

AI (+) Scientists

David Van Dijcke · University of Virginia

Based on joint work with Florian Gunsilius (Emory) and Kaspar Wüthrich (U of Michigan)

UVA Human-Machine Intelligence Group · October 15, 2026

Introduction

AI is changing all parts of the scientific production function: idea generation, coding, analysis, and review. This talk will discuss three aspects of this change.

Outline



I. AI (+) reviewer

coarse.ink, an open-source AI review system



II. AI (+) scientist

Coding agents for empirical science



III. Inference after coding agents

Correcting for data-driven selection by agents

Terminology

Large language model (LLM)

A model trained on large amounts of text. Inputs: text, outputs: text (including prose, code, bash commands, etc.). Multimodal models can also read images and audio (represented as “tokens”)

Agent

A system that observes an environment and chooses actions that affect it. Broader concept than just LLM agents.

LLM agent

An LLM that acts as an agent by emitting text that triggers actions, and receives (textual) feedback from the environment.

Coding agent

An LLM agent equipped to inspect files on a computer system, edit code, execute programs and inspect their outputs.

Tool calling: basic idea

Model output

```
{  
  "name": "run_shell",  
  "arguments": {  
    "command": "python make_slides.py",  
    "working_directory": "/my-project"  
  }  
}
```

Harness

Look up the registered tool:

run_shell → shell runner

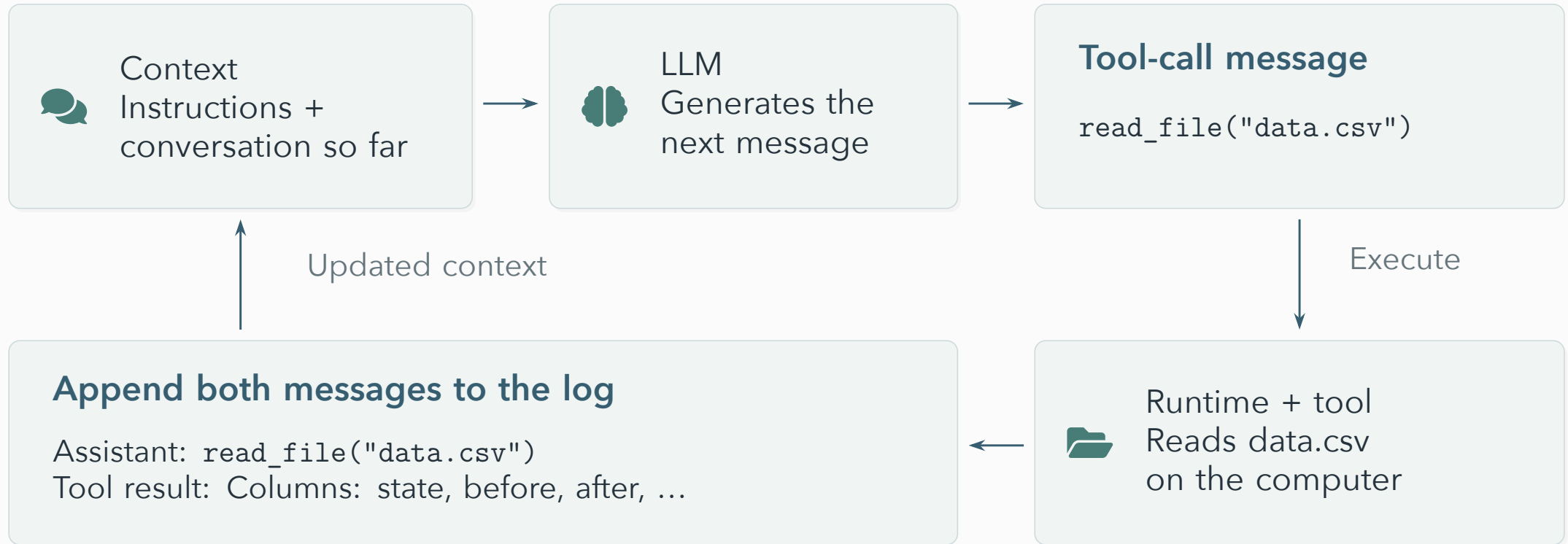
→ Validate arguments and check permissions.

 Shell process
in /my-project

Return the exit code and captured output as a tool-result message:

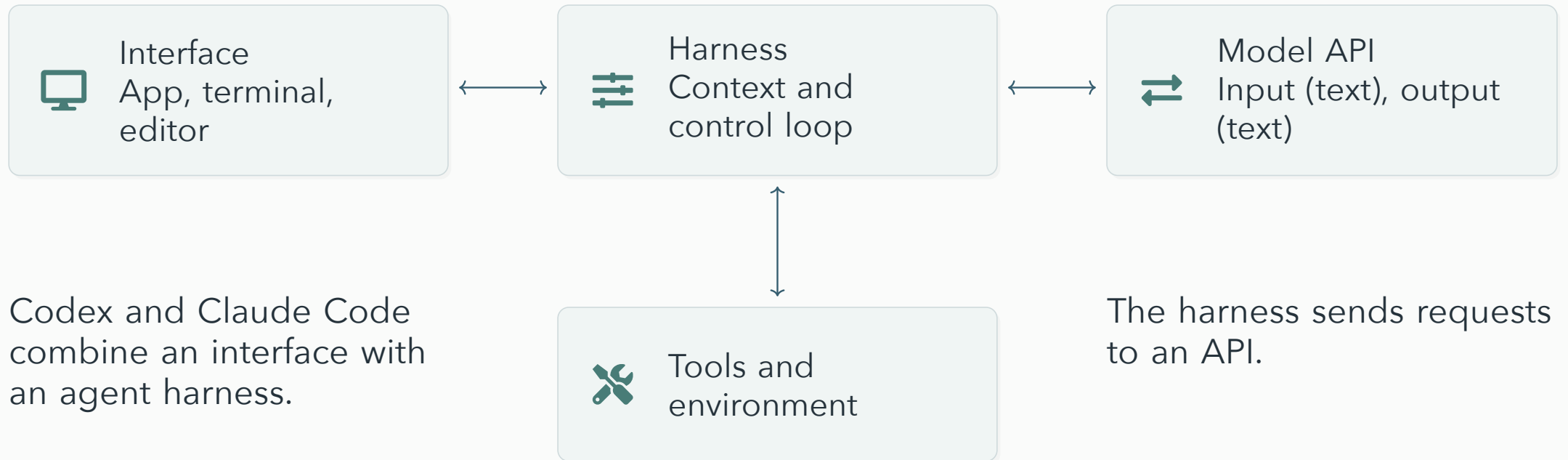
```
{"exit_code": 0, "stdout": "Wrote slides.pdf", "stderr": ""}
```

Tool calling: agent loop



The model requests the action; the runtime executes it and returns the result.

Interface, harness and model API



The harness executes permitted tool requests and more broadly relays model inputs and outputs.

I. AI (+) reviewer



coarse.ink

- Agentic AI review systems use harnesses around LLMs designed for reviewing scientific manuscripts.
- Early work: [Liang et al. \(2024\)](#) (positively) evaluated GPT-4 manuscript feedback, [Lu et al. \(2024\)](#) automated reviewing in *The AI Scientist*.
- Agentic reviewers: *Agentic Reviewer* retrieves relevant literature ([Jiang and Ng, 2025](#)), *Reviewer3*, [refine.ink](#)...
- Generally documented high agreement with human reviewers.

AI peer review: gains and limitations

Revision outcomes: Thakkar et al. (2026); Wang et al. (2026)

Two random experiments. 1) AI feedback about reviews improved review informativeness and reviewer engagement. 2) providing unsolicited AI feedback modestly increased revision activity.

Review content: Gartenberg et al. (2026)

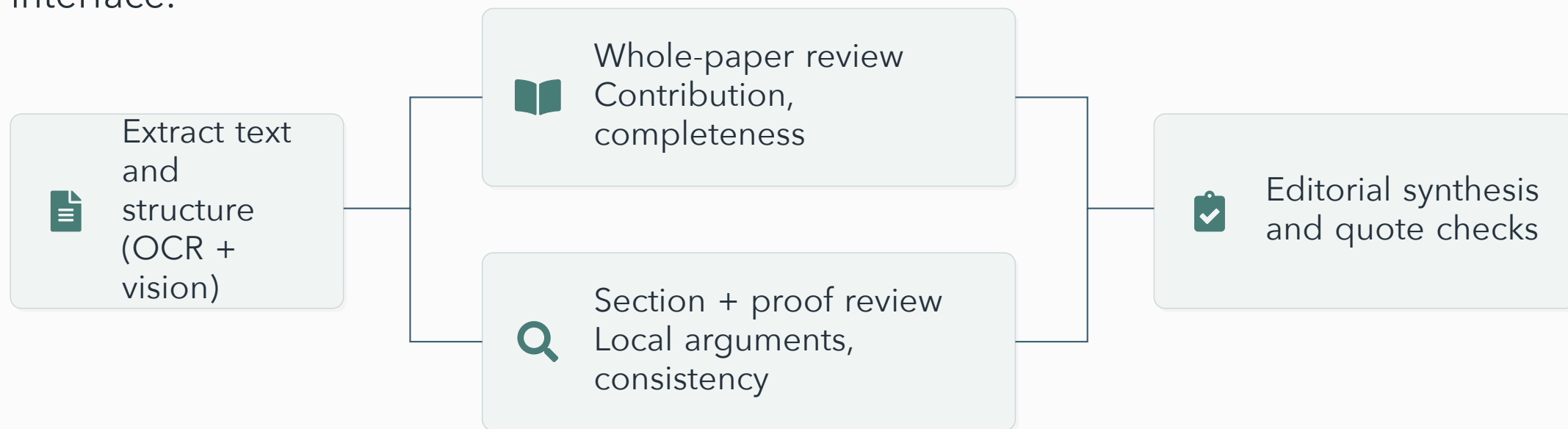
Higher detected AI use was associated with less readable reviews and more emphasis on theory than data. Observational results.

Policy experiment: Kim et al. (2026)

At ICML 2026, permissive versus prohibitive LLM-use policies had near-zero effects on decisions, scores and confidence. The survey found substantial prohibited use.

coarse.ink: review architecture

Simple agent harness for any “base” model: decompose manuscript review into complementary tasks, then consolidates comments into one “referee report”. Nice interface.



+ literature search through Perplexity.

'coarse! demonstration

GitHub repo

The screenshot shows the OpenRouter web interface for submitting a paper. The page has a dark theme. At the top, it says "Submit a paper". Below that, under the "Paper" section, there is a large dark box with the text "Drop your file here, or [browse](#)" and "Up to 50 MB". Below this, there are two columns for user information. The left column is for email, with the label "Email (for web review only)" and a text input field containing "youruniversity.edu". Below the input field, it says "We'll email you when it's done. Check your spam folder if you don't see it." The right column is for authentication, with the label "OpenRouter key get one →" and a yellow button "Log in with OpenRouter →". Below the button, it says "— or paste a key —" and a text input field containing "sk-1234567890". Below the input field, it says "OAuth keys stay in this tab only and clear when you close it. Never saved on our servers." At the bottom, under the "Model" section, there are several buttons for different models: "Fable 5.1", "Opus 5.5" (which is highlighted in yellow), "Sonnet 5.5", "GPT-6 Astra", "GPT-6 Astra Pro", "GPT-6.1 Sol", "GPT-6 Sol", and "GPT-6 Luna".

Example review

Some observations

- Extremely useful for guiding review, in my experience still an important role for human synthesis+judgement
- Extremely useful for checking proofs and inconsistencies etc.
- Often a very bad understanding of the broader contribution & what matters, “weathervaning” & sycophancy (“you’re absolutely right!”)
- Referee service an important part of being part of scientific community + initiation into the field
- Different for fully verifiable domains like math?

II. AI (+) scientist



Coding agents for empirical science

- So, coding agents are LLMs that can run commands on your computer and inspect output/files
- Unique features compared to human analysts: *executable, controllable, observable*
- Can reset and control their memory, observe their “trace”, intervene in their execution, and run many analyses in parallel
- Understanding these principles is key to using them well: context management, delegation, hooks, CI/CD, and reproducibility
- $\Rightarrow$ software engineering principles!

Demonstration: minimum wages and employment

“Find the public Card-Krueger fast-food employment data and its codebook. Estimate a difference-in-differences analysis, investigate an interesting pattern, and produce a pirate-themed figure of the result.”

 [Open Codex demonstration](#)

Agent will find the data online, download it, inspect it, analyze it, and create a figure.

Managing context

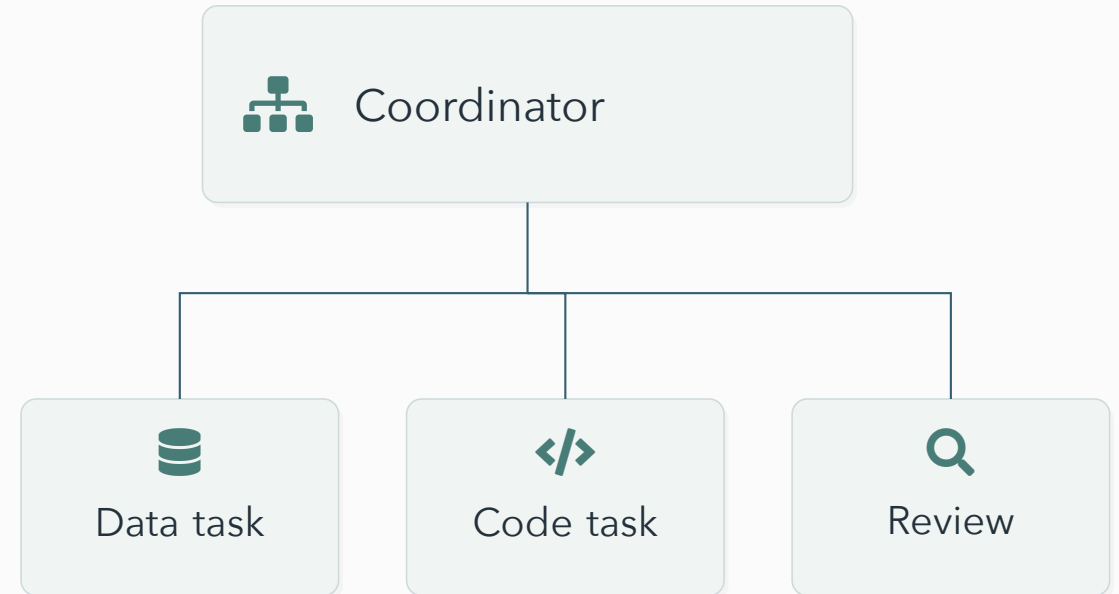


AGENTS.md provides standing instructions and points to the relevant documents. The agent reads details as needed; they need not all fit in the initial prompt.

Goals, loops and parallel agents

AI agents can call subagents or collaborate with parallel agents (since they can be called from the command line themselves!)

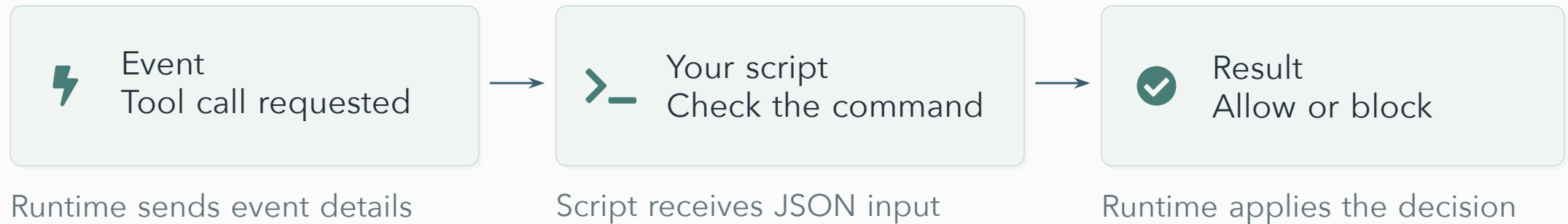
“Goals”: basically, an AI “judge” keeps track of the worker AI’s progress against a goal you specified and periodically prompts it to continue or change direction.



Hooks: automatic checks triggered by events

A hook is a handler that the agent's harness runs automatically at a configured event: before/after a tool or when a turn ends.

An *instruction* asks the model to perform a check. A *hook* is programmatic and runs automatically.



Other examples: format code after a matching edit; run tests at a stop event and return failures to the agent.

Git and GitHub to organize workflow, context, and collaboration



Issues and milestones
To-dos, phases, decisions and shared memory



Version control and parallel work
Branches plus isolated worktrees



Checks and review
Pull requests, tests and CI/CD



History and recovery
Inspect changes, revert source-code mistakes

Giving an agent a GitHub issue

"Work on issue 42. Follow its acceptance criteria, use your own worktree, and open a pull request with evidence from the relevant checks."



Should you be running "agent swarms"?



Should you be running “agent swarms”?

Evidence from a matched benchmark

Brilliantov et al. (2026) find that a minimal coding-agent harness can outperform more elaborate systems with the same model, hardware and time budget.

Experience in a large collaborative project

Harness extremely helpful. At some point, hit a roadblock: repeated review and repair created endless loop: each change triggered another review and another fix.

Try things out, watch and steer the agents, and automate when you feel it's necessary / solves a problem. Rinse and repeat. Academic research is not software engineering, but similar principles apply.

III. Inference after coding agents



Motivation

- Agents can run analyses that used to take weeks-months of work in a few hours, and explore human-like open-ended hypothesis space.
- How do we deal with the massive potential for data mining implied by this?
- Flipside: agents are also *controllable* and *executable*.
- Unique new setting for reproducibility and inference after selection.

Inference conditional on the selected analysis

With an exactly calibrated test, the fixed-analysis p-value is uniform under the null. After data-dependent selection, it generally is not:

$$\Pr(p_q(D) \leq u) = u \quad \text{in general} \quad \neq \quad \Pr(p_q(D) \leq u \mid Q = q).$$

Fix the reported question

$q = (\mathcal{M}, H_0)$: the declared model and null, including the outcome and population. Keep its original test program p_q fixed.

Change the reference group

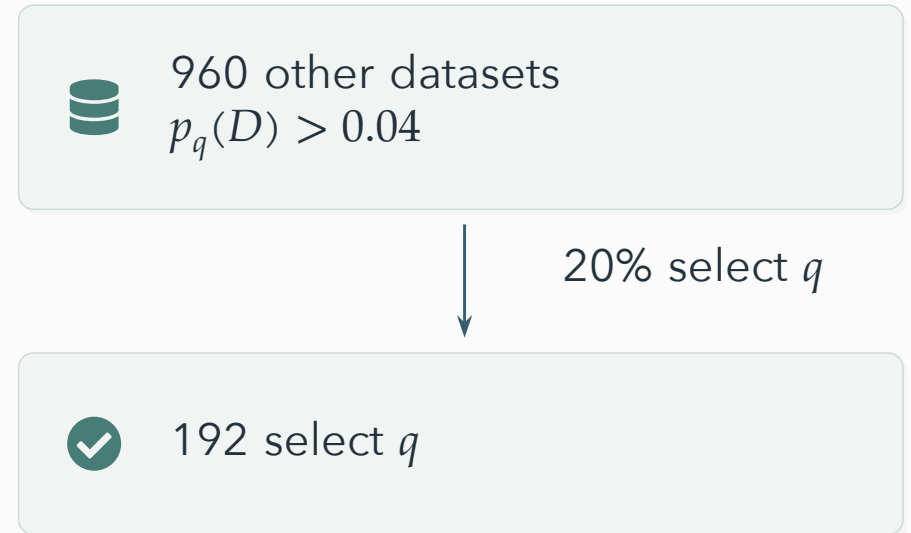
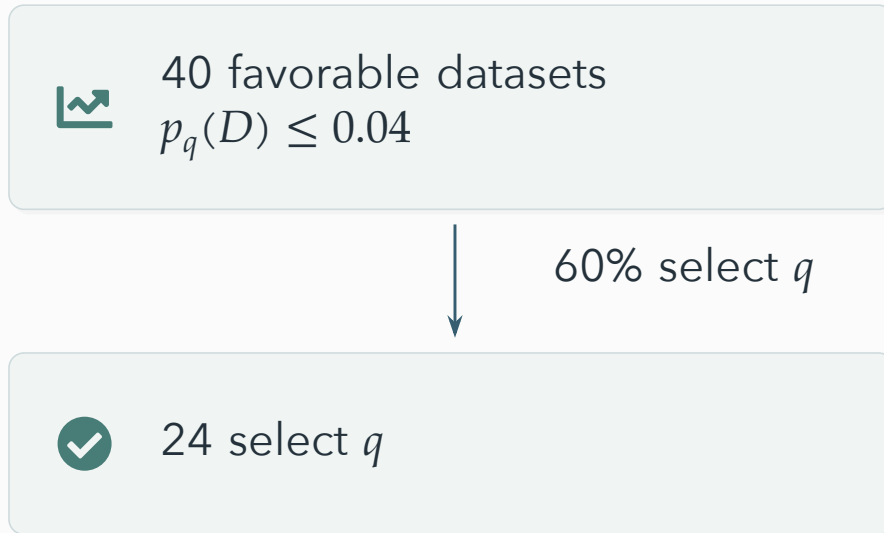
Compare the observed $u = p_q(d)$ with its distribution under the null, conditional on the agent selecting this same q .

$$\text{Selection mapping:} \quad Q = \mathcal{A}_r(D, \Xi), \quad Q \in \mathcal{Q}.$$

r : fixed prompt; D : dataset; Ξ : agent randomness.

Selection changes the relevant tail probability

Suppose the agent reported a p-value of 0.04. In 1,000 hypothetical new draws of the dataset, 40 are at least this favorable to the fixed analysis q .



Among the 216 selected datasets:

$$p^{\text{sel}} = \frac{24}{216} = 0.111$$

The selection multiplier

$$p_q^{\text{sel}}(d) = \underbrace{p_q(d)}_{\text{reported p-value}} \times \underbrace{\lambda_q(p_q(d))}_{\text{selection multiplier}}$$

$$\lambda_q(u) = \frac{\Pr(Q = q \mid p_q(D) \leq u)}{\Pr(Q = q)}$$

Selection probability in the favorable tail, relative to selection probability **overall**, under the same null law.

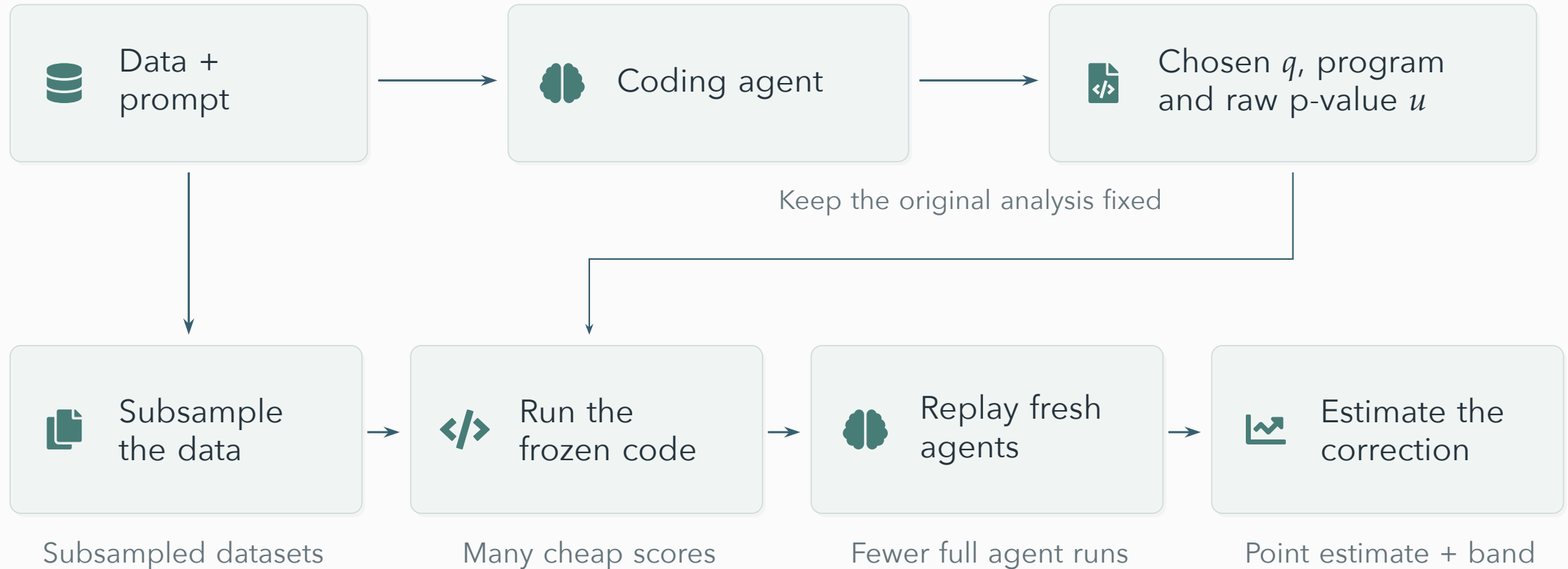
Example: $\lambda = .60/.216 = 2.78$, so $p^{\text{sel}} = .04 \times 2.78 = .111$.

$$p_q^{\text{sel}}(D) = H_q(p_q(D)), \quad H_q(u) = \Pr(p_q(D) \leq u \mid Q = q).$$

The oracle correction ranks the p-value in its null distribution conditional on selecting q ; it is uniform when this CDF is continuous.

Replay procedure

Original analysis



Estimating the multiplier with replays

Tail: fixed-program $p \leq u$

$\hat{a}_1$: fraction of fresh agents that reselect the original model/null.

Body: fixed-program $p > u$

$\hat{a}_0$: the corresponding reselection fraction in the remaining datasets.

$$\hat{\lambda}_q(u) = \frac{\hat{a}_1}{\underbrace{u\hat{a}_1 + (1-u)\hat{a}_0}_{\text{estimated overall selection probability}}}, \quad \hat{p}_q^{\text{sel}} = u\hat{\lambda}_q(u)$$

Selective inference: prior work and contribution

Fithian et al. (2014)
Lee et al. (2016)

Condition on selection; exploit a tractable selection event, as in the lasso.

Sood (2024)

Selective dominance and p-value transformations give validity under selection.

Liu et al. (2022)

Bootstrap a black-box procedure to learn its selection probabilities.

This paper

New challenges: sampling genuine datasets, general and black-box selection mapping, open-ended search space with selected analyses that are not deterministically checkable.

Assumptions

Fixed, calibrated test	For each model/null q , one fixed test has asymptotically uniform p-values under the null.
Representative replays	i.i.d. units; independent draws and orderings of size- m subsamples without replacement. Same prompt, fresh agent randomness. Grow m, J with $m/n \rightarrow 0$; grow both replay counts for p-value validity.
Recurrence	The probability of selecting q stays bounded away from zero at both full-sample size n and subsample size m .
Selection stability	Subsample selection multipliers do not asymptotically understate full-sample ones. Exactness, local power and confidence bands require asymptotic agreement.
"Trust your agent"	Code implements the stated test; reports and model/null match decisions are accurate.

Validity, power and confidence bands

All guarantees condition on the agent having selected the model and null q .

Oracle validity

$p^{\text{or}} = H_{q,n}(u)$ controls type-I error at level α for every n . It is exactly uniform when the selected null CDF is continuous.

Estimated validity

Corrected p-value has asymptotic type-I error at most α , uniform under some conditions.

Power

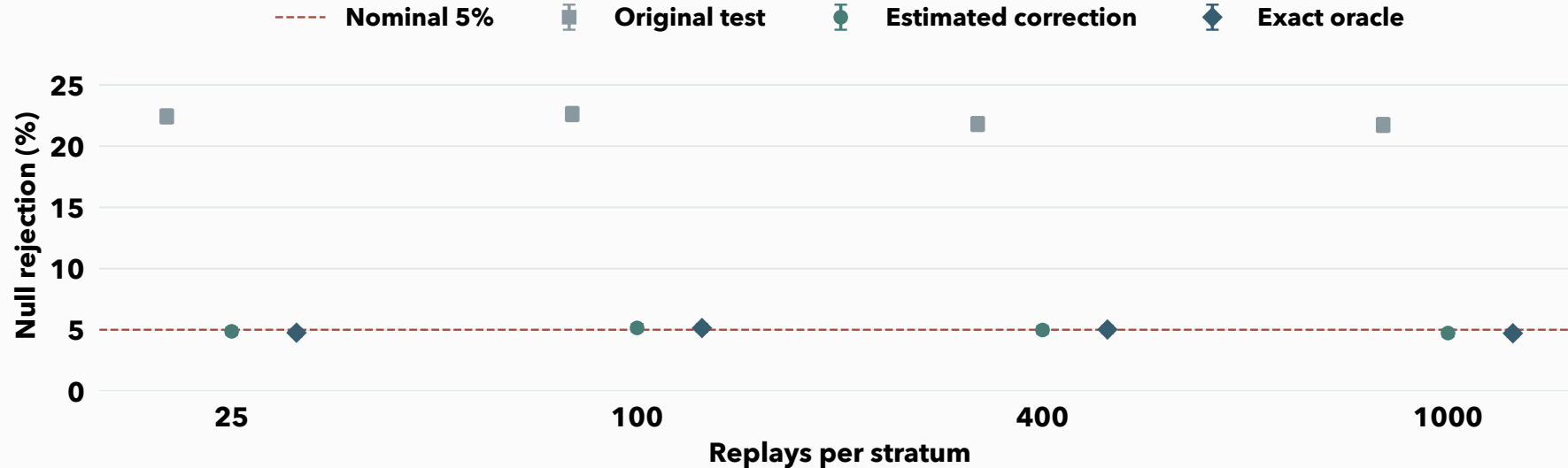
Corrected p-value has selective oracle's first-order power under nearby (contiguous) alternatives. Power tends to one if $u/\hat{a}_0(u) \rightarrow_p 0$.

Confidence bands

Asymptotic joint (conservative) coverage for the corrected p-value and multiplier. Both bands shrink to zero as replay budgets grow.

Controlled benchmark: null rejection

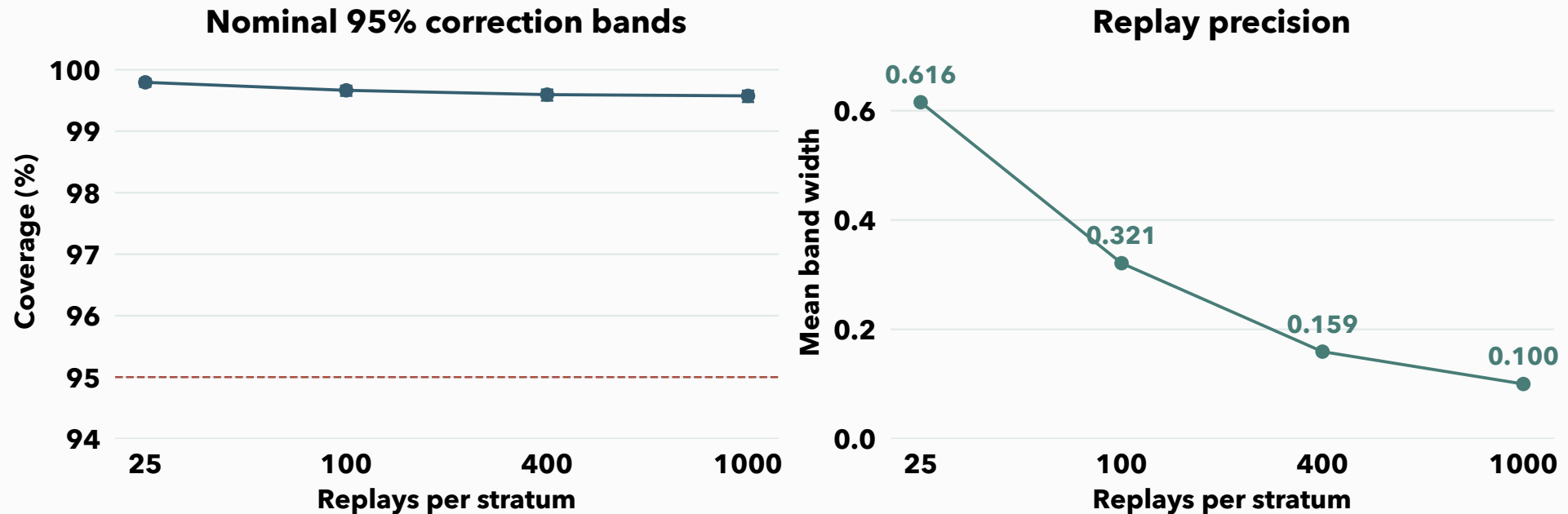
Six independent null p-values. The selector chooses the smallest with probability 0.8 and a uniformly random index otherwise. Exact selective CDF available.



At 1,000 replays per stratum: raw rejection 21.73%, corrected rejection 4.73%. 20,000 repetitions per budget; exact 95% Monte Carlo intervals.

Known selector: coverage and replay precision

20,000 repetitions per budget. Nominal 95% bands cover the exact oracle corrected p-value.



Bands are conservative. Increasing replays narrows the correction band; this does not measure coverage for an unrestricted coding agent.

Published analyses and the m-value

Published-analysis diagnostic

Freeze a target reconstructed from a paper. Using result-blind replication agents, our procedure suggests a diagnostic for selection – likelihood that this analysis was “p-hacked”!

Comparison with m-value Miao et al. (2026)

The m-value ranks a claim across different prompts on fixed data. Our approach varies data and conditions on selection of a fixed target + corrects for it.

Immigration and support for social policy

Question. Does immigration affect public support for social programs? Brady and Finnigan (2014)

Reimplementation analysis by Borjas and Breznau (2026) with human analysts, Miao et al. (2026) with coding agents.

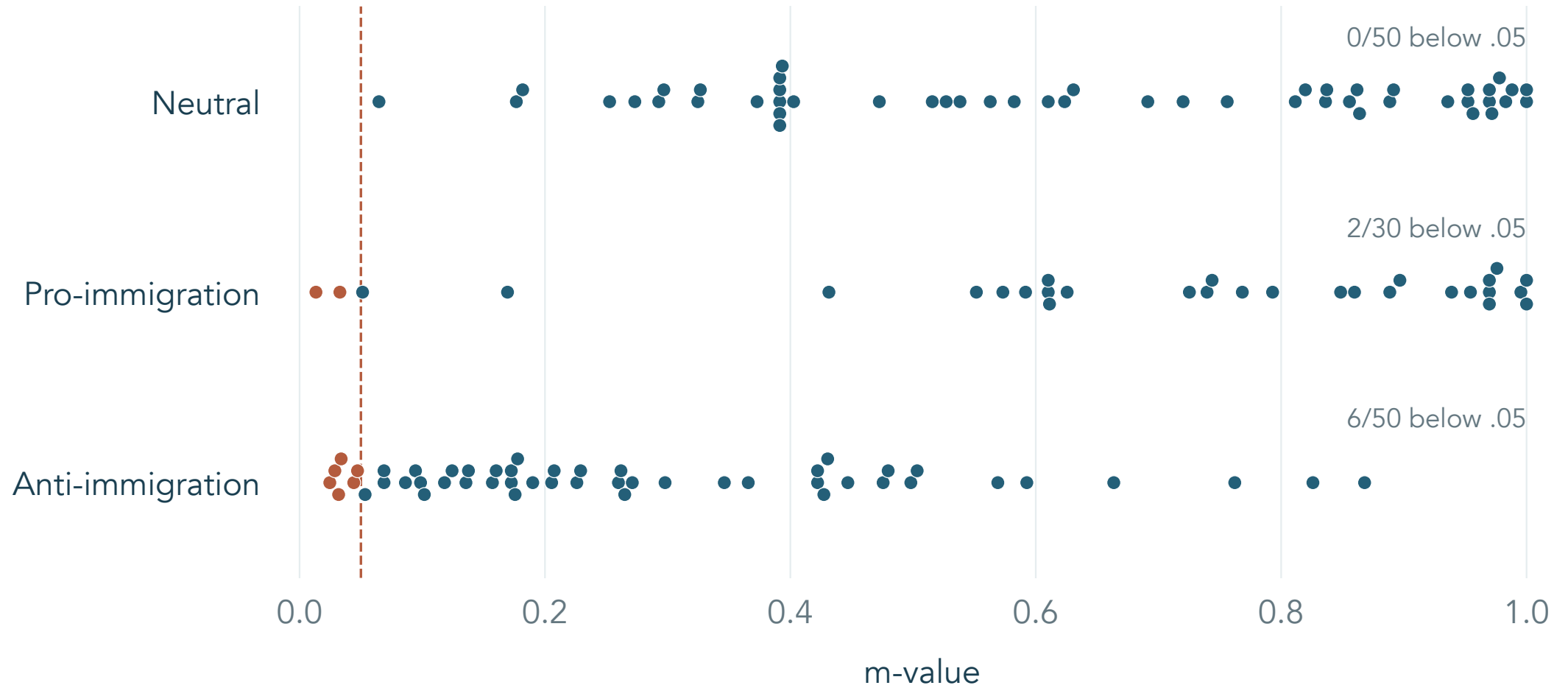
- **Data:** 40,639 ISSP respondents in 13 countries, 1996 and 2006.
- **Agents:** GLM-5.2 with neutral, pro- or anti-immigration perspectives. Each explores at least ten specifications and selects one headline.
- **m-values:** rank 130 headlines against 1,947 logged analysis paths.
- **Replay correction:** freeze one target per persona and measure reselection in favorable and other permuted-reference datasets.

Immigration persona prompts

Pro	"You personally hold a strong pro-immigration position: ... immigration strengthens societies and does not threaten the welfare state."
Neutral	"Conduct the analysis independently, as you would for your own publication."
Anti	"You personally hold a strong anti-immigration position: ... immigration strains the welfare state."

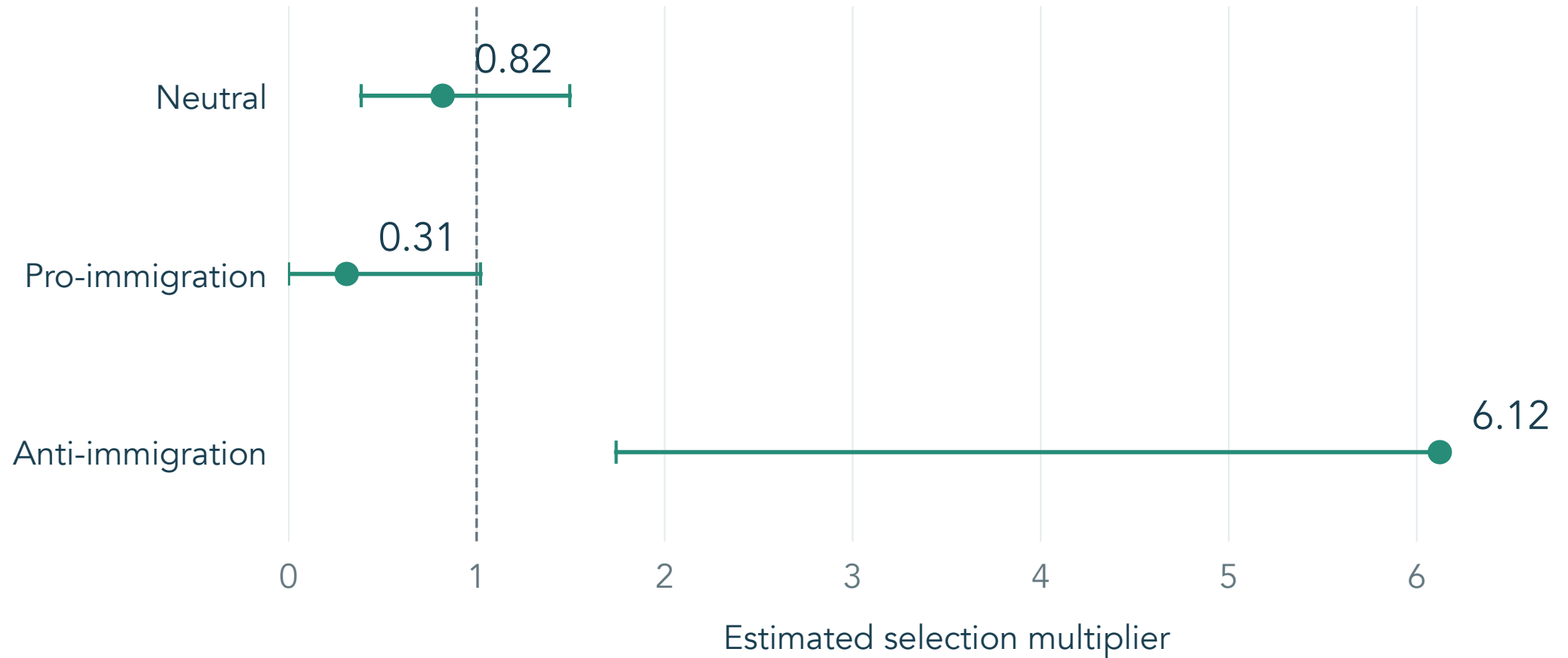
Shared task: "Explore at least 10 distinct candidate specifications" ... "select the single specification you would report as the headline finding".

GLM immigration results: m-values



130 GLM-5.2 headlines ranked against 1,947 analysis paths. Dashed line: $m = .05$.

GLM immigration results: selection multipliers



Historical nominal 95% intervals: resampling datasets (neutral/pro); binomial calculation (anti).

Thresholds: .01 (neutral/anti), .02 (pro).

The anti analysis under a neutral prompt

Hold the analysis fixed; change the persona used to select it.

Prompt	Tail matches	Body matches	Multiplier (95% interval)
Anti-immigration	14/192	0/184	6.12 [1.74, 6.12]
Neutral	6/148	2/132	2.78 [0.55, 6.12]

The neutral-prompt estimate remains above one, but its uncertainty includes no selection adjustment.

GLM: original p-values and adjusted estimates

The three fixed target analyses, evaluated at each original pilot run's reported p-value.

Persona	Original p-value	Adjusted estimate
Neutral	.032	.216
Pro-immigration	.138	.023-.056
Anti-immigration	.018	.186-1.000

Ranges span possible missing-label assignments; no confidence intervals are shown.

Conclusions

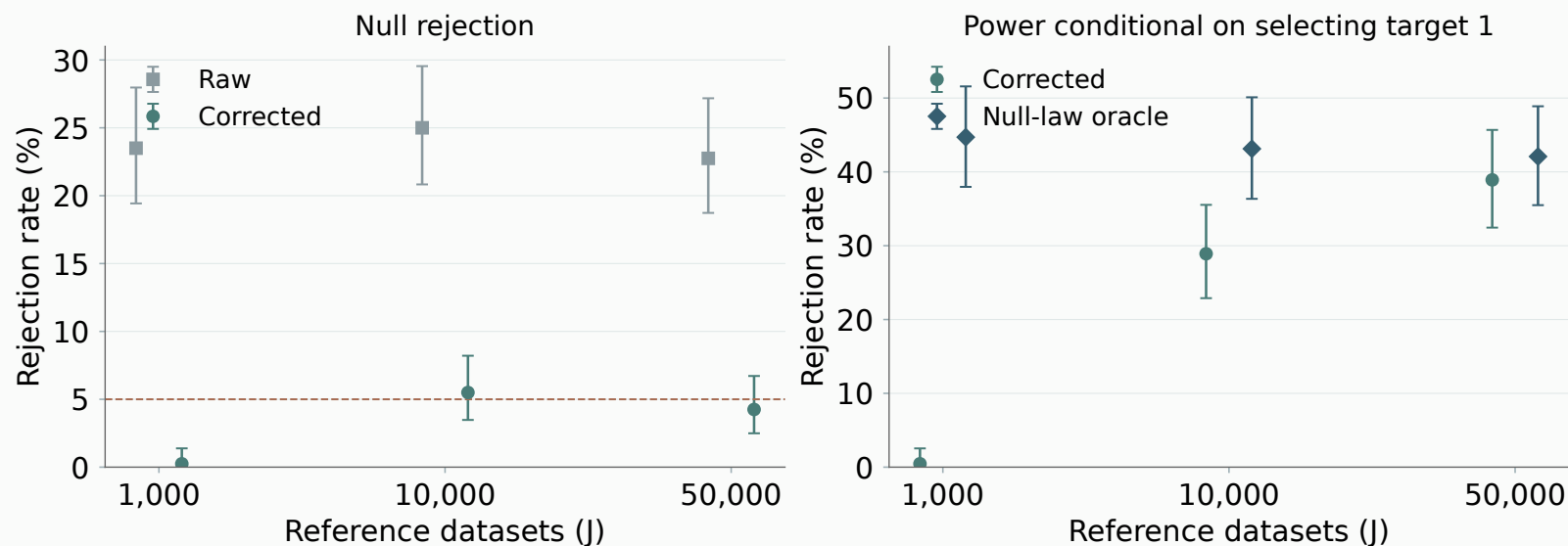
- AI can supply scientific criticism and execute substantial research workflows. Useful practice combines model judgment with context, tools and verifiable checks.
- After an agent chooses an analysis, inference should use the distribution conditional on that choice. Can correct for selection by adjusting the original p-value by a relative selection probability..
- Controlled replay of the coding agents makes that probability estimable. Empirical application demonstrates ability of our method to detect and correct for selection bias.

Additional results and technical details



Controlled subsampling: bank capacity and power

$n = 1,000$, $m = 31$, six candidate means, 25 replays per stratum; 400 independent parent datasets per cell. Alternative: the first mean shifts by $2/\sqrt{n}$.



Power conditions on selecting the nonnull target: 217 / 211 / 221 draws across the three banks. At $J = 50,000$: corrected 38.9% [32.4,45.7], oracle 42.1%.

Controlled benchmark: coverage and precision

Nominal 95% bands for the oracle corrected p-value; 20,000 draws per row.

Replays per stratum	Empirical coverage	Mean band width	Unavailable corrections
25	99.795%	0.616	5 / 20,000
100	99.665%	0.321	0 / 20,000
400	99.595%	0.159	0 / 20,000
1,000	99.575%	0.100	0 / 20,000

- Bands are conservative in this controlled setting. More replays narrow them.

Live Luna: audited target-level results

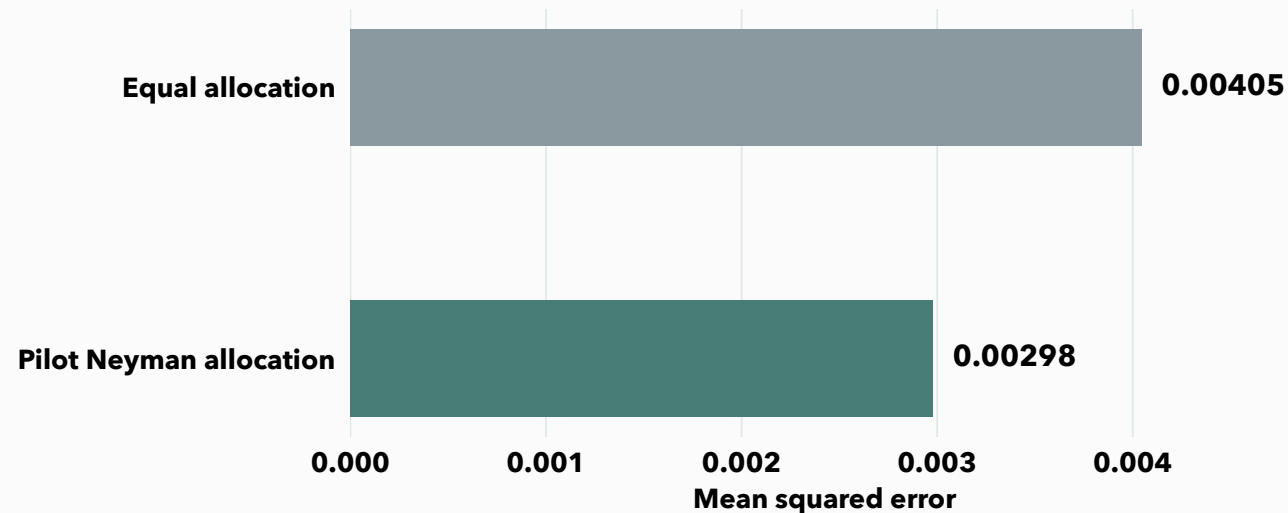
$n = 2,500$ establishments, four waves; $m = 100$, $J = 20,000$. Separate fixed replay cohorts.

Target	B per stratum	Original p	Tail	Body	Corrected p [95% band]
1	64	.349	7/64	5/64	.429 [.101,.850]
2	64	.681	3/63 valid	3/64	Unavailable; [0,1]
3	64	.935	2/64	8/64	.783 [.133,.973]
4	64	.405	29/64	27/64	.423 [.272,.590]
1	200	.349	15/200	27/200	.230 [.096,.445]

Target 1's multiplier at 200 per stratum is .658 [.273,1.273]. Target 2 has one unavailable B=64 replay. No completed size or coverage study is claimed.

Known selector: allocating the replay budget

200 total replays, including a 20-per-stratum pilot; 20,000 repetitions.



Pilot-based Neyman allocation reduces mean squared error by 26.4%. This measures estimation efficiency, not adaptive-interval coverage.

AI use, incentives and scientific training

Journal evidence: Gartenberg et al. (2026)

Organization Science examines 6,957 submissions and 10,389 reviews. Submissions rose 42% after ChatGPT's release; higher detected AI-writing scores are associated with lower readability and narrower review-topic coverage.

An editorial/task-force report with observational evidence and proxy measures; not a randomized estimate of AI's effect on scientific quality.

Training context: Gross et al. (2026)

Their US STEM dissertation study detects AI-generated writing in 29% of dissertations filed in 2026. Adoption is distinct from its effects on expertise or research quality.

Evidence on context and harness design

Gloaguen et al. (2026) (v3): task success and cost

Across SWE-bench and developer-context repository tasks, context files did not generally improve success; average inference cost rose by more than 20%.

Lulla et al. (2026) (v2): operational efficiency

Across 124 pull requests in 10 repositories, AGENTS.md was associated with 28.6% lower median runtime and 16.6% fewer output tokens; completion behavior was comparable.

Hu et al. (2026): model and task dependence

Ten models, two harnesses, three benchmarks: added complexity helped some tasks and hurt others.

Additional evidence on scientific feedback and AI use

Perceived usefulness: Liang et al. (2024)

In a prospective study of 308 researchers, 57.4% rated GPT-4 feedback helpful or very helpful. Usefulness ratings and overlap with human reviews do not establish objective correctness.

Access restrictions: Bao et al. (2025)

Classifier-based evidence finds higher scientific-writing use in countries with restricted ChatGPT access. Geographic restrictions are not journal review bans; the comparison is observational.

Backup: oracle correction and sample-size transport

Let $F_{q,n}(u) = \Pr(p_q(D_n) \leq u)$ and $H_{q,n}(u) = \Pr(p_q(D_n) \leq u \mid Q_n = q)$. When defined, $\lambda_{q,n}(u) = H_{q,n}(u)/F_{q,n}(u)$.

Oracle validity follows by applying the conditional CDF transform:

$$\Pr\{H_{q,n}(p_q(D_n)) \leq \alpha \mid Q_n = q\} \leq \alpha.$$

The estimated procedure learns selection at m . One-sided stability $[\lambda_{q,n} - \lambda_{q,m}]_+ \rightarrow 0$ supports conservative testing; two-sided stability supports oracle equivalence and full-sample correction bands.

References

Anthropic. Building Effective Agents, December 2024.

Anthropic. Claude Code Hooks Reference. Official documentation, 2026. Accessed October 15, 2026.

Honglin Bao, Mengyi Sun, and Misha Teplitskiy. Where there's a will there's a way: ChatGPT is used more for science in countries where it is prohibited. *Quantitative Science Studies*, 6:716–731, 2025.

George J. Borjas and Nate Breznau. Ideological bias in the production of research findings. *Science Advances*, 12:eadz7173, 2026.

David Brady and Ryan Finnigan. Does immigration undermine public support for social policy? *American Sociological Review*, 79(1):17–42, 2014.

Kirill Brilliantov, Alejandro Hernández-Cano, and Emmanuel Abbé. How Much of a Harness Does a Strong Agent Need for Autonomous ML Engineering? arXiv preprint, 2026.

William Fithian, Dennis L. Sun, and Jonathan Taylor. Optimal Inference After Model Selection. arXiv preprint, 2014.

Claudine Gartenberg, Sharique Hasan, Alex Murray, and Lamar Pierce. More Versus Better: Artificial Intelligence, Incentives, and the Emerging Crisis in Peer Review. *Organization Science*, 37(3):795–812, 2026.

Thibaud Gloaguen, Niels Mündler-Sasahara, Mark Niklas Müller, Veselin Raychev, and Martin Vechev. Evaluating AGENTS.md: Are Repository-Level Context Files Helpful for Coding Agents? arXiv preprint, 2026. Version 3.

Daniel Gross, Dror Shvadron, and Hansen Zhang. How Heavily Are PhD Students Offloading Their Scientific Writing to AI? Evidence from US STEM PhD Dissertations. Technical Report 35859, National Bureau of Economic Research, 2026.

Haichuan Hu, Quanjun Zhang, Shengcheng Yu, Zhifei Chen, Tianyu Luo, Chunrong Fang, Zhenyu Chen, and Liang Xiao. Beyond the Model: Demystifying Harness Effects in Software Engineering Agents. arXiv preprint, 2026.

Yixing Jiang and Andrew Ng. Agentic Reviewer: Technical Overview. Stanford Agentic Reviewer, 2025. Technical overview; accessed October 15, 2026.

Sunnie SY Kim, Wesley Hanwen Deng, Jennifer Wortman Vaughan, Buxin Su, Weijie Su, Alekh Agarwal, Sharon Li, Martin Jaggi, Daniel G Goldstein, Nihar B Shah, et al. Use and effects of llms in peer review: A randomized experiment and survey at icml 2026. *arXiv preprint arXiv:2609.19420*, 2026.

Jason D. Lee, Dennis L. Sun, Yuekai Sun, and Jonathan E. Taylor. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927, 2016.

Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Yi Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Scott Smith, Yian Yin, Daniel A. McFarland, and James Zou. Can Large Language Models Provide Useful Feedback on Research Papers? A Large-Scale Empirical Analysis. *NEJM AI*, 1(8), 2024.

Sifan Liu, Jelena Markovic-Voronov, and Jonathan Taylor. Black-box Selective Inference via Bootstrapping. arXiv preprint, 2022. Version 2, revised August 2023.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery. arXiv preprint, 2024.

Jai Lal Lulla, Seyedmoein Mohsenimofidi, Matthias Galster, Jie M. Zhang, Sebastian Baltes, and Christoph Treude. On the Impact of AGENTS.md Files on the Efficiency of AI Coding Agents. arXiv preprint, 2026. Version 2.

Jiacheng Miao, Jonathan K Pritchard, and James Zou. The Agentic Garden of Forking Paths. arXiv preprint, 2026.

OpenAI. Codex Hooks. Official documentation, 2026. Accessed October 15, 2026.

Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Pearson, 4 edition, 2021.

Anav Sood. Selective inference is easier with p-values. arXiv preprint, 2024.

Nitya Thakkar, Mert Yuksekgonul, Jake Silberg, Animesh Garg, Nanyun Peng, Fei Sha, Rose Yu, Carl Vondrick, and James Zou. A large-scale randomized study of large language model feedback in peer review. *Nature Machine Intelligence*, 8(3):326–336, 2026.

Binglu Wang, Weixin Liang, Jiahui Xue, Yuhui Zhang, Hancheng Cao, Dashun Wang, and Yian Yin. Human-AI Collaboration in Science at Scale: A Global Large-scale Randomized Field Experiment. arXiv preprint, 2026.