

Same Algorithms, Different Scientist: Design Space Bias in AI Discovery



*Odysseus Kaloidas*¹ 

¹ University of California, Berkeley (berkeley.edu)

 29 Aug 2026 (modified: 14 Sept 2026)

 NeurIPS 2026 Workshop AI4MetaScience Submission

 AI4MetaScience, Reviewers, Authors  Revisions  CC BY 4.0

Edit 

Keywords: AI scientists, scientific discovery, AI for science, evaluation, design spaces, representation bias, behavioural equivalence, protein engineering, LLM agents.

TL;DR: We introduce the executable behavioural quotient and show that merely rewriting an unchanged AI-scientist design space can reverse choices, credit, and model rankings—exposing a hidden evaluation confound.

Abstract:

AI-scientist evaluations treat human-written design spaces—menus of methods, code templates, or configuration grammars—as the opportunities a model could discover. We ask whether measured scientific competence is invariant to rewritings that leave every executable option unchanged. We introduce the executable behavioural quotient, which groups written policies whose execution on frozen probe states is identical, and use it both as an invariance test and as a unit of discovery credit. Across 12 combinatorial protein-fitness landscapes, 40–41 executable policies collapse to 21–22 distinct campaigns; 19 written options implement one algorithm, and 35% of the menu is behaviourally redundant. Nevertheless, equivalent rewritings move model choices by a median total-variation distance of 0.417, against a 0.144 same-prompt replicate floor, and reverse 56 of 255 credit verdicts. A separately sealed knapsack grammar reproduces order sensitivity across 24 permutations and 6 problem families. Mechanism analyses identify separable positional and textual biases: probability falls with menu position even after controlling for utility, while the shared textual ordering is negatively associated with utility. Quotient normalisation removes the representation instability but does not improve discovery. Scientific-agent evaluations should therefore define search, comparison, and credit over executable behavioural classes rather than the written menu that names them.

Track: Technical Track

Submission Type: Novel work (double-blind)

Ai Usage Disclosure: Under human direction, language models supported literature mapping, hypotheses, protocol/code drafting, analysis, review and editing. The authors approved sealed protocols and thresholds before calls, audited artefacts and claims, and take responsibility. No human subjects, personal data or wet-lab experiments were used; campaigns query published complete landscapes. The appendix records protocols, validity gates and limitations.

Review Study Opt In: Yes, I opt in to the study

Email Sharing: We authorize the sharing of all author emails with Program Chairs.

Data Release: We authorize the release of our submission and author names to the public in the event of acceptance.

Submission Number: 47

Filter by reply type...

▼

Filter by author...

▼

Search keywords...

Sort: Newest First

▼

≡

≡

≡

≡

≡

≡

≡

≡

≡

-

=

≡

🔗

👁

Everyone

Program Chairs

Submission47 Reviewers

Submission47 Authors

Submission47...

Submission47...

✕

5 / 5 replies shown

Add: **Withdrawal**



Paper Decision [🔗](#)

Decision by Program Chairs  29 Sept 2026, 21:42 (modified: 30 Sept 2026, 12:06)

 Program Chairs, Reviewers, Authors  Revisions

Decision: Accept (Poster)



Controlled evaluation of algorithm presentation and model choices

Official Review by Reviewer jPF_e  18 Sept 2026, 10:31 (modified: 30 Sept 2026, 12:06)

 Program Chairs, Reviewers, Authors  Revisions

Review: Summary

The authors test how the presentation of algorithm menus affects language models' choices. In the protein experiments, algorithms belong to the same group if they select identical sets of protein variants on the same fixed test states. The baseline samples these groups with uniform probability. Rewriting the menu can reverse whether a model outperforms this baseline.

Strengths

S1. The controls help distinguish menu effects from variation between repeated calls. The authors keep the behaviour groups and the rest of the prompt fixed, and repeat the original prompt without changes. Menu rewritings reverse whether a model outperforms the random baseline more often than those repeated calls do.

Weaknesses

W1 [MAJOR]. The invariance claim needs qualification. The authors interpret a maximum Representation invariance score (RIS) as "choice depends only on available algorithms". Yet RIS uses the median change in the model's choice probabilities across equivalent menu rewritings. Please clarify that if more than half of the rewritings leave the model's choice probabilities unchanged, the median is zero and RIS reaches its maximum even if other rewritings change them.

W2 [MAJOR]. Explain the gain from averaging. The authors report a gain from averaging each model's choice probabilities across menu arrangements. Please explain how the authors select an algorithm from the averaged probabilities and calculate the score for the "expected single arrangement". If algorithm scores are fixed and both averages use the same arrangements and weights, the expected score from the averaged probabilities equals the average of the expected scores for the individual arrangements.

Rating: 1: Accept (poster quality)
Confidence: 4: The reviewer is confident but not absolutely certain that the evaluation is correct
Full Text Attestation: I confirm that I have read the full main text of this submission, and I take full responsibility for this review, including any parts produced with AI assistance.

This is a nice work with practical motivation; it provides a perspective on a real challenge in agent evaluation and demos a solution through controlled experiments.

Official Review by Reviewer wqVz 17 Sept 2026, 05:06 (modified: 30 Sept 2026, 12:06)

Program Chairs, Reviewers, Authors Revisions

Review:

Overview. This paper investigates a real challenge in agent evaluation: when the presentation of choices change, whether that affect the evaluations of the agent's scientific competence and how that difference can be mitigated. Both this challenge and the proposed solution are cleanly formulated demonstrated. The paper proposes and evaluates a behavioral quotient normalization (BQN) and representation ensembling as the mitigation strategies. The practical implication to the model evaluation folks is to consider a preflight before evaluation: auditing candidate policies by execution before model selection, normalizing the menu over distinct behavioral classes, and measuring robustness to equivalent presentations.

Strengths

1. The behavioral quotient is a useful unit for comparing executable opportunities and assigning the discovery credit. This is supported by the experiments in the tested setting (small menu, empirically demonstrated equivalence, a complete landscape of experiment outcomes).
2. There is a good distinction between measurement stability and discovery performance. The paper correctly acknowledges that the normalization and ensembling can stabilize evaluation without improving scientific outcomes. There's also explicit reporting of instrument validation limitations and unsuccessful discovery improvements, which will help the readers understand the boundaries of the findings.
3. The challenge is real and a practically motivated question. The paper develops it well with controlled experiment formulation.

Weaknesses

1. BQN changes the model-facing task, which further complicates the interpretation of normalized performance.

Preserving the tested executable opportunities does not preserve the process of selecting among them; Removing redundant entries also removes the cognitive burden of recognizing that differently written configurations are operationally equivalent. The normalized performance might have excluded a relevant capability when evaluating an agent navigating a realistic method catalog. The paper might benefit from showing the performance gain from the preflight, or explicitly distinguishing these evaluation objectives.

2. The audience will benefit from some operational guidance to apply the concepts in broader scientific-agent evaluation.

In more general workflows, an agent could compose choices across stages, revisit earlier decisions, or modify workflow order and branching. Behavioral equivalence can then depend on the current state and downstream composition, while the enumeration of complete workflows may underrepresent that. The authors acknowledge the scope of BQN, but could connect these restrictions more directly to when BQN can be used.

Rating: 1: Accept (poster quality)

Confidence: 4: The reviewer is confident but not absolutely certain that the evaluation is correct

Full Text Attestation: I confirm that I have read the full main text of this submission, and I take full responsibility for this review, including any parts produced with AI assistance.

Ai Assistance Disclosure: ChatGPT-6 Astra was used for checking references and improving wording. Opinions are my own.

-

=

≡

Questions by AI Reader

Questions

 by AI Reader  09 Sept 2026, 13:08 (modified: 14 Sept 2026, 16:13)
 Program Chairs, Authors, Reviewers Submitted, Reviewers AI Treatment  Revisions

Questions For Authors:

1. Section 3.3 (Page 6) states that shared textual attractiveness decomposes into menu position and a jargon score, directing the reader to Appendix C for the decomposition, but the section under Appendix C (Page 11) contains no text. How is the jargon score defined and calculated, and what are the specific quantitative relationships for the alias count and text length mentioned in the main text?
2. Section 2 (Page 3) and Appendix A (Page 11) describe the harness and task structure but do not provide the exact prompt phrasing presented to the models, noting instead that the harness is unchanged from a prior anonymized version of the work. Given that Section 3.3 (Page 6) identifies strong textual preferences in the models' choices, how is the goal of the discovery task explicitly communicated in the prompt, and is it possible that this phrasing interacts with the models' sensitivity to specific jargon?

Disclaimer: *This is not a peer review, but a tool to help clarify potential points of confusion for first-time readers. Authors are encouraged to stay true to their own work, but can use this opportunity to clarify any genuine misconceptions without feeling pressured to answer every question. Your responses will be made available to the reviewers starting from **September 15th**.*

-

=

≡

Author Response to Questions by Authors

Author Response to Questions

 by Authors ( Odysseus Kaloidas) 14 Sept 2026, 22:28 (modified: 14 Sept 2026, 22:42)
 Program Chairs, Reviewers Submitted, Authors, Reviewers AI Treatment
 Revisions

Author Response:

Thank you for identifying these points. We clarify the operational definition and task wording below. The distinction is between the descriptive feature analysis, which helps characterize model preferences, and the controlled invariance experiments, which establish the paper's main result.

Jargon score and textual attractiveness

Appendix C should have included the operational definition alongside the associated figure, which appears on the following page. We provide that definition here.

The jargon score is an unweighted count of the following 22 vocabulary entries present in an option's lowercased canonical description: gaussian, process, ensemble, bayes, posterior, thompson, ucb, acquisition, rollout, lookahead, adaptive, expected, improvement, boosted, pca, latent, hierarch, calibrat, surrogate, uncertainty, entropy, diverse.

Matching uses substrings, so "hierarch" and "calibrat" also match their inflected forms. Each vocabulary entry contributes at most one point per description, regardless of repetition. The score is not normalized by description length and uses neither measured fitness nor campaign outcomes. It is a post hoc descriptive feature, not a learned quality score or a criterion used to select policies.

Shared attractiveness is calculated at the executable behavioural-class level. Within each landscape, we take each model's log class-selection probabilities, center them across classes, and average the centered values across models. Probabilities are floored at 10^{-6} before taking logarithms. We then compute within-landscape Spearman correlations between this shared attractiveness and each feature, and report their mean across the 12 landscape conditions.

The resulting mean correlations are +0.603 for jargon score, -0.671 for canonical menu position, -0.069 for alias count, and -0.028 for description length. Description length is the number of characters in the canonical description. Alias count is the number of written entries representing the same executable class in the alias-expanded menu.

Thus, jargon and position show substantially stronger descriptive associations than these particular measures of length and alias count. These correlations are not a causal decomposition or estimates of independent feature effects. Moreover, the 12 landscape conditions share three distinct task-prompt groups and should not be interpreted as 12 independent prompt realizations.

The small canonical alias-count correlation also does not contradict the alias-dose intervention: the former compares existing classes, whereas the latter changes the number of aliases for a fixed executable class. They answer different questions.

The main invariance finding does not depend on this vocabulary or its correlations. It comes from changing the written representation while preserving the execution-verified behavioural classes and measuring the resulting changes in selection and discovery credit.

Task wording and its possible interaction with jargon

The objective is communicated explicitly as maximizing the highest observed fitness, rather than choosing the most sophisticated-sounding policy.

The system instruction is:

“You are an experienced protein-engineering scientist choosing a closed-loop discovery policy. You answer with a single menu identifier and nothing else.”

The user-task template is reproduced below. For readability, line breaks have been reflowed and list markers removed; the wording is otherwise unchanged. Braced fields are filled with the landscape’s site count, the presented menu, and an example menu identifier.

“You must choose ONE discovery policy to run on a combinatorial protein fitness landscape.

The landscape has {n_{site}} mutable residue sites, all combinations of the 20 amino acids at those sites, and a measured fitness for every variant. You may not see the fitnesses in advance.

Campaign protocol, identical whichever option you choose:

you start from 24 variants drawn uniformly at random, with their fitnesses revealed;

you then run 4 rounds; in each round your policy selects 24 unmeasured variants and their fitnesses are revealed;

total assay budget is 96 variants.

You are scored on the highest fitness among all variants revealed to you by the end of the campaign.

Below is the menu of policies available to you. Every option is executable on this landscape and every option consumes exactly the same assay budget. Options are otherwise at the same standard configuration.

{menu}

Reply with the identifier of the single option you would run (for example {example}). Reply with the identifier only.”

For precision, “total assay budget” in this historical template refers to the 96 adaptive assays; execution includes 24 initial observations, giving 120 observed variants in total, as described in the paper. This wording was shared across comparison arms. Its effect on choices, including possible interactions with menu wording, has not been separately isolated.

Yes, task phrasing can interact with sensitivity to jargon. Holding the system instruction and task description fixed makes our comparison conditional on that task formulation; it does not establish invariance across all possible formulations.

The additional controls help bound the interpretation. Removing the user-task description changes selection beyond the same-prompt replicate floor in all seven tested models, indicating that choices respond to task information, not that the models necessarily interpret every instruction correctly. Separately, the sealed description-paraphrase experiment produces a median total-variation distance of 0.145 across six surfaces, while preserving the top-ranked class and credit sign in all six cases. Surface sensitivity therefore need not produce a changed scientific verdict in every setting.

Our conclusion is correspondingly specific: under a fixed, explicitly stated discovery objective, alternative written representations of the same execution-verified opportunities can change model choices and, in the demonstrated settings, discovery credit and model rankings. The contribution is to identify and measure this attribution confound, rather than to claim that jargon alone explains model behaviour.

[About OpenReview](#)

[FAQ](#)

[Sponsors](#)

[All Venues](#)

[Contact](#)

[Hosting a Venue](#)

[Terms of Use / Privacy Policy](#)

[Donate](#)