the eight new predictors exactly reproduces its own saved predictions, and its complete public preprocessing block matches the original public transformation.

The four nested views are P, P plus the two auxiliary fields, P plus the 128 MLP statistics, and P plus both. For each view the full public transform is preserved; appended coordinates use training-only imputation, standardization and missingness indicators without new prototype summaries. Original chemistry/episode folds select regularization, and DW calibration rows alone select temperature. The analysis executes 104 logistic fits (96 cross-validation and eight final fits) plus four DW temperature fits; it makes no neural forward or environment call. All eight final heads have distinct saved identifiers. No new head replaces a predictor in an online result.

D Action Choice on Identical Candidate Pools

The comparison uses 4128 candidate pools from all 126 original MADE/DW-P P+C episodes: sixty MADE and 66 DW-P. It is conditional on the states and proposals recorded under P+C and does not generate alternative trajectories. All two-proposal decisions count toward the total denominator. An

eligible pair contains two valid, distinct executable actions supported by both predictors; differences in rationale or candidate index alone do not qualify. Selection retains the original support check over all valid candidates, the common fallback and the tie rule.

In MADE, action comparison uses the tool and arguments and ignores rationale. Online internal-feature values were not stored, so we recompute P scores but use saved P+C scores. Every original P+C selected index is reproduced. The calculation retains the original batch of 100 repeated rows rather than substituting a single-row matrix multiplication. This reproduces choices from the available scores; it does not recompute P+C scores from internal features.

For DW-P, each predictor scores a candidate using the original batch of 100 identical rows. All 3600 candidate P+C logits and probabilities exactly match their recorded values, and all 1800 original selected indices are reproduced. Action comparison omits fields ignored by the environment and treats absent and null values as equivalent only where the environment adapter does so. It assumes no state-dependent equivalence. These records come from the earlier Proteomics cohort, not the later Easy/Normal study.

The supplied records contain hashes of executable inputs, validity and support indicators, and both predictors' scores and selected indices, but no outcomes for unexecuted alternatives. A standalone Python script verifies eligibility, action differences, denominators and table totals, and checks that each selected score is minimal under the support and tie rules. The hashes distinguish executable inputs, not necessarily their effects in the environment.

Table 9: Action disagreement between P and P+C on 4128 candidate pools from 126 P+C episodes. Eligible pairs contain two valid, distinct executable actions supported by both predictors. All observed changes occur in eligible pairs; the same action-change count is divided by all decisions or eligible pairs. Higher rates mean more disagreement, not better outcomes. The comparison is conditional on P+C states and proposals.

Cohort	B	Decisions	Eligible	Changes	All (%)	Elig. (%)
MADE	10	408	134	54	13.24	40.30
MADE	20	785	249	99	12.61	39.76
MADE	30	1135	413	150	13.22	36.32
DW-P	10	200	72	14	7.00	19.44
DW-P	20	400	156	28	7.00	17.95
DW-P	30	600	258	45	7.50	17.44
DW-P	100	600	277	66	11.00	23.83

D.1 Support and Duplicates in the Executed Cohorts

Table 10 summarizes the P+C trajectories. The supplementary tables retain Two and P, the separate original Easy cohort, and every denominator. Support only means that the specified score can be computed. When both proposals encode the same executable action, there is no distinct action to choose between. Different executable inputs can still have the same environmental effect. None of these counts assigns an outcome to an unexecuted candidate.

Table 10: P+C support and action changes on its own recorded trajectories. Support uses all candidates for MADE/DW-P and valid candidates for standard DW. Duplicate denominators are two-valid pairs. Changes are relative to the first valid candidate (First) or the fixed Two ordering at that same pool; these comparators are not interchangeable.

Cohort	B	Decisions	Support	Duplicates	Changes	Relative to
DW-P	10	200	400/400	128/200	40/200	First
DW-P	20	400	800/800	244/400	71/400	First
DW-P	30	600	1200/1200	341/600	138/600	First
DW-P	100	600	1200/1200	323/600	105/600	First
MADE	10	408	490/816	38/371	139/407	First
MADE	20	785	987/1570	108/749	298/784	First
MADE	30	1135	1516/2270	156/1085	421/1135	First
DW Easy amended	100	2000	4000/4000	1157/2000	418/2000	Two
DW Normal	1000	20000	39999/39999	11043/19999	4557/20000	Two

E Scientific Outcomes, Heterogeneity and Failures

The original analysis includes 503 completed and validated episodes: 239 MADE and 264 DW-P. Eight additional score-shuffle episodes are compared with their matching original runs. The MADE four-arm study is one episode short of its planned total: the missing P result for Mg–Sn–Sr, B10, seed 501 is not assigned a value. Excluding its three available counterparts from the four-arm plots leaves 500 trajectories; the separate Single/P+C comparison retains twenty pairs at each MADE budget. These experiments cover two selected chemistries or worlds, not the full benchmarks.

All arms use frozen Qwen/Qwen3.5-4B, checkpoint revision `851bf6e806efd8d0a36b00ddf55e13ccb7b8cd0a`. Reported seed spread is conditional on this checkpoint and the fixed fitted heads.

Uncertainty and units of replication. Intervals use 2000 paired bootstrap resamples of policy seeds within each fixed chemistry or world. When a comparison pools budgets, all observations sharing a policy seed within a stratum are resampled together. The intervals are conditional on these chemistries/worlds and fitted heads; two strata cannot establish transfer to a task population. Table 11 reports P+C minus Single and P at every registered budget, regardless of sign; the numerical supplement contains all six pairwise arm contrasts. An interval that includes zero does not establish equivalence. An all-zero observed success sample produces an all-zero empirical interval; it does not show that success is impossible in the population. We make no retrospective power or confirmatory significance claim.

Within-stratum SDs in the numerical supplement use sample variance across policy seeds. Paired-difference SDs describe the spread of same-seed differences. Chemistry/world-specific means are separately shown, so cross-task heterogeneity is not mislabeled as seed spread. These comparisons do not treat individual actions as independent experimental replicates.

Table 11: Paired P+C minus the indicated comparator. W/T/L counts P+C wins, ties and losses. Intervals use 2000 policy-seed bootstrap resamples within each fixed chemistry/world; they do not quantify unseen-task generalization. No missing outcome is imputed.

Cohort	B	Versus	n	Metric	Δ	Conditional 95% CI	W/T/L
MADE	10	Single	20	SUN	0.1000	[-1.000, 1.350]	10/2/8
MADE	20	Single	20	SUN	0.3500	[-1.550, 2.650]	6/5/9
MADE	30	Single	20	SUN	-2.4000	[-4.850, 0.050]	7/2/11
MADE	10	Single	20	AUDC	-0.0050	[-0.120, 0.125]	11/1/8
MADE	20	Single	20	AUDC	0.0059	[-0.076, 0.111]	10/0/10
MADE	30	Single	20	AUDC	-0.0443	[-0.142, 0.053]	8/0/12
MADE	10	P	19	SUN	0.0000	[-0.686, 0.737]	6/5/8
MADE	20	P	20	SUN	0.1000	[-0.850, 1.150]	9/3/8

Table 11 (continued)

Cohort	B	Versus	n	Metric	Δ	Conditional 95% CI	W/T/L
MADE	30	P	20	SUN	-0.7500	[-3.150, 1.550]	8/2/10
MADE	10	P	19	AUDC	-0.0347	[-0.112, 0.039]	6/2/11
MADE	20	P	20	AUDC	-0.0040	[-0.052, 0.048]	8/0/12
MADE	30	P	20	AUDC	-0.0031	[-0.076, 0.067]	12/0/8
DW-P	10	Single	20	Progress	0.0375	[0.000, 0.075]	5/14/1
DW-P	20	Single	20	Progress	0.0375	[0.006, 0.075]	6/13/1
DW-P	30	Single	20	Progress	-0.0125	[-0.056, 0.031]	4/11/5
DW-P	100	Single	6	Progress	0.1458	[0.083, 0.208]	5/1/0
DW-P	10	P	20	Progress	0.0063	[0.000, 0.019]	1/19/0
DW-P	20	P	20	Progress	0.0000	[-0.019, 0.019]	1/18/1
DW-P	30	P	20	Progress	-0.0250	[-0.050, 0.000]	1/14/5
DW-P	100	P	6	Progress	0.0000	[-0.083, 0.083]	3/0/3

Table 12: Legacy DW-P execution-success percentages on all completed arm cohorts. The final column gives full task completions per arm. At B10/B20/B30 each arm has twenty trajectories; B100 has six (two worlds, seeds 501–503). This legacy request/tick accounting is separate from the later official-task adapter.

B	n/arm	Single	Two	P	P+C	Completed/arm
10	20	36.00	47.00	58.00	53.50	0/20
20	20	38.25	48.75	58.00	59.50	0/20
30	20	39.33	49.00	62.50	62.67	0/20
100	6	36.50	62.17	71.67	66.50	0/6

E.1 Variation Across MADE Chemistries

Table 13 retains the primary within-chemistry P+C–P differences. Complete arm means, within-arm seed spread and P+C–Single differences remain in the numerical supplement. The two chemistries were previously examined during development; repeated seeds do not establish transfer to new chemistries.

Table 13: P+C minus P within each MADE chemistry. SD is the sample standard deviation of paired policy-seed differences; n_{pair} is the observed number of pairs. All signs are retained.

Chemistry	B	n_{pair}	ΔSUN	SD	ΔAUDC	SD
Au-K-Tb	10	10	-0.30	0.82	-0.0650	0.1096
Mg-Sn-Sr	10	9	0.33	2.40	-0.0011	0.2322
Au-K-Tb	20	10	1.60	2.46	0.0905	0.1447
Mg-Sn-Sr	20	10	-1.40	2.37	-0.0985	0.0925
Au-K-Tb	30	10	-0.90	4.18	0.0092	0.1427
Mg-Sn-Sr	30	10	-0.60	7.03	-0.0153	0.1953

E.2 Sensitivity to a Separate Technical Recovery

After the original 503-episode analysis, we conducted one separately preregistered recovery attempt for the missing MADE case: Mg–Sn–Sr, B10, seed 501, with the public-feature selector. The first attempt had stopped because an action-position check compared tuple and list representations. Its recorded calls and process shutdown were verified before the new attempt. The recovery used the corrected implementation with the same frozen model, risk predictor, task, budget and seed, and allowed no further retry based on the outcome. It obtained SUN 6, mSUN 0.6 and AUDC 0.72. This result is reported separately from the original failure.

Table 14: Separate B10 technical-recovery sensitivity. The amended intersection adds the new P result and its three already completed counterparts; no baseline was rerun. B20 and B30 are unchanged. Original tables and curves retain their original case masks.

Cohort	n	Metric	Single	Two	P	P+C
Original	19	SUN	4.58	4.16	4.63	4.63
Amended	20	SUN	4.75	4.40	4.70	4.85
Original	19	AUDC	0.4826	0.4732	0.5074	0.4726
Amended	20	AUDC	0.5015	0.4960	0.5180	0.4965

The amended dataset has 240 completed MADE trajectories and sixty complete four-arm cases, compared with 239 trajectories and 59 cases originally. Only the B10 comparison changes. Using the same 2000 paired bootstrap resamples of policy seeds within each chemistry, amended B10 P+C minus P is $+0.15$ SUN (conditional 95% interval $[-0.60, 0.90]$) and -0.0215 AUDC ($[-0.0970, 0.0535]$). Both intervals include zero, so the amendment does not change the conclusion that the added features do not consistently improve scientific outcomes.

The original interrupted attempt and the clean recovery both retain their incurred costs in Appendix F and the numerical supplement. The recovery does not replace an unknown original outcome.

E.3 Completed Registered Standard DiscoveryWorld Cohorts

All eighty planned Normal episodes completed, giving twenty cases with all four arms. Easy has 78 completed original episodes and two interrupted P runs. Two clean recovery attempts were registered and completed separately. We report the original eighteen-case comparison and the amended twenty-case comparison separately. The interrupted runs and their known generation, action, tick and RPC costs remain in the record; the recovery attempts do not replace their unknown outcomes.

Each complete cohort covers four official themes, five API seeds per theme and one policy seed under the specified frozen agent. It is not the full original DiscoveryWorld benchmark or a reproduction of its original agents. All arms are matched on the same cases within each row. Normal P+C minus P is -1.5986 percentage points; an exploratory four-theme cluster interval is $[-8.4444, 2.5397]$. Normal Two has one task completion in twenty cases; the other three arms have none. Easy has no task completions in any arm. Scores, task completion and action success remain distinct, and unevaluated knowledge scoring is not assigned zero.

Table 15: Standard DiscoveryWorld snapshot (25 September, 19:05 CST). Official task scores are multiplied by 100. Each row uses the same complete cases for all four arms. Easy original preserves two censored cells; Easy amended includes two separately registered clean attempts. Normal has 80/80 completed episodes.

Cohort	n	Single	Two	P	P+C
Easy original	18	15.97	27.08	23.38	26.62
Easy amended	20	17.50	26.25	25.42	28.96
Normal complete	20	22.42	27.58	24.16	22.57

Table 16: Official task completions and mean per-episode action-success percentages on the same standard-DW cases as Table 15. Entries are task completions / action-success percentage. A parser rejection is an unsuccessful request for this descriptive action-rate denominator, not a labeled official dispatched action for risk-model evaluation.

Cohort	Single	Two	P	P+C
Easy original	0/49.06	0/50.44	0/49.44	0/54.50
Easy amended	0/49.10	0/49.00	0/49.05	0/53.80
Normal complete	0/32.02	1/34.53	0/37.76	0/35.91

Theme-level primary contrasts. The amended complete-case P+C–P score difference is $+3.5417$ percentage points, with five positive, twelve tied and three negative case differences. An exploratory exact resampling distribution over four theme clusters gives the interval $[-8.125, 15.000]$. The separate original eighteen-case contrast is $+3.2407$, with interval $[-9.2105, 17.6471]$. These intervals apply to the selected four themes; five API seeds and one policy seed per theme do not establish broad task or policy-seed generalization. Table 17 reports the variation across themes for both Easy and Normal, retaining favorable and unfavorable differences.

Table 17: P+C minus P by official DiscoveryWorld theme, in score percentage points. Each row uses five matched API seeds. Easy is the amended cohort; the original eighteen-case result is reported separately. Normal uses its original completed cohort. These eight rows summarize the same cases as the aggregate table.

Difficulty	Theme	n_{pair}	P	P+C	Δ
Easy amended	Plant Nutrients	5	45.00	65.00	+20.00
Easy amended	Rocket Science	5	46.67	33.33	-13.33
Easy amended	Space Sick	5	10.00	17.50	+7.50
Easy amended	translation	5	0.00	0.00	+0.00
Normal	Plant Nutrients	5	13.33	15.56	+2.22
Normal	Rocket Science	5	51.43	54.29	+2.86
Normal	Space Sick	5	7.89	8.42	+0.53
Normal	translation	5	24.00	12.00	-12.00

F Computational Cost and Exploratory Diagnostics

F.1 Cost Accounting

Generation tokens, reference/capture forwards, dispatched environment actions and explicit world ticks are separate quantities. Table 18 consolidates the directly recorded standard-DW totals. The numerical supplement supplies per-arm and per-budget costs for all cohorts, including MADE and DW-P, and the corresponding observed-cost curves. No cost has been removed from the underlying records. The original failed MADE P attempt incurred 96 initialization ORB evaluations, one tool attempt, three returned generations and 591 sampled-token steps. Its complete inference totals remain unknown. The separate recovery incurred 96 initialization and ten candidate ORB attempts, 17 tool attempts and 73,360 returned prompt-plus-completion tokens, without capture forwards.

Table 18: Recorded cohort totals, not compute-matched efficiency estimates. DW rows count explicit post-load actions and ticks. The original and amended Easy cohorts overlap and must not be summed. Episode totals exclude unmeasured installation/shared-loading work and interrupted attempts listed below. A dash means that the count is not reported here.

Cohort	Episodes	Dispatches	Explicit ticks	Prompt tokens	Output tokens	Capture/ reference
DW Easy amended	80	8000	8000	53,852,402	274,984	8000
DW Normal	80	79,523	79,523	555,583,173	2,413,521	79,998

The two original interrupted Easy attempts additionally incurred 768,168 known prompt tokens and 4365 output tokens; one generation and one RPC did not return. Unknown costs are not zero and are not replaced by clean-recovery expenditure. The supplement retains the full cost records and distinguishes time measured within each episode from time measured for the whole job.

Cost curves use the last observed score within each declared joint cost domain and never extend an episode past its endpoint. In standard DW, the number of complete cases $n(x)$ can decrease at scalar cost x, changing the task mixture; x is not the feature vector $x(h, q)$. MADE/DW-P curves use their fixed shared observed domain. Wall time depends on uncontrolled hardware, load and runtime overhead. These are descriptive profiles and do not establish a causal speedup.

F.2 Prediction on Each Selector's Own Trajectories

The two selectors execute different actions and visit different states, so their own-trajectory AUROC and Brier scores do not measure additional predictive information on identical examples. The full tables are provided in the Numerical Supplement, Section 4 (Prediction Quality and Action Selection), together with NLL, episode-macro metrics and exploratory intervals. MADE uses stable-and-new discovery labels, and DW uses known execution responses. Unexecuted alternatives remain unlabeled.

F.3 Exploratory Score-Shuffle Diagnostic

Eight registered MADE episodes cover B10/B20, two chemistries and seeds 501/502: four matched pairs per budget. The control retains generation, capture, support and fallback, but independently permutes the two whole P+C scores. It does not selectively remove internal features. Shuffle minus P+C is $+0.0500/+0.1250$ mSUN and $-0.0100/+0.0875$ AUDC at B10/B20. SUN wins/ties/losses are $2/2/0$ and $3/0/1$, respectively. The budgets are reported separately, without a pooled efficacy estimate. These small conditional comparisons establish neither broad score-ordering efficacy nor that internal information is useless. The numerical supplement retains both budgets, every metric and conditional interval, all eight cases and measured costs. Original P comparisons have seven complete cases; the separately registered recovery is a sensitivity analysis (Appendix E.2).

G ScienceWorld: Prediction and Action-Selection Diagnostics

This appendix uses ScienceWorld (Wang et al., 2022) to examine RQ1 and recorded action-selection behavior. It evaluates prediction of a specified response-message label and differences between selected actions; it does not establish scientific utility. All results below concern the original P and P+C predictors. P uses public-state and generated-output features, including generation-probability summaries; P+C adds the original 128 action-position MLP statistics and two auxiliary measurements. Each predictor retains its own fitted preprocessing and head. This comparison does not isolate the MLP statistics from the auxiliaries, feature transformation, or fitted predictor.

G.1 Recorded Action Protocol and Label

The registered cases contain two official test variations from each of ten task families. Each case was run at native budget parameters $B \in \{10, 20, 30, 100\}$ under Single, Two, P and P+C, with policy seed 924501. Budgets use separate rollouts, not prefixes of one trajectory. The common-example

and common-pool analyses below use all eighty P+C rollouts: twenty cases at each budget. These are twenty case identities, not eighty independent tasks, and there is no replication across policy seeds.

The frozen Qwen3.5-4B policy uses sampling temperature 0.6, top-$p = 0.95$, top-$k = 20$, a 4096-token input cap and 128-token generation cap, with thinking disabled. The implementation uses FP32 and eager attention, with embedding and output head on the CPU. Inputs retain the public observation, look, inventory, task description, budget counters and the last eight actual interactions. The second proposal also sees the first packet and a fixed alternative instruction. There are exactly two sequential proposals in the relevant two-proposal arms; only one selected packet is dispatched. Environment rewards, oracle valid-action lists and hidden evaluator state are excluded from online inputs.

The native step-limit parameter is B, with a separate guard of $2B$ attempted action requests. This follows the official T5 termination profile, not the complete T5 agent or its training. The pinned ScienceWorld 1.3.0 wrapper checks $\mathrm{moves} > B$ after dispatch. Invalid, clarification and free actions need not advance this counter, while waits can advance it multiple times. A malformed packet still consumes a request. These rules determine the observed action sequences; B is not an exact count of dispatched actions.

For a verified native response, the frozen `parser_rejection_v1` target is a fixed-message proxy. After surrounding whitespace is stripped, $y = 1$ only for `No known action matches that input.` or a full match to the registered `ERROR: Unknown response` / `Value out of range` ambiguity-response pattern ending in `Action cancelled.` Other verified messages receive $y = 0$; locally rejected malformed packets and unknown transitions have no label. The rule does not exhaust native action errors: for example, `Unknown action.` is not one of its recognized messages. A zero proxy label therefore does not establish that an action was valid or executed successfully. The exact matcher is distributed with the figure's source data.

The action codec identifies the original JSON `action` string through `/action`. Its token IDs are not retokenized. For each action-value token at position j, the extractor uses the state at $j-1$, pooling MLP input/output mean and root-mean-square statistics over those positions and hidden coordinates. Thirty-two layers yield 128 statistics; final-position next-token entropy and the maximum absolute reference/capture-logit difference supply the two auxiliary fields. Reference and capture passes run without gradients, with the original final-logit parity tolerances of 10^{-6} absolute and 10^{-5} relative. Parser validity, action-token alignment and predictor support are distinct from the native response label.

G.2 Training, Calibration and Frozen Predictors

Forty fitting variations and twenty calibration variations use the official training split and are disjoint from the test cases in Table 20. Official development variations are unused. Preparation uses the single-proposal agent and records internal features for aligned actions. Each variation runs once; each budget receives ten fitting and five calibration episodes. For budget order $(10, 20, 30, 100)$, the fitting-budget index is the sum of the zero-based task and within-role variation indices modulo four; calibration doubles the task-index term.

The sixty preparation episodes yield 2466 fitting and 1305 calibration records, with 1008 and 631 positive proxy labels, respectively. Both predictors use the same 3771 records; capture requires 7542 reference/capture forward passes. Installation and model checks are separate from these preparation records.

P contains 31 numeric fields and a training-only vocabulary of twenty action words plus an unknown category, giving 52 raw fields. P+C has 182 raw fields. Vocabulary, imputation, scaling and failure prototypes are fitted within each training fold. Preprocessing is unweighted; logistic fitting and its negative log-likelihood (NLL) objective use episode weights $n_{\mathrm{train}}/(En_e)$. The transformed dimensions are 107 and 367. The maps are view-specific: adding C recomputes the prototype distance and cosine in the larger space.

Each view uses four folds, holding out one of each task's four fitting variations per fold. The inverse-regularization grid is $\{0.001, 0.01, 0.1, 1\}$; selection minimizes NLL averaged equally across held-out episodes, with smaller values breaking exact ties. The solver is `lbfgs`, with an intercept, no

class weighting, tolerance 10^{-8} and at most 2000 iterations. This gives 32 cross-validation fits and two final fits. Each view then fits one positive temperature on its separate calibration records by minimizing episode-weighted NLL over $\log T \in [-4, 4]$. Calibration-set NLL is not held-out predictive evidence. No predictor or temperature is fitted separately by evaluation budget.

Table 19: The two original frozen ScienceWorld predictors. The same saved head is used at every evaluation budget; these are fitting settings, not evidence of predictive performance. Full head hashes, feature order and vocabulary are retained in the accompanying source metadata.

View	Raw fields	Transformed	C_{reg}	T	Head SHA-256 prefix
P	52	107	1	1.9113	`3bfe169e49b4`
P+C	182	367	1	2.4233	`2c70055cb165`

The source, model identity, feature schema, finite coefficients and available training/calibration support at every budget were checked before evaluation. The original head files identify the same `parser_rejection_v1` target. Prediction uses the saved transformation, coefficients and positive temperature; no evaluation-time fitting occurs. Risk ranking requires a nonempty valid set with every valid candidate supported. Otherwise the common ordering applies to all valid candidates, or the benchmark handler applies to an empty set. The fixed ordering also resolves score ties. The P selector checks alignment without capturing activations, so a capture-record flag is not a measure of its predictive support.

Table 20: Fixed ScienceWorld variations, preserving the registered order used in preparation-budget assignment. Fitting and calibration use the official training split; evaluation uses the official test split. The names are task descriptions, not action commands.

ID	Task	Fitting variations	Calibration	Test
1-2	Melting	2, 7, 11, 8	5, 3	27, 24
2-1	Temperature measurement	60, 72, 30, 236	126, 230	464, 520
3-4	Unknown-substance conductivity	258, 228, 219, 170	110, 40	590, 479
4-2	Identify nonliving objects	149, 148, 57, 56	9, 23	261, 243
5-1	Grow a plant	51, 29, 14, 47	43, 27	95, 124
6-1	Mixtures	11, 8, 6, 0	9, 3	24, 29
7-3	Lifespan comparison	19, 32, 51, 42	44, 52	108, 115
8-1	Life-stage identification	5, 4, 1, 3	2, 0	13, 12
9-3	Inclined-plane friction	9, 28, 20, 68	46, 12	137, 138
10-2	Plant genetics	182, 238, 213, 125	235, 60	443, 364

G.3 Prediction on Identical Executed Actions

This post-hoc analysis applies both original fixed predictors to the same 4754 executed actions with known labels from the eighty P+C rollouts. Both predictors can score every record, so none is excluded. No fitting or calibration label enters this analysis. Recomputed P+C logits and probabilities exactly match their recorded values. Public inputs, action tokens and labels remain unchanged, and the replay requires no new language-model inference, environment interaction, fitting or calibration.

Table 21 compares probability error and discrimination on identical examples. Brier measures probability-prediction error, not calibration alone; AUROC is the area under the receiver operating characteristic curve. These comparisons concern the fixed response-message proxy on actions selected by P+C. They do not evaluate actions or histories that P would have produced in a different rollout.

The pattern varies across budgets. At B30, AUROC is 0.8018 for P+C and 0.7290 for P; at B100, P+C reduces Brier by 0.0272. At B10, its Brier is instead higher by 0.0631. All four budgets and both directions are retained in Figure 4 and Table 22.

The interval procedure was specified after observing the point estimates. It uses 10,000 paired bootstrap replicates under two schemes: resampling the two episodes within each of ten fixed task

Table 21: Fixed original P and P+C predictors on the same labeled, executed actions. Every row uses twenty P+C trajectories on the same twenty test variations; n counts actions, not independent trials. The target is the specified rejection-message proxy.

B	n	$y = 1$	Brier P	Brier P+C	AUROC P	AUROC P+C
10	345	104	0.0650	0.1281	0.9362	0.8596
20	606	118	0.1064	0.0942	0.7250	0.7940
30	970	226	0.1274	0.1225	0.7290	0.8018
100	2833	779	0.1373	0.1101	0.7863	0.7974

families, or resampling ten task families while retaining each family's two episodes. Individual actions are not independently resampled. The intervals are exploratory and unadjusted for multiple comparisons. They reflect neither repeated policy seeds nor new task families. B20 conclusions depend on the resampling scheme; an interval including zero does not establish equivalence.

Table 22: P+C minus P on identical executed actions. Both interval schemes use 10,000 paired resamples; the analysis is exploratory and unadjusted. Lower Brier and higher AUROC are preferred. The two schemes are sensitivity analyses of the same data, not independent replications.

B	Metric	Difference	Within-family episode CI	Task-family CI
10	Brier	+0.0631	[0.0047, 0.1230]	[0.0040, 0.1491]
10	AUROC	-0.0766	[-0.1578, 0.0118]	[-0.1618, 0.0216]
20	Brier	-0.0121	[-0.0233, -0.0012]	[-0.0225, 0.0003]
20	AUROC	+0.0689	[0.0463, 0.0927]	[-0.0013, 0.1076]
30	Brier	-0.0050	[-0.0152, 0.0057]	[-0.0227, 0.0129]
30	AUROC	+0.0728	[0.0253, 0.1215]	[0.0240, 0.1238]
100	Brier	-0.0272	[-0.0468, -0.0057]	[-0.0467, -0.0100]
100	AUROC	+0.0111	[-0.0793, 0.0953]	[-0.0653, 0.1144]

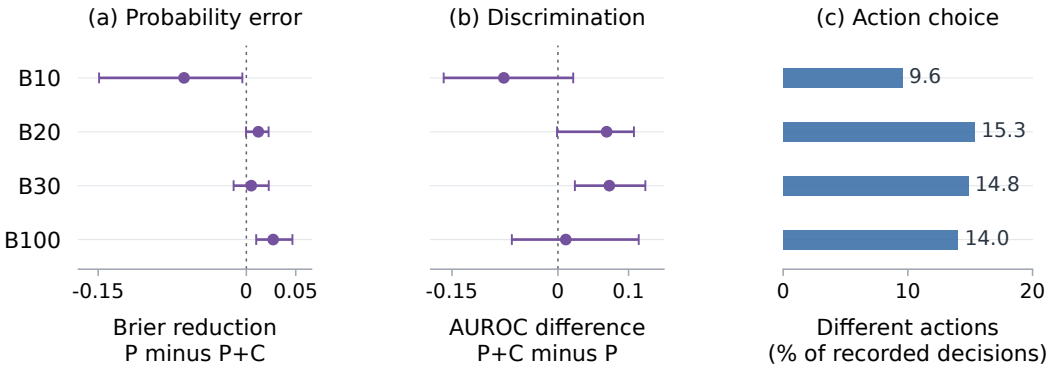

Figure 4: ScienceWorld action-level comparisons across all four budgets. Panels (a,b) evaluate the same labeled P+C-executed actions: (a) Brier reduction, P minus P+C; (b) AUROC difference, P+C minus P. Positive values favor P+C's prediction. Bars in these panels are exploratory, unadjusted 95% paired task-family bootstrap intervals. Panel (c) reports the percentage of recorded pools where P and P+C select different executable action strings; larger values mean more disagreement, not better actions. The figure contains no measure of scientific utility.

G.4 Choices on Identical Recorded Candidate Pools

The comparison uses the same 4754 P+C candidate pools, retaining the original support, fallback and tie rules. Both fixed predictors score each stored proposal with the original inference implementation. All 9508 P+C candidate logits and probabilities exactly match the saved values, and all 4754 original P+C selected indices are reproduced. Comparison uses the exact action string sent to the environment, not the candidate index or rationale.

An eligible pair contains two parser-valid, distinct executable strings supported by both predictors. Table 23 reports disagreement using both the all-decision and eligible-pair denominators. Table 24 separately gives feature support and duplicates; predictor support does not certify the native action's success or the predictor's calibration. The complete action-mechanism records for the four original arms remain in the figure's source data, but the matched comparisons here concern P+C's recorded state distribution.

Table 23: Action disagreement on the same recorded P+C pools. Eligible pairs contain two distinct valid action strings supported by both predictors. The same number of different selections is divided by all decisions or by eligible pairs; a higher rate means more disagreement, not better actions.

B	Decisions	Eligible pairs	Different actions	All (%)	Eligible (%)
10	345	176	33	9.57	18.75
20	606	344	93	15.35	27.03
30	970	528	144	14.85	27.27
100	2833	1698	397	14.01	23.38

Table 24: Support and duplicates in the P+C trajectories used above. Validity means acceptance by the action-object parser; support means the fixed predictor can score a candidate. Neither certifies successful environment execution. Duplicate pairs have identical executable strings.

B	Valid / generated candidates	Supported / valid candidates	Duplicate / two-valid pairs
10	690/690	690/690	169/345
20	1212/1212	1212/1212	262/606
30	1940/1940	1940/1940	442/970
100	5666/5666	5666/5666	1135/2833

The predictors select different strings in 9.57–15.35% of recorded decisions across budgets. Duplicates explain why a changed candidate index can leave the executable action unchanged; distinct strings can also have the same environmental effect. These are diagnostics of prediction and selection opportunity. No unexecuted candidate is assigned an outcome, and neither a lower message-proxy risk nor a different selected action is treated as evidence of improved scientific utility.