DO LLMS HAVE RESEARCH TASTE? AN ANALYSIS USING *BÓLÈ*BENCH

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) are increasingly used for recursive self-improvement, in which they generate and implement research ideas intended to advance their own capabilities. With finite compute, researchers must decide which ideas to pursue before their outcomes are known. This judgment reflects research taste developed through experience. Researchers increasingly rely on LLMs to make these decisions, yet their research taste remains poorly measured. We introduce BoleBench, a benchmark that grades foresight with hindsight. We start from parallel rollouts, where different agents attempt to improve a shared baseline on a fixed set of ML research tasks under matched compute budgets. An LLM reads abridged, outcome-free descriptions of two attempts at the same task and predicts which scored higher. The executed outcomes serve as ground truth without requiring human outcome labels. BoleBench is a pipeline that turns executed attempts into test items covering five categories of research decisions, tests pick judgments at four information levels, and scores each judge on multiple axes. New item banks are built from post-cutoff corpora as models retrain. On AI4AI pairs with a clear boldness gap, a simple pick-the-bolder heuristic outperforms every full-description judge. On the subset where the bolder attempt loses, no judge exceeds chance. We release BoleBench, 4,059 blinded pairs from four corpora. On the ten shared anchor pairs, human and model majorities made identical choices, including the same three errors. On the 16-judge leaderboard the newest frontier model leads at 70.8%, against a nine-judge majority of 67.1%. On AI4AI, judges add little accuracy beyond the boldness heuristic. Asked to pick the best of ten attempts at one task, the top judge is right 36.0% of the time against a chance rate of 10%, so current judges have much room to improve.

1 INTRODUCTION

Automated research increasingly depends on choosing which ideas deserve limited compute. "Autoresearch" systems propose changes, test feasibility, and execute experiments (Jordan & contributors, 2024; Chi et al., 2026), while research preference models already allocate GPU budgets (Foster et al., 2026). As generating candidates becomes cheaper, small improvements in selection can accumulate across a research trajectory. *Research taste*, the ability to identify promising directions before outcomes are known, is therefore important for automated discovery.

Prior work predicts winners from text, ranks directions within agentic search, and selects candidates without execution (Wen et al., 2025; Mule et al., 2026; Foster et al., 2026; Zheng et al., 2026). However, existing evaluations often omit failures or executed alternatives, become vulnerable to memorization as models retrain, and obscure distinct skills, shared errors, and selection value behind aggregate accuracy. High scores may therefore reflect simple preferences, such as choosing bolder proposals. Moreover, more than half of the labeled AI4AI attempts change only schedules, hyperparameters, or other run-side factors (Fig. 8).

We introduce BoleBench[1], the first benchmark to combine executed alternatives, retained failures, renewable item banks, multi-track diagnosis, and out-of-loop selection replay. Judges choose between blinded, outcome-free descriptions of attempts executed from the same starting state under matched budgets. They commit before seeing the scores, and hidden evaluators grade their choices

[1]Named for Bo Le, the legendary appraiser of horses who could pick a winner before it raced.

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.

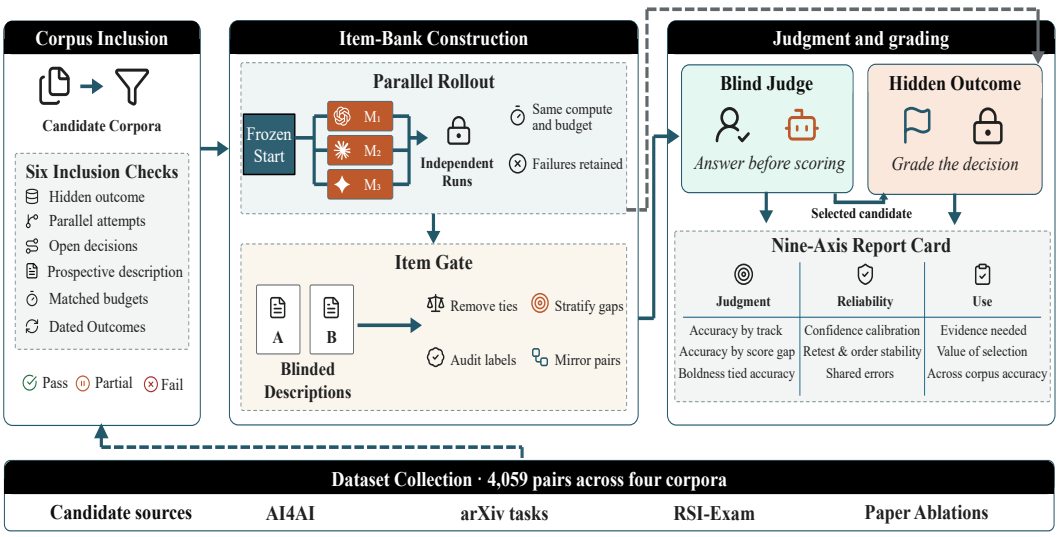


Figure 1: **From executed attempts to graded research judgment.** Corpora pass six inclusion checks. Hidden evaluators score parallel attempts, which are converted into blinded pairs. Judges answer before outcomes are revealed and receive a nine-axis report card. Four corpora share this pipeline. Track and information-level analyses use AI4AI. Model logos illustrate possible executor families.

against realized outcomes. Both alternatives are executed, so every comparison has an observed counterfactual. The framework audits leakage, removes near-ties, stratifies outcome gaps, and rebuilds dated banks from fresh corpora as training windows advance (Fig. 1). The current release contains 4,059 pairs across AI4AI, arXiv tasks, RSI-Exam, and paper ablations (Section 4.1).

We operationalize research taste as *predicting realized outcomes from prospective information beyond consensus, prestige, and simple heuristics*, following accounts emphasizing problem choice under slow feedback (Hamming, 1986; Nanda, 2025; Mahajan et al., 2025). Our key diagnostic is the *heuristic-silent margin*: accuracy above chance on equally bold attempts, where "pick the bolder" cannot choose (Fig. 4). Initial validity checks support this operationalization (Cronbach & Meehl, 1955).

On 1,127 AI4AI pairs with clear boldness gaps, "pick the bolder" scores 77.9%, beating every full-description judge. The strongest judges reach 60–64% when boldness is tied, and none exceeds chance when the bolder attempt loses. Eight judges from four families jointly miss 4.5% of audited pairs, roughly 135 times the independent-error expectation. Despite this shared reliance on boldness, judge selection matters. At $k{=}10$, the best judge recovers roughly three quarters of the oracle-over-random gain, while the weakest falls below random.

We make three contributions:

1. **Can judges identify promising ideas before seeing results?** BoleBench grades blinded choices against matched, executed alternatives, including failures, and formalizes when outcomes can measure foresight (Section F).
2. **How can the benchmark stay useful as models learn old results?** Fresh, dated releases apply checks for matched execution, blinding, leakage, ties, and outcome gaps.
3. **What does accuracy alone miss?** We use five decision tracks, four select-best information levels, a nine-axis report card, and one-of-k replay. These measure heuristic reliance, judgment beyond those cues, shared errors, and selection value. Directly choosing the best of ten remains difficult, with the best judge reaching 36% against 10% chance.

Table 1: **BoleBench vs. the ten nearest prior works: each ingredient exists somewhere, none combines them.** (✓) = partial or qualified; *Failures visible* = abandoned or failed attempts appear in the ground truth; *Selection value* = judge-based selection shown to improve realized outcomes.

Work	Ground truth	Prospective	Failures visible	Sub-skills	Info ablation	Contam. control	Human baseline	Selection value
Wen et al. '25	published result tables	✓	✗	✗	(✓)	✓	✓	✗
RPM '26	own search trees	✓	✓ biased	✗	✓	(✓)	✗	✓ in-loop
PBE '26	own MLE-bench runs	✓	(✓)	✗	✓	✗	✗	✓ in-loop
Mule et al. '26	leaderboards	✓	✗	✗	(✓)	✓	✗	✗
HindSight '26	bibliometric	(✓) leaky	✗	✗	✗	✗	✗	✗
ForeSci '26	bibliometric	✓	✗	✓	✗	(✓)	✗	✗
TastyBench '25	citation velocity	(✓)	✗	✗	✓	✗	✗	✗
Abstention '26	constructed labels	mixed	✗	(✓)	✗	(✓)	✗	✗
ReviewGuard '26	future citations	✗	✗	✗	✗	(✓)	✓	✓ retro.
SoundnessBench '26	reviewer soundness scores	✓	(✓)	✗	✗	(✓)	✗	✗
BoleBench (ours)	**parallel-attempt execution**	✓	✓	✓ 5 tracks	✓ 4 levels	✓ releases+probes	✓ 60 pairs	✓ replay

2 Related Work

Predicting outcomes, and taste as engineering. Language models can predict research outcomes prospectively. Wen et al. (2025) report 77% pairwise accuracy on their full test set; on its NLP subset, their system scores 64.4% versus 48.9% for human experts. Mule et al. (2026) forecast leaderboard outcomes with a consistency-aware metric whose random baseline is 25%, unlike our pairwise 50%. Park et al. (2025) forecast a benchmark's scores from its task description alone (the PRECOG corpus, scraped from published papers), and Luo et al. (2025) cover neuroscience. All four rely on publication-filtered ground truth without abandoned attempts. Prospective judgment already serves engineering, selecting candidates in agent search (Foster et al., 2026) and judging unexecuted solutions at 61.5% pairwise accuracy on the judge's own agent family, though only inside one self-generated scaffold; the inputs there resemble our four information levels (Zheng et al., 2026). BoleBench grades independent systems' parallel attempts under a hidden third-party evaluator and replays selection outside any judge's loop.

Judging artifacts: reviewers and idea feasibility. Reviewer models predict noisy, retrospective committee consensus (Cortes & Lawrence, 2021). ReviewGuard predicts citations; its rescue-rate study of selection value is genuine though retrospective (Rasool et al., 2026). ForeSci, our closest sub-skill precedent, grades four decision families against later publication (Tian et al., 2026). SoundnessBench finds pervasive proposal-stage optimism bias (Ho et al., 2026). TASTE asks the question closest to ours. It consists of 92 pairs of model-written AI safety proposals, labeled by expert preference after pair discussion and a confidence filter. The best model reaches 60% on this dataset against 77% estimated human agreement (Baig et al., 2026). Its proposals are never executed, so preference is the only available ground truth there. BoleBench instead judges decisions in running research against executed outcomes, not attention or preference. Feasibility is a special case of select-best; reject-below-baseline and weigh-evidence exist only mid-run, since publication-filtered data records neither discarded attempts nor the interim evidence that guided them.

Abstention and ungraded generation. Abstention benchmarks score knowing when not to act, on constructed labels where frontier models barely clear the 50% degenerate-policy ceiling (Liu et al., 2026). Our reject-below-baseline track is a harder test. It asks whether an abandoned direction, run to completion elsewhere, was worth abandoning. Judges' ratings need not predict executed outcomes. Expert-judged novelty of LLM-generated ideas does not survive execution (Si et al., 2025). LLM-judged novelty anti-correlates with bibliometric impact (Spearman -0.291) (Jiang, 2026), though the generator's training cutoff falls within the period used to measure outcomes.

Psychometrics and judge homogeneity. Our evaluation draws on established methods: item-response models and item-bank compression (Maia Polo et al., 2024), factor structure (Ilić & Gignac, 2024), and low-validity-environment criteria for trusting intuition (Kahneman & Klein, 2009). This battery of reliability, convergent, discriminant, and known-groups checks, run before any model is judged, has, to our knowledge, not previously been applied to a judgment construct. Errors correlate across LLM families (Kim et al., 2025; Goel et al., 2025). We quantify shared er-

rors in research judgment against parallel-attempt ground truth, examine reliance on the boldness heuristic, and measure the outcomes of judge-based selection. No work in Table 1 combines these elements. None reports that a one-line heuristic outperforms every LLM judge, or that judges fall below chance where the boldness prior favors the eventual loser.

3 THE BOLEBENCH FRAMEWORK AND RELEASE

BoleBench provides an evaluation procedure and a benchmark of executed research attempts. Each question in Section 1 varies one part of the same comparison (Fig. 1). A *task* τ is a frozen starting state with a hidden evaluator. An *attempt* a is one system's run on τ, with a prospective description $d(a)$ and a hidden score $\sigma(a)$ revealed after execution. Descriptions state the idea and its implementation with all outcomes removed. Judgment therefore concerns the executed attempt, not the abstract idea. An idea-only variant retains the mechanism while masking constants and identifiers to test dependence on implementation detail (Section B and Table 3). A *judge* J reads two blinded attempts on the same task, selects one, and reports a confidence. The hidden scores grade its choice.

Corpus and item construction. Corpora must provide hidden outcomes, parallel attempts from a frozen start, an open decision space, prospectively describable choices, dated outcomes, and matched budgets (Fig. 1). AI4AI-Bench and RSI-Exam meet all 6 criteria. Exclusions and partial matches are documented with the release. AI4AI contains 271 matched-budget attempts on 10 ML-research tasks, scored by sealed evaluator after every judge's training cutoff (Chi et al., 2026). Failed attempts remain in reject-below-baseline. The arXiv-task corpus contains 398 scored attempts by 15 models on 35 sealed tasks (Section B and Fig. 2). Pair preparation blinds descriptions, audits leakage, excludes near-ties, adds committee labels for boldness and change family, and reverses A/B order to test position consistency. This yields 1,842 AI4AI pairs. A pre-registered, seed-frozen sieve provides a cheaper ranking subset, while accuracy claims use appropriate full bank or an explicitly identified stratum. A subset labeled with text-based boldness proxies is used only to size the sieve. All reported heuristic-silent results instead use the 333 pairs tied under committee boldness labels.

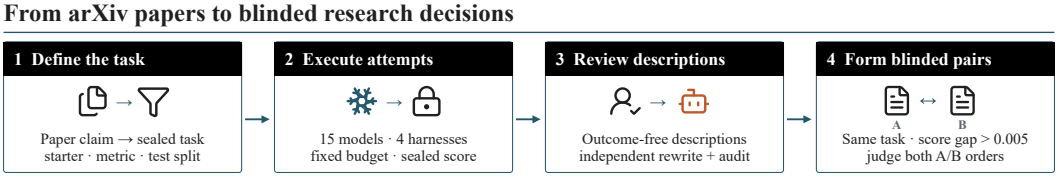


Figure 2: **Building the arXiv-task corpus.** Papers become sealed tasks for fixed-budget execution by 15 models. Outcome-free descriptions are reviewed and paired. Of 418 reviewed descriptions, 398 attempts form 1,897 released pairs.

Decision tracks and information. Each item tests one of five decisions. *Select-best* asks which of two attempts scores higher. *Reject-below-baseline* asks whether one attempt beats an expert-level baseline. *Select-ambitious* tests whether the judge correctly weighs a bold attempt (high-risk, high-reward) against a safe one. *Weigh-evidence* supplies exploration evidence, including traps where proxy evidence favors the eventual loser. *Override-prior* uses clear-boldness-gap pairs where the bolder attempt lost and the boldness prior is wrong. Chance is 50% on every pairwise track. AI4AI provides 1,842 select-best pairs, 271 reject-below-baseline items (one per attempt), 560 weigh-evidence items (158 traps), and 249 override-prior pairs (22% of the 1,127 clear-boldness-gap pairs). Each select-best item has four information levels: a 350-character gist, the full blinded description, description plus exploration evidence, and description plus system identity. Comparisons remain within stratum, so what the judge reads is not confounded with which items it sees.

Scoring and validation. Each judge receives a nine-axis report card covering its profile across the five tracks, information need, resolution (gap-stratified accuracy), heuristic-silent margin (accuracy beyond the boldness prior on heuristic-silent pairs), calibration (Brier/ECE), reliability (test–retest and mirror robustness), independence (error correlation across judges), selection value (one-of-k replay against random and oracle), and cross-corpus generalization (Section 4.1). Performance varies

across these measures, so a single ranking, including our leaderboard, can recommend the wrong judge for a particular use. Because judges commit before outcomes are revealed, their choices can be scored against realized outcomes. A single execution remains noisy and cannot separate chance from missing information. Appendix F gives the Bayesian argument, Brier-loss justification, and identification limits. Reliability, convergent, discriminant, and known-groups checks precede model comparison. Across three 200-item replicates, agreement ranges from 0.85 to 0.95, κ from 0.71 to 0.89, and accuracy stays within ± 1.3 points. A memorization quiz and release tags probe contamination. Identity ablation tests the effect of revealing the authoring system.

4 RESULTS

Judge cohorts, baselines, and controls are detailed in Section A. Judge strengths vary across the five decisions, so we read the aggregate leaderboard alongside the track profile (Fig. 9). Change-family payoffs appear in Fig. 8.

Figure 3 shows performance where the boldness prior provides no preference. On the 333 boldness-tied pairs, the best judge reaches 63.5%, while the weakest score near chance. The rule that scores 77.9% on clear-gap pairs cannot choose between equally bold attempts. Across all 4,059 pairs (Fig. 7 and Section D), claude-fable-5-1 leads at 70.8%. No judge clearly separates from the nine-judge majority, and the weakest approach the length baseline. We examine how information, the boldness prior, and decision skills affect accuracy and selection value. Chance is 50% for pairwise winner prediction.

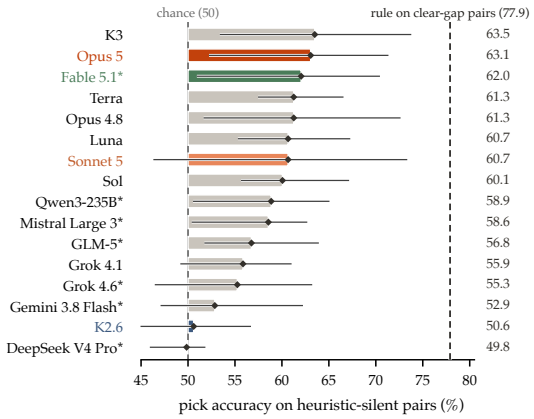


Figure 3: **Standings on boldness-tied pairs.** Pick accuracy on the 333 boldness-tied AI4AI pairs (full description; per-judge coverage 326–333, missing predictions excluded), where the one-line pick-the-bolder rule predicts nothing. For scale, the dashed line marks the rule's accuracy on clear-gap pairs. Whiskers are 95% task-clustered intervals; asterisks mark judges added after the original grid; colors highlight judges discussed in the text. The pooled leaderboard over all 4,059 pairs is in Fig. 7 (Section D).

The 350-character gist already supports much of each judge's accuracy, and full descriptions add at most five points. Adding exploration evidence improves accuracy for every grid judge, by up to 12.4 percentage points (Table 2). The identity condition instead supplies the description with system identity and reasoning effort, without exploration notes. Seven of eight judges score lower in this condition than with exploration evidence. For gpt-5.6-sol, the corresponding accuracies are 67.3% with the full description, 79.7% with exploration evidence, and 65.8% with identity. Opus 5 reaches 78.8% in both latter conditions. Because these conditions provide different information, their difference does not isolate the effect of revealing executor identity.

Most of the headline accuracy comes from the prior. On 1,127 boldness-different pairs, "pick the bolder" scores 77.9 and beats every full-description judge. On 333 boldness ties, the strongest judges retain only $+10$–14 points over chance, with Kimi-K2.6 near zero (Fig. 4).

Takeaway 1. Most select-best accuracy comes from a boldness prior. A one-line boldness rule beats every full-description judge on clear-gap pairs. When boldness is tied, the best judges are only 10 to 14 points above chance.

Decision tracks show failures that aggregate accuracy hides. On 158 misleading-evidence items, fourteen of fifteen judges score 23–45%, versus 80–94% when evidence points correctly. On 249 override-prior pairs where bolder attempt loses, all sixteen score 36.9–50.0% (Fig. 9 and Section H).

Rankings also reverse across tracks. Claude-opus-4-8 scores 60.6% on select-best but leads reject-below-baseline at 68.4% balanced accuracy. Claude-sonnet-5 leads select-ambitious at 73% but falls

to 23% on weigh-evidence traps. An overall select-best leaderboard therefore cannot replace testing judges on the specific decision a workflow requires. Full per-judge results appear in Section H.

Errors coincide far beyond independence. After audited bank defects are removed, all eight judges from four families (n=1,820) fail on 4.5% of pairs, roughly 135× the 0.033% expected under independent errors. The excess persists across model families. The failure taxonomy appears in Section H.

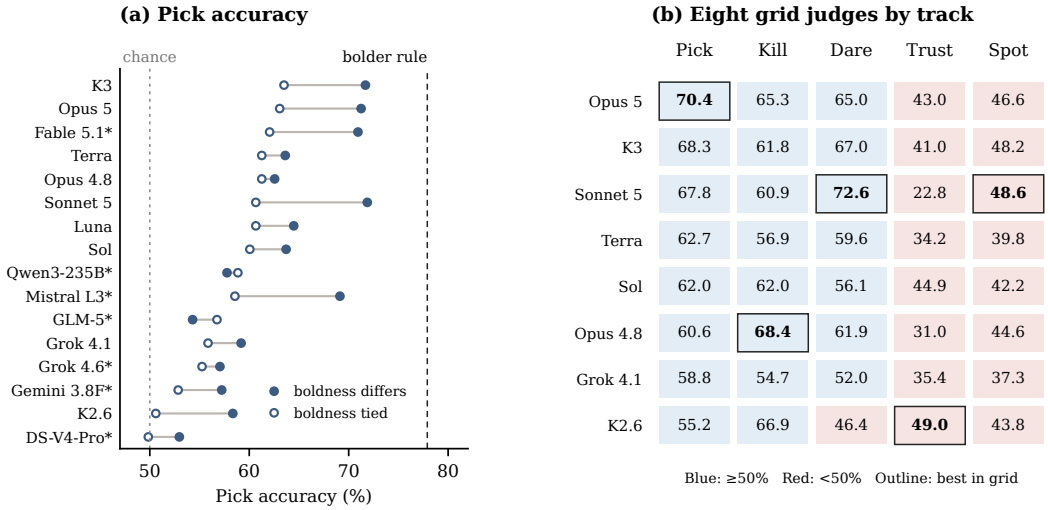


Figure 4: **Select-best accuracy reflects a boldness prior, and judge rankings vary across tracks.** (a) Clear-boldness-gap pairs (n = 1,127) and boldness-tied pairs (n = 333) for sixteen judges. Asterisks mark added judges. (b) Accuracy (%) for eight information-grid judges with complete five-track runs. Pick denotes select-best, Kill reject-below-baseline, Dare select-ambitious, Trust weigh-evidence, and Spot override-prior. Reject-below-baseline reports balanced accuracy. Weigh-evidence is restricted to trap items, and override-prior to pairs where the bolder attempt lost.

Takeaway 2. Aggregate accuracy hides track-specific failures. Fourteen of fifteen judges fall to 23% to 45% when exploration evidence is misleading, and no judge beats chance when the bolder attempt loses.

Table 2: **Exploration evidence improves accuracy across the grid judges.** Pick accuracy by information level on a stratum-balanced 600-pair subsample (chance = 50%; best value per column in bold; full-bank accuracies in the AI4AI column of Table 6).

Judge	gist (350c)	full	+exploration	+identity
claude-opus-5	**73.5**	**75.5**	78.8	**78.8**
kimi-k3	70.8	70.7	76.8	72.5
claude-sonnet-5	67.0	71.0	73.2	70.8
gpt-5.6-sol	63.7	67.3	**79.7**	65.8
gpt-5.6-terra	61.8	65.8	76.5	64.8
claude-opus-4-8	59.5	62.2	73.2	63.7
grok-4.1-fast	58.0	62.7	69.7	62.0
Kimi-K2.6	55.0	55.3	62.4	59.4

Executor and self-bias. In Fig. 6, we compare each judge's choices with realized outcomes on the *same* one-executor pairs. All sixteen judges under-select the strongest executor, claude-opus-5, by 4.5–29.8 points relative to its actual win rate, and over-select the weaker claude-sonnet-5 by 7.0–18.6 points. Own-model effects are mixed, not a uniform preference: kimi-k3 picks its own attempts 12.8 points above their win rate, whereas gpt-5.6-sol picks its own 9.5 points below. Because names are hidden here, the figure measures associations with executor-linked descriptions. It does not

measure a direct response to model names. The separate identity condition in Table 2 supplies system information, with further details in Section H.

Takeaway 3. Blinding does not remove executor-linked bias. Every judge discounts the strongest executor and favors the same weaker one, but own-model effects differ in direction.

Accuracy and selection value. Pairwise accuracy treats all errors equally, but selection mistakes cost more when outcome gaps are large (Section F). We therefore replay each judge's pairwise calls as a one-of-k selection policy and score the retained attempt. The leading policy, Opus 5, recovers about three quarters of the oracle-over-random improvement at k=10. Increasing the pool to twenty improves its retained score, but the oracle improves more, widening their score gap from 0.032 to 0.045. The recovered share of the available gain consequently falls from approximately 76% to 71%. Thus even an improving selection policy can leave a growing gap to the best available outcome. Other judges peak and then decline as pools grow, and the weakest falls below random selection (Fig. 5 and Table 4). These patterns show why increasing the number of candidates must be evaluated together with the policy used to select among them.

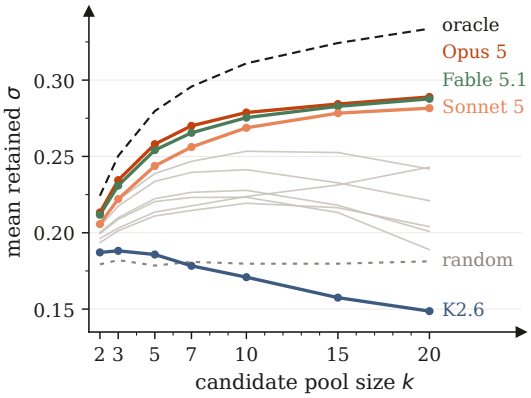


Figure 5: Selection quality diverges with pool size. Mean retained score σ in one-of-k replay for Opus 5, Fable 5.1, Sonnet 5, and Kimi-K2.6 (colored), other judges (gray), and oracle and random baselines. Exact values appear in Table 4.

We also asked judges to select the best of ten attempts directly from shuffled pools spanning AI4AI and arXiv tasks, each with a unique best attempt by sealed score. The leading judge selects the best in 36% of pools, above both chance and the longer-description baseline, and recovers just over half the available score gain (Table 5). Across the sixteen judges, direct selection averages 2.8 percentage points below Copeland replay using each judge's pairwise choices on the same pools. Direct selection also shows position bias. Some judges select the first candidate substantially less often than it is best.

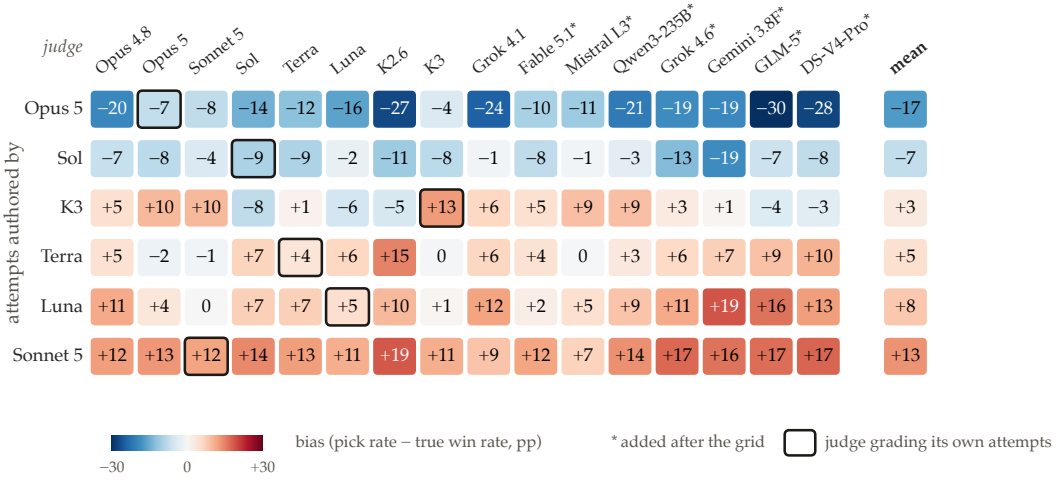


Figure 6: **Judges favor and discount the same authors.** Rows identify executor models and columns identify judges. Each cell maps a judge to an executor: the judge's pick rate minus that executor's actual win rate on pairs with exactly one attempt from that executor (n=141–648 per cell). Identities were hidden; outlined cells mark a judge evaluating its own model's attempts. The right column gives the mean across judges.

Takeaway 4. Pairwise accuracy alone does not determine selection value. Larger candidate pools improve outcomes for some judges but push others below random selection. Selection policies therefore need evaluation against random and oracle baselines.

4.1 BOLEBENCH ACROSS CORPORA

Transfer across corpora. We applied the same inclusion criteria to 35 tasks derived from 2026 arXiv papers, yielding 398 attempts and 1,897 blinded pairs judged in both A/B orders (Section B). Another 221 RSI-Exam and 99 paper-ablation pairs bring BoleBench to 4,059 pairs. Table 6 reports every judge on every corpus. Judge accuracies correlate at Spearman 0.62 between AI4AI and arXiv tasks over the canonical nine. Across the eight canonical judges shared by each comparison, correlations are 0.73 between AI4AI and RSI-Exam and 0.72 between RSI-Exam and arXiv tasks. Claude-opus-5 and claude-fable-5-1 remain in the top four on every agent-attempt corpus, while Kimi-K2.6 stays in the bottom three. Paper-ablation accuracies reach 76.8–94.9%, but rankings correlate less closely with other corpora (0.12–0.57). The longer description wins 51.5% of AI4AI pairs, 59.0% of arXiv-task pairs, and 74.5% of ablation pairs, so accuracy must be interpreted alongside this heuristic. The nine-judge majority reaches 67.1% overall, 66.6% on both AI4AI and arXiv tasks, and 70.1% on a balanced 99-per-corpus subset. The weakest judge remains within a few points of the length baseline. We also evaluated seven models released after the original grid on the whole bank. Claude-fable-5-1 leads at 70.8% pooled, while the added open-weight judges, qwen3-235b, glm-5, and DeepSeek-V4-Pro, rank in the bottom half. Thus judge ordering broadly transfers across agent-attempt corpora, although absolute accuracy depends on the corpus and its surface cues. Paper ablations provide weaker evidence of this transfer because high accuracy coincides with a strong length baseline. Both relative rankings and gains over simple heuristics matter.

Robustness and executor bias. We next examine whether stable rankings imply stable individual choices. All sixteen judges evaluated the arXiv-task pairs in both orders. Order-swap consistency ranges from 0.51 for mistral-large-3, which chooses the first description on 72–75% of pairs in either order, to 0.88 for claude-opus-5 and kimi-k3. Claude-fable-5-1 and gemini-3.8-flash reach 0.87. Strict accuracy, requiring correctness in both orders, peaks at 0.66 for claude-fable-5-1. On pairs containing exactly one own attempt, self-favoritism ranges from -0.19 to $+0.16$ after descriptions are rewritten in a common style. Kimi-k3, whose attempts were added last, scores $+0.03$. Rewriting lowers self-favoritism for nine of twelve judges with original agent-written descriptions, suggesting that prose style contributes to some measured self-preference under blinding (Section B). We further re-judged all 1,897 arXiv-task pairs using descriptions that retain the mechanism and intended change but mask constants as `[value]` and identifiers as `[name]`, reducing characters by 63% (Section B and Table 3). Across the canonical nine and claude-fable-5-1, pooled accuracy moves from 64.9 to 64.4%, a paired difference of -0.4 points $[-2.4, +1.4]$, despite 21.0% of choices flipping. Only claude-fable-5-1 has a decline whose interval lies below zero, from 71.6 to 69.2%. Rankings remain stable at Spearman 0.95, while the length baseline rises from 59.0 to 63.1%, narrowing the margin over the heuristic. Self-favoritism also persists after masking: claude-sonnet-5 remains at $+0.16$, while grok-4.1-fast moves from $+0.11$ to $+0.17$. Stable accuracy and rankings therefore conceal changes in individual decisions, and prose style and explicit identifiers do not fully explain executor-linked preferences.

Rationale analysis. Requesting rationales provides no average improvement. Across nine judges on a balanced 296-pair subset, mean accuracy falls 0.9 points, with individual changes from -4.4 to $+3.0$, while order consistency is unchanged. For the two judges exposing hidden reasoning, accuracy is lower with above-median reasoning length, at 62.6% versus 77.9% for K3 and 57.8% versus 69.8% for Kimi-K2.6. Every rationale cue family appears more often on errors, and judgments whose rationales mention ambition-related keywords are correct in 56.7% of cases. These associations do not establish that longer reasoning causes errors (Sections B and H).

Human comparison. Eleven valid submissions from ten researchers average 63.6% on a 60-pair subset, above six of eight matched judges. The pairs were divided into 20-pair slices with ten common anchors. Two speed-runs and one disavowed first attempt were excluded, and individual accuracy ranges from 45 to 85%. On the shared anchors, human and model majorities make identical choices, including the same three errors. This small sample provides a human reference

point but cannot reliably rank individual researchers against judges. The shared mistakes also show that agreement with the human majority does not establish correctness against executed outcomes. Protocol details and individual variation appear in Section H.

5 Analysis

After removing nine defective descriptions, 81 items remain jointly missed by all eight grid judges. The taxonomy points to overlooked execution discipline, misplaced priors, risk miscalibration, overfitting assumptions, and surface-text cues (Section H). Different model families therefore do not guarantee independent judgments, although shared difficulty can also explain excess joint error (Section F). Whether a cue helps depends on the research setting. Update-rule changes beat the expert baseline in all 32 observed attempts, compared with 35% for hyperparameter tweaks (Fig. 8). These improvement rates differ from pairwise accuracy, but motivate examining judgment by change family. Eighteen shared misses involve useful priors applied in the wrong regime. Likewise, trap subsets reveal failures when normally useful cues mislead, but cannot alone establish irrational updating when contrary evidence is unavailable.

Blinding also leaves systematic preferences. All sixteen judges underselect Opus 5 attempts and overselect Sonnet 5 attempts relative to actual win rates, while own-model effects vary (Fig. 6). Some self-preference persists after rewriting and identifier masking, so names and original prose style do not fully explain it. These associations do not establish model recognition. Nor does the lower accuracy of seven grid judges under identity information isolate a naming effect, because that condition omits exploration notes. These distinctions matter for deployment. Larger candidate pools improve some judges' selections while pushing others below random (Fig. 5). Directly choosing the best of ten remains difficult: the leading judge succeeds in 36% of pools against 10% chance, and direct selection averages 2.8 points below Copeland replay across sixteen judges (Table 5). More candidates help only if the selection policy can recover their value, so pairwise accuracy needs to be read alongside realized outcomes.

6 Conclusion

BoleBench evaluates research judgment against executed outcomes using blinded, matched attempts, five decision tracks, four information levels, and selection replay. Validity checks audit the item bank, and fresh corpora allow the benchmark to be renewed as models retrain. The results show a gap between predicting winners and reliably selecting experiments. On AI4AI pairs with clear boldness differences, a simple boldness rule beats every full-description judge, while no judge exceeds chance when the bolder attempt loses. Judges retain a modest advantage when boldness is tied, but eight judges across four model families share errors roughly 135 times more often than expected under independence. Even with names hidden, all sixteen judges underselect the strongest executor and overselect the same weaker one, so blinding does not eliminate executor-linked preferences. These weaknesses have practical consequences. At a candidate pool size of ten, the best replay policy recovers about three quarters of the oracle-over-random gain, while the weakest falls below random selection. Asked to choose directly among ten attempts, the leading judge identifies the best in 36% of pools against 10% chance, leaving substantial room for improvement. Judge rankings broadly transfer across the agent-attempt corpora, although strengths vary by decision track and requesting rationales does not improve average accuracy. The presentation checks also show why stable rankings are insufficient. Masking implementation details changes 21% of individual choices while leaving pooled accuracy nearly unchanged, so aggregate stability can conceal sensitivity in particular decisions. Likewise, human and model majorities agree on the ten shared anchors but make the same three errors. Agreement and stability are therefore useful diagnostics, but neither alone establishes that a judge consistently selects the attempts with better executed outcomes. Reliable research judgment therefore requires evaluation beyond an overall ranking, including performance when simple cues fail, shared mistakes, and the realized outcomes of selected experiments.

Reproducibility Statement

We release the complete evidence chain behind every number in the paper: the four item banks (blinded descriptions with executed outcomes as the answer key), every judge prediction file behind the tables and figures, the seed-frozen sieve, and the scripts that rebuild the pooled bank and the leaderboard from the raw runs. The construction procedure for tasks and trajectories will be open-sourced in full: the corpus inclusion gates (Section 3), the arXiv task distillation and fixed-budget rollout procedure with its description review and rewrite protocol and per-attempt ledger (Section B), and the human-baseline exam with its scoring and exclusion rules. Sampling steps state their frozen seeds where they occur, and a datasheet accompanies the item bank.

AI Use Statement

Large language models are both the object of study and, where stated in the paper, components of the research pipeline: they served as the judges under evaluation and powered the reviewer and rewrite sessions used to produce the blinded descriptions (Section B). LLM assistants also helped screen candidate items and generate rollout data used to construct a question bank supporting subsequent research ideation and experimental execution. Separately, LLM assistants supported manuscript preparation by drafting and revising text under the authors' direction, checking numerical consistency across sections, tables, and figures, and generating figure code. The final research questions, benchmark design, data-selection decisions, analyses, and conclusions were determined by the authors. All AI-assisted outputs were reviewed, edited, and approved by the authors, who take full responsibility for the content.

Ethics Statement

BoleBench evaluates judgments about executed research attempts. Its results concern the sampled tasks, descriptions, and evaluation conditions, and should not be treated as a general ranking of researchers or research ability. The human comparison is small and is reported as a reference point. The observed shared errors and sensitivity to presentation also warrant caution when using model judgments to allocate research resources.

References

Hasan Baig, Hailey Joren, and Joe Benton. TASTE: Can AI models judge AI safety research proposals? Anthropic Alignment Science Blog, 2026. URL https://alignment.anthropic.com/2026/taste/.

Yizhe Chi, Wenyi Li, Deyao Hong, Xiaoqiu Wang, Mingju Gao, Kaisen Yang, Bingxiang He, Youjie Zheng, Calvin Xiao, and Qinhuai Na. AI4AI-Bench: Benchmarking LLM agents in algorithmic design for recursive self-improvement. *arXiv preprint arXiv:2608.20318*, 2026.

Corinna Cortes and Neil D Lawrence. Inconsistency in conference peer review: Revisiting the 2014 neurips experiment. *arXiv preprint arXiv:2109.09774*, 2021.

Lee J Cronbach and Paul E Meehl. Construct validity in psychological tests. *Psychological Bulletin*, 52(4):281–302, 1955.

Thomas Simon Foster, Bassel Al Omari, Tingchen Fu, Thomas Mann, Carl Domond, et al. AI research preference models. *arXiv preprint arXiv:2608.13940*, 2026.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

Shashwat Goel et al. Great models think alike and this undermines AI oversight. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025. arXiv:2502.04313.

Richard W Hamming. You and your research. Talk at Bell Communications Research, 1986.

Sy-Tuyen Ho, Minghui Liu, Huy Nghiem, and Furong Huang. SoundnessBench: Can your AI scientist really tell good research ideas from bad ones? *arXiv preprint arXiv:2605.30329*, 2026.

David Ilić and Gilles E Gignac. Evidence of interrelated cognitive-like capabilities in large language models: Indications of artificial general intelligence or achievement? *Intelligence*, 106:101858, 2024. arXiv:2310.11616.

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

Keller Jordan and contributors. modded-nanogpt: NanoGPT speedrun. https://github.com/KellerJordan/modded-nanogpt, 2024.

Daniel Kahneman and Gary Klein. Conditions for intuitive expertise: A failure to disagree. *American Psychologist*, 64(6):515–526, 2009.

Elliot Kim, Avi Garg, Kenny Peng, and Nikhil Garg. Correlated errors in large language models. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025. arXiv:2506.07962.

Xun Liu et al. Agentabstain: Do LLM agents know when not to act? *arXiv preprint arXiv:2607.10059*, 2026.

Xiaoliang Luo et al. Large language models surpass human experts in predicting neuroscience results. *Nature Human Behaviour*, 9:305–315, 2025.

Parv Mahajan, Yilin, and yix. TastyBench: Toward measuring research taste in LLM. LessWrong, Dec 2025, 2025.

Felipe Maia Polo et al. tinybenchmarks: evaluating LLMs with fewer examples. In *ICML*, 2024. arXiv:2402.14992.

Srujan P Mule, Aniketh Garikaparthi, and Manasi Patwardhan. Teaching language models to forecast research success through comparative idea evaluation. *ACL Findings*, 2026. arXiv:2605.21491.

Neel Nanda. My research process: Understanding and cultivating research taste. AI Alignment Forum, 2025. URL https://www.alignmentforum.org/posts/Ldrss6o3tiKT6NdMm/my-research-process-understanding-and-cultivating-research.

Jungsoo Park, Ethan Mendes, Gabriel Stanovsky, and Alan Ritter. Anticipatory evaluation of language models, 2025. URL https://arxiv.org/abs/2509.20645.

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

Chenglei Si, Tatsunori Hashimoto, and Diyi Yang. The ideation-execution gap: Execution outcomes of LLM-generated versus human research ideas. *arXiv preprint arXiv:2506.20803*, 2025.

Qiuyu Tian, Haojie Yin, Yingce Xia, Youyong Kong, and Zequn Liu. ForeSci: Evaluating LLM agents for forward-looking AI research judgment. *arXiv preprint arXiv:2606.00644*, 2026.

Jiaxin Wen, Chenglei Si, Chen Yueh-Han, He He, and Shi Feng. Predicting empirical AI research outcomes with language models. In *NeurIPS*, 2025. arXiv:2506.00794.

Jingsheng Zheng, Jintian Zhang, Yujie Luo, Yuren Mao, Yunjun Gao, Lun Du, Huajun Chen, and Ningyu Zhang. Can we predict before executing machine learning agents? *arXiv preprint arXiv:2601.05930*, 2026.

A Experimental Setup

Judges and cohorts. Three cohorts recur. The eight information-grid judges span four model families (one accessed via proxy) and have runs at every pick-information level. The canonical nine add gpt-5.6-luna (full level only) and supply the majority votes. The sixteen leaderboard judges add seven post-grid models (Section 4.1); decision-track coverage varies by judge, and claude-fable-5-1 has no AI4AI exploration-level run. We use abbreviated release names, e.g., claude-opus-5 as Opus 5 and kimi-k3 as K3.

Baselines. Baselines are random choice, the one-line boldness heuristic, and the longer-description baseline (pick whichever description is longer). Researchers judged a matched 60-pair subset as a human baseline (Section 4.1).

Controls and inference. A memorization quiz probes contamination; identity ablation runs on the eight information-grid judges, not every leaderboard judge. Outcome ties are excluded, and a task-cluster bootstrap estimates uncertainty. Information-level comparisons use a stratum-balanced 600-pair subsample; full-bank accuracies appear in the AI4AI column of Table 6.

B The arXiv-Task Corpus: Construction and Review

Corpus. The arXiv-task corpus holds 35 tasks, each distilled from one arXiv paper posted in 2026 (17 STEM fields). The paper's central quantitative claim becomes a task with a starter program, one hidden metric, a sealed test split and a verifier. The system attempting the task sees the task statement and the starter but never the sealed data. A companion paper (in preparation) describes how papers are selected and turned into verified tasks; here the tasks are taken as given (Fig. 2). Fifteen models attempted the tasks under four coding harnesses (a Claude Code panel and three external harnesses), with several attempts per task. The fifteenth, kimi-k3, was added after the first review round. It attempted every task once under one of the external harnesses, on the same budgets as the other models, and 20 of its 35 attempts met the same eligibility criteria. They added 196 pairs; their descriptions underwent the same review and rewrite procedure. Each attempt is scored on the sealed split, and scores are normalized so the shipped starter maps to 0 and a strong reference to 1. Two tasks whose environment image exposed sealed statistics and one attempt whose final score depended on machine load were excluded. After these exclusions, 398 attempts remain in the shipped pairs (418 candidate descriptions were reviewed, including those on the excluded tasks).

Descriptions. Each attempt ships as a structured, outcome-free description of the graded program (mechanism, what changed versus the starter, what stayed, constants, risk signals). Every description was checked by a reviewer (a model session with read-only access to the graded program, the starter and the agent's own notes). A second session then rewrote the description from the program with the reviewer's findings in hand. An independent third session re-checked the rewrite under the same rules, with at most one further revision round. Of the 418 reviewed descriptions, 301 were accepted by their last reviewer as written and 117 had the last reviewer's flagged sentences removed; none failed fidelity. Of the 398 shipped attempts, 288 were accepted as written and 110 had flagged sentences removed. Descriptions contain no outcomes, file citations, or names of verifiers, models or harnesses. The per-attempt ledger and per-pair records are released with the item bank.

Pairs and judging. Pairs of attempts at the same task whose normalized scores differ by more than 0.005 form 1,897 items (899 across harnesses, 998 within). Gold labels follow the higher sealed score. Twenty-one judges answered the pairs in both A/B orders with the same prompt and a fixed per-judge output cap (Kimi-K2.6 needed 32k tokens on these longer descriptions). They comprise the corpus's own fourteen-judge grid plus the seven judges added after it; the grid overlaps with the sixteen-judge leaderboard cohort but differs from it. One judge, claude-opus-5, returned no parseable answer on 6 order-judgments; these stay missing. On this corpus, the nine-judge majority has 66.6% accuracy (n=1,897), mean judge accuracy is 63.7%, and the longer-description baseline is 59.0%. Over these 21 judges, order-swap consistency runs from 0.51 to 0.89 (gemini-3.8-flash), with DeepSeek-V4-Flash at 0.57. The low end is mistral-large-3, which picks the first-shown description in 72 to 75% of pairs in either order, so its picks mostly reflect position. Rewriting

every description in one house style lowered self-favoritism for nine of the twelve judges with own attempts, relative to the same pairs in their original agent-written form (claude-opus-5 $-0.19 \rightarrow -0.07$, gemini-3.8-flash $+0.11 \rightarrow +0.05$, Kimi-K2.6 $+0.10 \rightarrow +0.01$, claude-sonnet-5 $+0.20 \rightarrow +0.16$). Judges gpt-5.6-terra, gpt-5.6-luna and claude-haiku-4-5 moved the other way (by 0.03, 0.05 and 0.01). Attempts by kimi-k3 have no original-form counterpart; its self-favoritism on the rewritten descriptions is $+0.03$ (n=196).

Rationale run. A 296-pair subset (about 75 per corpus, stratified by score gap) was re-judged by the canonical nine, this time with a rationale requested before the JSON answer. Judging ran in both orders with the per-corpus prompts otherwise unchanged. We store rationales and token counts, plus any hidden reasoning the API returns. Against the no-rationale picks on the same pairs, accuracy changed as follows: claude-sonnet-5 -4.4, claude-opus-5 -3.1, kimi-k3 -3.1, grok-4.1-fast -2.2 (on the 225 pairs its no-rationale grid covers), Kimi-K2.6 -0.3, gpt-5.6 (three runs) within ± 0.7, claude-opus-4-8 $+3.0$.

Idea-only judging. To test how much of the judgment rests on implementation detail, we rewrote every arXiv-task description to state only the idea (the mechanism and what the change does relative to the starter). Numeric constants were replaced by `[value]` and identifiers by `[name]`. Mean length falls from about 3,860 to about 1,450 characters (63% fewer). The same 1,897 pairs, prompt and answer format were used. The canonical nine and claude-fable-5-1 judged the idea-only pairs in normal order; claude-opus-5, gpt-5.6-sol and claude-fable-5-1 also judged them in swapped order. Comparisons pair each judge's idea-only and full-description picks on the same items, with intervals from a 1,000-resample bootstrap over the 35 tasks (Table 3). Pooled over the ten judges, accuracy moves from 64.9 to 64.4% (-0.4 points [-2.4, $+1.4$]) while 21.0% of picks flip, and judges keep, on average, 97% of their margin over chance. Only claude-fable-5-1 has a paired-difference interval below zero, and its loss falls in the widest-gap tercile of pairs (-4.3 points [-8.0, -1.5]). The pooled differences by score-gap tercile are $+0.2$, -0.9 and -0.7 points. Order-swap consistency is unchanged for the three judges run in both orders (claude-opus-5 0.88 to 0.89, gpt-5.6-sol 0.86 to 0.86, claude-fable-5-1 0.87 to 0.87); claude-fable-5-1's strict both-order accuracy falls from 65.7 to 63.2% [-4.9, -0.3]. We also split pairs by whether an attempt crashed or otherwise failed (1,794 clean, 103 with a failed attempt). The pooled difference is -0.6 points [-2.5, $+1.4$] on clean pairs and $+2.5$ [-2.8, $+7.0$] on failure pairs. Removing execution evidence therefore does not cost accuracy on failure pairs, where that evidence should matter most. Self-favoritism on pairs with exactly one own attempt persists once identifiers are masked: claude-sonnet-5 $+0.16 \rightarrow +0.16$ (n=260), grok-4.1-fast $+0.11 \rightarrow +0.17$ (n=327), claude-opus-5 $-0.07 \rightarrow -0.16$ (n=218), gpt-5.6-luna $-0.19 \rightarrow -0.23$ (n=252), claude-opus-4-8 $+0.08 \rightarrow +0.03$ (n=292). The change in own-pick rate between the two renderings is within ± 0.06 for every judge except claude-opus-5 (-0.08).

Context from other benchmarks. Neighboring research-taste evaluations report accuracies in the 61.5–77% range, but these raw scores cannot be compared with ours or with one another. Their ground truths differ (publication records, leaderboards, a scaffold's own executions), as do their chance floors (near 25% for consistency-aware accuracy, 50% for pairwise accuracy). None of these setups exposes failed attempts. Fine-tuned judges reach 77% pairwise accuracy on published idea pairs (Wen et al., 2025), 77.1% consistency-aware accuracy on leaderboard forecasts (Mule et al., 2026), and 61.5% pairwise accuracy on execution-skipping solution preference (Zheng et al., 2026). In-loop research preference models sit at 65–69% (Foster et al., 2026). The tasks, ground truths and chance baselines behind these headline accuracies all differ, so the numbers are not directly comparable.

C Selection Replay at Extended Pool Sizes

Table 4 tabulates the curves of Fig. 5. For each AI4AI task and each pool size k, we sample k of the task's executed attempts (2,000 draws per task, seed frozen). Random picks an attempt uniformly, and the oracle keeps the sampled attempt with the highest sealed score. Each judge runs a round robin over its full-level pairwise calls and picks the candidate with the most wins, with ties going to the first sampled candidate. Values are the mean sealed score σ of the retained attempt; on this scale the shipped expert baseline sits at 0.1. Judges differ in how they scale with pool size. The top

Table 3: **Idea-only judging on the arXiv-task corpus.** Pick accuracy (%) on the same 1,897 pairs when the judge reads the full description versus the idea-only rendering (constants and identifiers masked). Difference = idea-only minus full, with a 95% task-bootstrap interval; flip = share of items on which the pick changed. Items are in normal A/B order. claude-opus-5 has 1,895 paired items, and the longer-description baseline excludes two pairs tied in length. Chance = 50%; – = not applicable.

Judge	Full	Idea-only	Difference [95% CI]	Flip (%)
claude-fable-5-1	71.6	69.2	-2.4 [-4.7, -0.1]	18.5
gpt-5.6-sol	70.0	68.8	-1.2 [-3.9, $+1.6$]	20.8
kimi-k3	69.8	71.1	$+1.3$ [-1.6, $+3.8$]	16.6
claude-opus-5	67.7	68.0	$+0.3$ [-2.6, $+3.3$]	16.4
gpt-5.6-terra	65.2	65.2	$+0.1$ [-3.1, $+3.3$]	19.2
claude-sonnet-5	64.1	64.5	$+0.4$ [-2.3, $+3.2$]	19.8
grok-4.1-fast	62.0	61.8	-0.2 [-3.5, $+3.2$]	26.3
gpt-5.6-luna	61.2	60.1	-1.1 [-4.2, $+2.0$]	24.5
Kimi-K2.6	59.0	57.4	-1.6 [-4.2, $+1.3$]	26.4
claude-opus-4-8	58.0	58.1	$+0.2$ [-2.7, $+3.2$]	22.0
Pooled (10 judges)	64.9	64.4	-0.4 [-2.4, $+1.4$]	21.0
Canonical nine-judge majority	66.6	67.1	$+0.5$ [-2.7, $+3.5$]	–
Longer-description baseline	59.0	63.1	$+4.0$ [$+1.6$, $+6.6$]	–

three curves keep climbing through k=20, and most mid-table judges peak near k=10 and decline as pools grow. Kimi-K2.6 declines beyond k=3 and crosses below random at k=7.

Direct pick of the best of ten. Table 5 uses a different protocol from the replay above. Each of 603 pools holds ten real attempts at one task and has a unique best by sealed score (200 AI4AI pools from 10 tasks, 403 arXiv-task pools from 26 tasks). The candidates are shuffled, one prompt shows all ten blinded descriptions, and the judge names the one it expects to have scored highest. Rates are task-weighted (every task counts once), with 1,000-resample task bootstraps stratified by corpus. Pools where a judge returned no valid answer are dropped from its denominator (one pool for claude-fable-5-1, none for the other judges). The Copeland column scores each attempt in a pool by wins minus losses under the same judge's pairwise picks, with unjudged pairs contributing zero and m-way ties earning $1/m$. The bank holds on average 23.1 of the 45 AI4AI pairs per pool and 41.4 of 45 on the arXiv tasks, so perfect pairwise picks recover the best in only 86.9% of pools. Averaged over the 16 judges, the direct pick falls 2.8 points below the Copeland replay (6.3 points on the 448 pools where perfect pairwise picks recover the best uniquely). The direct pick and the Copeland replay name the same attempt in 43.9% of pools and rank the judges alike (Spearman 0.86). Direct rates also track pooled pairwise accuracy (Spearman 0.89). Position bias is strongest for claude-sonnet-5 and grok-4.1-fast, which choose the first-listed candidate in 5.3% of pools, against 9.5% of pools where it is the best ($p < 10^{-3}$). The distribution of chosen positions is also non-uniform for mistral-large-3, qwen3-235b-2507, claude-opus-4-8, glm-5 and DeepSeek-V4-Pro (chi-square $p < 0.01$).

Table 4: **Mean retained σ by selection policy and pool size k.** One value per policy and pool size, with the best judge in each column shown in bold. Oracle and random provide reference baselines. Claude-fable-5-1 was evaluated on the same items after the original grid.

Selection policy	k=2	k=3	k=5	k=7	k=10	k=15	k=20
Oracle (ceiling)	0.224	0.250	0.280	0.296	0.311	0.324	0.334
claude-opus-5	**0.213**	**0.234**	**0.258**	**0.270**	**0.279**	**0.284**	**0.289**
claude-fable-5-1	0.212	0.231	0.254	0.265	0.275	0.283	0.288
claude-sonnet-5	0.206	0.222	0.244	0.256	0.269	0.278	0.282
kimi-k3	0.206	0.222	0.239	0.247	0.253	0.253	0.242
gpt-5.6-sol	0.204	0.218	0.234	0.240	0.241	0.233	0.221
gpt-5.6-terra	0.200	0.210	0.222	0.226	0.228	0.218	0.201
gpt-5.6-luna	0.199	0.209	0.220	0.223	0.223	0.213	0.189
claude-opus-4-8	0.196	0.203	0.214	0.218	0.224	0.231	0.243
grok-4.1-fast	0.193	0.201	0.211	0.215	0.219	0.216	0.204
Kimi-K2.6	0.187	0.188	0.186	0.178	0.171	0.158	0.149
Random	0.179	0.182	0.179	0.181	0.180	0.180	0.181

Table 5: **Asked to pick the best of ten attempts, the best judge is right 36.0% of the time, against 10.0 for chance.** The evaluation uses 603 pools of ten real attempts at one task (200 AI4AI, 403 arXiv-task), each with a unique best by sealed score. For each pool, one prompt shows all ten blinded descriptions in shuffled order, and the judge names the one it expects to have scored highest. Direct pick is the task-weighted share of pools in which the named attempt is the best; brackets give the 95% task-bootstrap interval. Gap share is the share of the random-to-oracle gain in sealed score that the pick recovers (random 0, oracle 100). Copeland replay scores each attempt in the pool by wins minus losses under the same judge's pairwise picks; with perfect pairwise picks it reaches only 86.9 because the bank judged about half of the AI4AI pairs. Bold marks each column's best judge.

Judge	Direct pick (%)	Gap share (%)	Copeland replay (%)
claude-fable-5-1	**36.0** [25.5, 47.0]	53.1	35.7
kimi-k3	35.4 [25.4, 46.1]	53.3	**39.5**
claude-opus-5	33.5 [24.0, 44.2]	48.9	36.4
gpt-5.6-sol	31.9 [22.2, 42.6]	**53.7**	35.7
claude-sonnet-5	30.6 [21.9, 39.5]	47.9	37.4
gpt-5.6-terra	30.5 [20.6, 40.8]	49.4	34.1
mistral-large-3	28.4 [20.3, 37.2]	36.5	27.8
grok-4.6	26.1 [17.4, 35.5]	43.6	23.2
qwen3-235b-2507	24.6 [17.0, 33.2]	40.7	32.4
grok-4.1-fast	23.4 [16.7, 30.7]	42.7	32.0
gemini-3.8-flash	22.5 [15.5, 29.7]	37.5	29.7
claude-opus-4-8	22.4 [15.6, 29.4]	38.4	20.1
glm-5	22.2 [15.3, 29.4]	36.1	25.2
Kimi-K2.6	21.4 [14.5, 28.8]	38.1	20.5
gpt-5.6-luna	20.3 [14.4, 26.7]	45.6	26.7
DeepSeek-V4-Pro	20.3 [14.2, 27.1]	28.0	18.1
Chance	10.0	0.0	–
Longest description in pool	21.3 [12.5, 30.7]	19.7	–
Canonical-nine plurality of direct picks	32.5 [22.9, 42.7]	51.3	–
Copeland with perfect pairwise picks (ceiling)	–	96.8	86.9

D PER-CORPUS LEADERBOARD AND CONSISTENCY

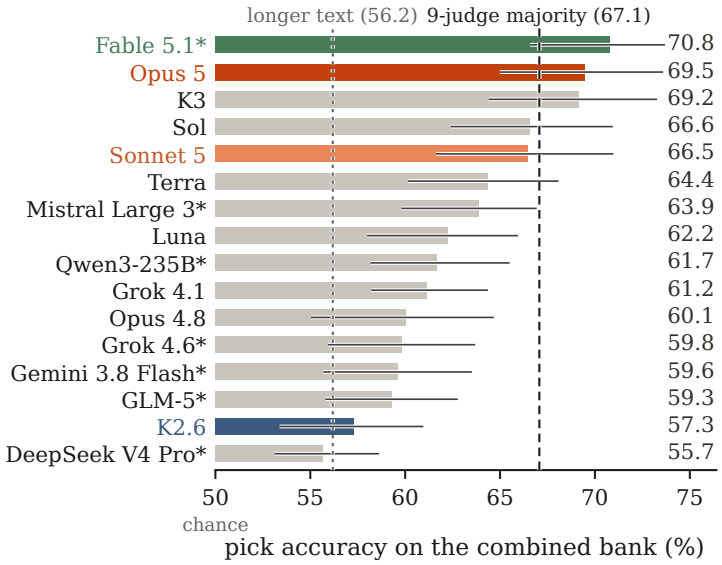


Figure 7: **Sixteen judges on 4,059 pairs.** Pooled pick accuracy with 95% task-clustered intervals. Full results appear in Table 6. The main text presents the heuristic-silent standings on boldness-tied AI4AI pairs (Fig. 3).

Table 6: **The judge ordering is stable across corpora, and the newest frontier model has the highest pooled accuracy.** Pick accuracy (%) per corpus; "all" pools each judge's available pairs from the 4,059-pair bank (chance = 50%). Balanced = 99 pairs per corpus; swap = agreement between the two A/B orders and self-fav. = P(pick own attempt) − P(own attempt wins), both measured on the arXiv-task corpus. Groups sorted by pooled accuracy; – = not run (grok-4.1-fast on RSI-Exam) or undefined (self-favoritism needs own attempts in the corpus, and the seven added judges have none). Longer-description baseline: 51.5 / 59.0 / 62.6 / 74.5 / 56.2 (AI4AI / arXiv tasks / RSI-Exam / abl. / all); canonical nine-judge majority: 66.6 / 66.6 / 62.4 / 93.9 / 67.1.

Judge	AI4AI	arXiv tasks	RSI-Exam	abl.	all	balanced	swap	self-fav.
Canonical nine								
claude-opus-5	70.4	67.7	66.5	92.9	69.5	71.0	0.88	−0.07
kimi-k3	68.3	69.8	62.4	88.9	69.2	68.7	0.88	+0.03
gpt-5.6-sol	62.0	70.0	63.8	91.9	66.6	71.7	0.86	−0.04
claude-sonnet-5	67.8	64.1	63.8	93.9	66.5	69.2	0.81	+0.16
gpt-5.6-terra	62.7	65.2	59.7	89.9	64.4	68.4	0.85	−0.12
gpt-5.6-luna	62.6	61.2	56.6	87.9	62.2	65.9	0.79	−0.19
grok-4.1-fast	58.8	62.0	–	88.9	61.2	64.3	0.75	+0.11
claude-opus-4-8	60.6	58.0	57.9	94.9	60.1	66.2	0.78	+0.08
Kimi-K2.6	55.2	59.0	51.1	76.8	57.3	60.6	0.81	+0.01
Added after the original grid								
claude-fable-5-1	69.5	71.6	65.6	91.7	70.8	68.6	0.87	–
mistral-large-3	66.8	59.7	63.8	90.9	63.9	70.2	0.51	–
qwen3-235b-2507	57.2	64.9	61.1	84.8	61.7	64.6	0.76	–
grok-4.6	56.3	62.1	57.0	87.9	59.8	65.4	0.84	–
gemini-3.8-flash	56.2	61.6	57.0	89.9	59.6	65.7	0.87	–
glm-5	54.3	62.6	61.1	83.8	59.3	67.7	0.74	–
DeepSeek-V4-Pro	51.2	58.2	57.9	84.8	55.7	58.8	0.63	–

E CHANGE-FAMILY PAYOFF AND JUDGE ACCURACY

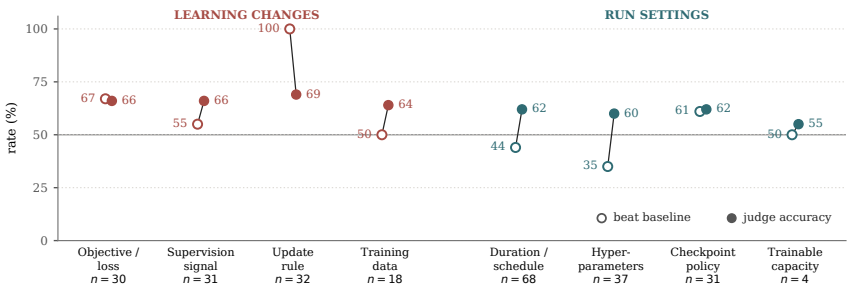


Figure 8: **Judge assessments of research changes track realized payoff only loosely.** Open markers show the fraction beating the expert baseline; filled markers show mean judge accuracy on associated pairs. Labels give the number of attempts in each change family (n=4–68).

F WHY HINDSIGHT CAN GRADE FORESIGHT: A BAYESIAN ACCOUNT

Judges commit before hidden outcomes are revealed, so execution outcomes can be used to score their choices and probabilities. Brier loss is strictly proper. In expectation, truthful probabilities minimize loss (Gneiting & Raftery, 2007). A good forecast can still lose on one run, and a single execution cannot separate outcome noise from missing information.

To ask what judgment adds beyond simple cues, write the visible information as $X = (H, Z)$. Here, H contains predefined cues and Z contains the remaining evidence. With $p_H = \Pr(Y = 1 \mid H)$ and $p = \Pr(Y = 1 \mid H, Z)$, Bayes' rule gives

$$\log \frac{p}{1-p} = \log \frac{p_H}{1-p_H} + \log \frac{f(Z \mid Y = 1, H)}{f(Z \mid Y = 0, H)}. \tag{1}$$

The question is whether Z changes the odds once the cues are known. Marginal cue accuracy does not determine this conditional weight. Our hard-choice analyses measure departures from the boldness heuristic. The sections below give the Brier decomposition, identification limits, and cautions for trap subsets and shared item difficulty.

Setup. Fix a distribution over tasks and pairs that meet the inclusion criteria, with $Y = \mathbf{1}\{\sigma(a) > \sigma(a')\}$. At information level ℓ, the judge sees X_ℓ, reports a probability $q_{J,\ell}(X_\ell)$, and commits before learning Y. If the reported confidence in the chosen answer is c, use $q = c$ after choosing a and $q = 1 - c$ after choosing a'. Write $p_\ell = \Pr(Y = 1 \mid X_\ell)$ and $R_\ell(q) = \mathbb{E}[(q(X_\ell) - Y)^2]$. Comparisons across levels use the same item distribution. These are population identities; pairs need not be independent. Sampling uncertainty must nevertheless account for shared attempts and tasks. If the judge is stochastic, assume its randomization is independent of Y conditional on its input, and include that randomization in the expectations. Selecting pairs by score gap changes the target distribution, so the results concern eligible pairs, not all possible proposals. For reject-below-baseline, Y denotes whether an attempt beats the expert baseline; the same scoring argument applies. Select-ambitious, weigh-evidence, and override-prior refer to their selected subsets, not the full pair distribution.

Why Brier loss. For a fixed input, write $p = p_\ell(X_\ell)$ and let $r \in [0, 1]$ be a reported winning probability. Since Y is binary,

$$\mathbb{E}[(r - Y)^2 \mid X_\ell] = p(1 - r)^2 + (1 - p)r^2 = p(1 - p) + (r - p)^2.$$

The first term does not depend on the report. The second is nonnegative and zero only when $r = p$. This proves strict propriety. Substituting the judge's report and averaging over inputs gives

$$R_\ell(q_{J,\ell}) = \mathbb{E}[p_\ell(1 - p_\ell)] + \mathbb{E}[(q_{J,\ell} - p_\ell)^2]. \tag{2}$$

We use Brier loss because this calculation separates uncertainty left by the input from error in the reported probability. The loss also stays between zero and one, even for completely confident reports.

Other strictly proper losses also reward accurate probabilities; the squared form makes this separation explicit. A good forecast can still score poorly on a single run. The guarantee concerns expected loss, and the residual uncertainty includes both missing information and execution randomness.

Conditional Bayesian odds. Suppress the level index and write the available information as $X = (H, Z)$, where $H = h(X)$ contains the chosen baseline cues and Z is the remaining information. Bayes' rule gives

$$\frac{\Pr(Y = 1 \mid H, Z)}{\Pr(Y = 0 \mid H, Z)} = \frac{\Pr(Y = 1 \mid H)}{\Pr(Y = 0 \mid H)} \frac{f(Z \mid Y = 1, H)}{f(Z \mid Y = 0, H)}.$$

Here f denotes a probability mass function or density. Taking logs yields (1) wherever the terms are finite. No independence between cues is required. This formulation assumes H is determined by the supplied input. Committee labels used only by the evaluator support stratified diagnosis; they need not be cues explicitly shown to a judge. In particular, evidence must be evaluated conditional on the cues already observed; its marginal accuracy does not determine its conditional weight. In our observed subsets, the boldness heuristic is correct on 77.9% of 1,127 eligible pairs, and exploration evidence on 402/560 items (71.8%). These marginal accuracies use different subsets and cannot determine the likelihood ratio in (1).

Improvement beyond baseline cues. Let $p_H = \Pr(Y = 1 \mid H)$ and $p = \Pr(Y = 1 \mid X)$. Applying the Brier decomposition to p_H and to q_J cancels the common irreducible term, giving

$$R(p_H) - R(q_J) = \mathbb{E}[(p - p_H)^2] - \mathbb{E}[(q_J - p)^2]. \tag{3}$$

Its first term measures predictive signal beyond H; its second is the judge's error in using the full input. This comparison uses the best predictor based on H. Beating an arbitrary heuristic can instead reflect better calibration or better use of those same cues. Accuracy on boldness-tied pairs therefore measures performance on a restricted subset; it does not establish an upper bound on every other form of useful judgment. For empirical use, specify H in advance and fit the baseline on separate tasks. A gain over a fitted baseline remains useful, but it need not equal the gain over p_H. Its interpretation depends on how well the fitted baseline approximates that conditional probability. We do not estimate (3) here; that would require held-out probabilities and a baseline fitted on separate tasks. Committee boldness labels instead define an annotated diagnostic baseline; judges need not see those labels.

Departures from a heuristic. Restrict to pairs where both the judge's hard choice $\widehat{Y}_J$ and the boldness heuristic h_b are binary. On agreement, their correctness indicators are equal. On disagreement, exactly one is correct. Splitting the accuracy difference over these two events gives

$$\mathrm{Acc}(J) - \mathrm{Acc}(h_b) = \Pr(\widehat{Y}_J \neq h_b) \left[2 \Pr(\widehat{Y}_J = Y \mid \widehat{Y}_J \neq h_b) - 1 \right]. \tag{4}$$

If disagreement has probability zero, the difference is zero. Otherwise, a judge improves accuracy precisely when its departures are correct more than half the time. On these 1,127 clear-gap pairs, no full-description judge beats the heuristic, so departures are correct at most half the time.

Heuristic-silent margin. On the 333 pairs tied under committee boldness, set $b = 0$ and define

$$m_J = \Pr(\widehat{Y}_J = Y \mid b = 0, \text{ full description}) - \tfrac{1}{2}. \tag{5}$$

The observed $+10$ to $+14$ points, with Kimi-K2.6 near zero, measure performance where boldness cannot choose. They do not bound the information in a judge's forecast or its ability on other pairs.

Value of additional information. If the next input retains all earlier information, for example $X_{\ell+1} = (X_\ell, Z)$, the tower property gives $\mathbb{E}[p_{\ell+1} \mid X_\ell] = p_\ell$. Consequently,

$$R_\ell(p_\ell) - R_{\ell+1}(p_{\ell+1}) = \mathbb{E}[(p_{\ell+1} - p_\ell)^2] \geq 0. \tag{6}$$

To verify the equality, expand the square and substitute $\mathbb{E}[p_{\ell+1} p_\ell] = \mathbb{E}[p_\ell^2]$. Additional information cannot harm an optimal predictor. It can harm an actual judge, whose error in estimating the posterior may increase. This comparison requires nested inputs and matched item distributions. Rewriting a description or replacing exploration evidence with identity does not necessarily satisfy the nesting assumption. The theorem alone therefore does not explain changes between the implemented levels.

Pairwise selection value. Let $G = |\sigma(a) - \sigma(a')|$ and let C_J indicate a correct choice. The selected score is $\min\{\sigma(a), \sigma(a')\} + GC_J$, whereas a uniform random choice has conditional mean $\min\{\sigma(a), \sigma(a')\} + G/2$. Subtracting and averaging gives

$$V_J - V_{\text{random}} = \mathbb{E}[G]\big(\text{Acc}(J) - \tfrac{1}{2}\big) + \text{Cov}(G, C_J). \tag{7}$$

The covariance form follows by expanding $\mathbb{E}[GC_J]$. This identity also holds as an average over a fixed bank. Errors on large gaps cost more than errors on small gaps. For more than two candidates, selection value also depends on the selection rule and must be evaluated under the rule actually used.

Conditioning on a failed cue. Suppose boldness is the only available signal and $\Pr(Y = h_b \mid h_b) = 0.779$ in either direction. The calibrated, accuracy-optimal policy follows it on every pair and therefore scores zero on the subset selected precisely because boldness was wrong. Thus override-prior accuracy alone cannot establish faulty Bayesian updating. The same argument applies to weigh-evidence traps, which are selected because exploration evidence favors the losing attempt.

Shared difficulty and shared errors. Excess joint error does not by itself identify a shared prior. For a simple counterexample, let an item's difficulty $\Theta \in [0, 1]$ be random, and suppose M judges fail independently conditional on Θ, each with probability Θ. Then Jensen's inequality gives

$$\Pr(\text{all } M \text{ judges fail}) = \mathbb{E}[\Theta^M] \geq \big(\mathbb{E}[\Theta]\big)^M. \tag{8}$$

For $M > 1$, nonconstant difficulty makes the inequality strict. Joint error above the product of marginal error rates establishes dependence, but does not identify its cause.

What the outcomes establish. Observed outcomes permit direct scoring of frozen forecasts. Without further assumptions or repeated executions, they do not separate posterior estimation error from irreducible uncertainty, nor do they identify a run-noise ceiling. Repeating judge calls measures variation in forecasts, not in the executed outcomes. Here Y records which executed attempt scored higher, not which proposal has higher latent expected quality. Finally, descriptions reconstructed from completed programs support outcome-blind evaluation. Reading them as judgments made before execution also requires that the information they contain could have been available then. Removing explicit results alone does not establish such availability.

G IMPLICATIONS

For deployment, choose a judge using the measures in Section 3 that match the intended use. A research preference model (Foster et al., 2026) should be evaluated on the relevant decisions and task settings, including the outcomes of its selections. Check its calibration and accuracy on pairs tied in boldness before relying on its overall win rate.

Automated research beyond ML, such as AI4Science pipelines and self-improving fleets, faces a systemic version of this risk. Because taste and execution dissociate, cheap judges will likely be paired with strong executors. The shared-error analysis limits the benefit of combining judges; it does not by itself identify their internal priors. Correlated errors across LLM families are well established in general (Kim et al., 2025; Goel et al., 2025). Here, errors may correlate through reliance on the same boldness cue or through shared item difficulty (Eq. (8)). The attempts graded here come from agents, but nothing in the instrument depends on who wrote them. The same protocol can grade judgments about human proposals, given a parallel-attempt corpus of such proposals.

H ADDITIONAL RESULT DETAILS

Information levels and pair accounting. Every measured judge exceeds chance from the 350-character gist alone, with claude-opus-5 reaching 73.5%. Full descriptions add at most five points (Table 2). Exploration evidence helps every grid judge, by up to 12.4 points for gpt-5.6-sol (67.3% to 79.7%). The identity condition supplies system identity and reasoning effort but omits exploration notes. Seven of eight judges score lower than with exploration evidence, including gpt-5.6-sol at 65.8%. Claude-opus-5 reaches 78.8% in both conditions. These differing inputs prevent isolating the effect of revealing identity. The 1,842 AI4AI pairs comprise 1,127 with clear committee-labeled boldness gaps, 333 boldness ties, and 382 in neither split.

Decision tracks. Fifteen leaderboard judges have exploration runs; provider filtering excludes claude-fable-5-1. On misleading-evidence items (158/560), fourteen score 23–45%, versus 80–94% when evidence points correctly. Kimi-K2.6 scores 49% and 66.6%, respectively. This 17.6-point gap compares with 39–70 points for every other judge and is consistent with weaker evidence-following. All sixteen score 36.9–50.0% on override-prior pairs (249/1,127). Rankings vary across tracks (Fig. 9). Claude-opus-4-8 scores 60.6% on select-best but leads reject-below-baseline at 68.4% balanced accuracy. The grid-judge AUCs span 0.58–0.70. Kimi-K2.6 ranks last on select-best among the canonical nine but second on reject-below-baseline among the fifteen complete judges (66.9%). Claude-sonnet-5 leads select-ambitious (73%) but ranks last on weigh-evidence traps (23%).

Shared failures. After excluding nine bank defects, 4.5% of n=1,820 pairs are missed by all eight grid judges, versus 0.033% under independent errors (roughly 135×). Before the audit, the figures were 4.9% and 150×. Joint errors across families reach 1.3–2.0× the independence expectation, a range that includes the within-family value of 1.7×. Correct-answer agreement is also similar across and within families (0.68 versus 0.73). The three OpenAI-family judges score 62.0, 62.6, and 62.7% on AI4AI. Removing nine defective descriptions from the original 90-item taxonomy leaves 81 shared misses: overlooked execution discipline (22), useful priors applied in the wrong regime (18), risk miscalibration (15), misapplied overfitting assumptions (15), surface-text artifacts (6), and cases labeled as having no recoverable signal (5).

Executor-style effects. Most author-bias variance lies with the executor rather than the judge (Fig. 6). All sixteen judges underselect the strongest executor, claude-opus-5, by 4.5–29.8 points relative to its actual win rate and overselect claude-sonnet-5 by 7.0–18.6 points. Own-model effects differ. Kimi-k3 tops its own column at +12.8 points, while gpt-5.6-sol judges its own attempts more harshly than most peers (−9.5). With identities hidden, these associations concern executor-linked descriptions and do not establish model recognition.

Cross-corpus diagnostics. The arXiv-task corpus contains 35 tasks, 398 attempts, and 1,897 pairs. RSI-Exam and paper ablations add 221 and 99 pairs, bringing the bank to 4,059. Judge-rank correlations are 0.62 for AI4AI–arXiv over the canonical nine, and 0.73 for AI4AI–RSI-Exam and

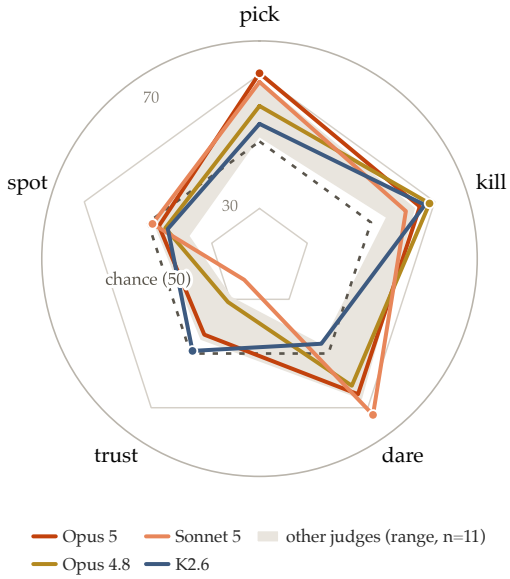


Figure 9: **Judge rankings vary across tracks.** Fifteen judges have complete runs. Fable 5.1 has no weigh-evidence score because provider filtering blocked 770 of 799 exploration prompts. Pick, Kill, Dare, Trust, and Spot denote select-best, reject-below-baseline, select-ambitious, weigh-evidence, and override-prior, respectively. Track leaders are shown in color, and the other eleven judges in gray. The dashed contour marks 50% chance. The radial scale starts at 15%.

0.72 for RSI-Exam–arXiv over the eight canonical judges available in each comparison. Claude-opus-5 and claude-fable-5-1 remain in the top four across agent-attempt corpora, and Kimi-K2.6 in the bottom three. Paper-ablation accuracies span 76.8–94.9%, with cross-corpus correlations of 0.12–0.57. Longer-description accuracy is 51.5% on AI4AI, 59.0% on arXiv tasks, and 74.5% on ablations. The nine-judge majority reaches 67.1% overall, 66.6% on both AI4AI and arXiv tasks, and 70.1% on the balanced subset. Of seven post-grid models evaluated on the full bank, claude-fable-5-1 leads at 70.8%, while qwen3-235b, glm-5, and DeepSeek-V4-Pro rank in the bottom half. ArXiv order consistency spans 0.75–0.88 for the canonical nine and 0.51–0.88 for all sixteen judges, with mistral-large-3 lowest. After rewriting in a common style, self-favoritism spans -0.19 to $+0.16$, with kimi-k3 at $+0.03$. Rewriting reduces self-favoritism for nine of twelve judges with original-form comparisons (Section B and Table 6).

Rationale run. The canonical nine re-judged a balanced 296-pair, four-corpus subset in both orders and gave a two-to-three-sentence rationale before each pick. Accuracy changes average -0.9 points, ranging from -4.4 for claude-sonnet-5 to $+3.0$ for claude-opus-4-8. Mean order consistency is unchanged. Above-median hidden-reasoning length accompanies lower accuracy for kimi-k3 (62.6% versus 77.9% below median) and Kimi-K2.6 (57.8% versus 69.8%; about 37k reasoning characters per pair). Every coded cue family appears more often on errors. Normal-order judgments tagged for ambition-related keywords are correct in 122/215 cases (56.7%). A tag records a mention and does not imply endorsement. These associations do not establish causation.

Human baseline. Twelve researchers judged a 60-pair AI4AI subset in blinded 20-pair slices sharing ten anchors. The automatic exclusion criteria were a median response time below 15 seconds or a total time below five minutes. After excluding two speed-runs and one disavowed first attempt, eleven submissions from ten researchers remain, including two retained retakes that re-exposed anchors. Accuracy averages 63.6% (140/220 judgments), with an individual range of 45–85%, versus 40.0–68.3% for matched grid judges. The human mean exceeds six of eight judges. Five submissions beat at least six judges, and three beat none. At this sample size, accuracy does not differ clearly by seniority. Human and model majorities make identical choices on the ten anchors, including seven correct answers and three shared errors.