How Feedback Shapes Search: Candidate Competition in LLM Agents

Anonymous authors
Paper under double-blind review

Abstract

Large language model (LLM) agents use feedback to guide search, but **better proposals need not yield a better final solution**. We study two LLMs searching fixed sets of image and text classifiers trained before search. With evaluation and selection fixed, showing the correct condition label for each score improves average proposal quality for the requested condition in all four original model–task settings. Paired post hoc comparisons support larger effects on proposal quality than on final quality in two settings, with limited confirmation from new responses. In text classification, the same comparison format encourages one model to revisit tried configurations or explore untried ones depending on score visibility. We construct searches with identically distributed complete proposal-quality sequences but different final outcomes under one selection rule, and characterize when improving a final proposal cannot reduce selected quality. **Evaluating feedback therefore requires tracking candidate competition and improvements preserved through selection.**

1 Introduction

Large language model (LLM) agents use tests, scores and critiques to guide repeated revisions of prompts and programs (Yuksekgonul et al., 2025; Agrawal et al., 2026b). These searches aim to turn additional computation into better solutions. Their success ultimately depends on the quality of the candidate selected as the final answer.

A new proposal can improve on the previous attempt without outscoring the best candidate already found. A proposal that does outscore it may perform worse on unseen examples. Prior work measures feedback gains in searches that retain the best-scoring candidate (Chakraborty et al., 2025) and diagnoses losses from search and final selection (Toledo et al., 2025). **With candidate scores and the selection rule fixed, we ask how feedback changes which proposals challenge existing solutions and whether replacements improve performance on unseen examples.**

We study this question using classifiers trained before search. Using the same classifiers and evaluation data keeps each candidate's score fixed across feedback conditions. An LLM agent chooses a classifier, receives feedback on its performance, and makes its next choice. At the end, we return the highest-scoring classifier tried, using scores even when hidden from the agent. We compare searches starting from the same choices but receiving different performance information. We test proposals and the final choice on separate data that never guide search or selection (Figure 1).

We use two LLMs, DeepSeek-V4-Flash-0731 (DeepSeek) and GPT-5.6-Terra (Terra), on two classification tasks. Both agents search for classifiers that recognize objects in blurred or noisy CIFAR-100 images, or predict individual annotators' emotion labels for comments in GoEmotions (Krizhevsky, 2009; Demszky et al., 2020).

We first compare feedback that correctly identifies which image corruption or annotator each score refers to with feedback that swaps these labels. In the original searches, correct labels improve average proposal test performance on the assigned target for both models on both tasks. Paired post hoc comparisons support larger effects on average proposal quality than on final selected quality in two of the four settings; confirmation from new responses on reused materials is limited (Section 3). Better proposals may still score below an existing winner. We therefore trace which candidates challenge it and whether replacements help or harm.

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.

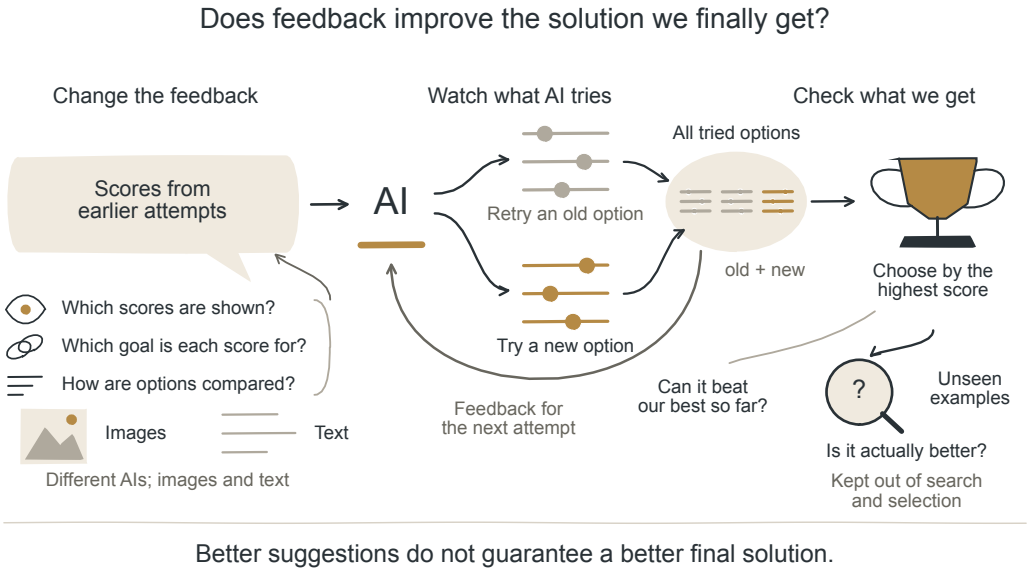


Figure 1: Conceptual overview: from feedback to the final selected solution. We vary score coverage, score labels and comparison format, then track proposals and the final choice among tried options. Held-out data measure proposal and final-choice quality without guiding generation or selection.

In a separate study on the emotion task, we compare scores listed in the order classifiers were tried with a display that sorts disclosed scores, marks the highest disclosed score, and shows each disclosed score's difference from it. **For Terra, this display favors revisiting tried classifiers when some scores are hidden, but trying new classifiers when all tried scores are shown.** The reversal repeats with fresh responses on the same materials; the increase in new choices is less stable with new comments. For DeepSeek, an exploratory comparison with all tried scores shown finds that this display reduces how often proposals outscore the best classifier in the starting history, even though every proposal is new (Section 4). Thus, the frequency of trying new candidates can remain unchanged while their ability to challenge an existing solution changes.

Even a complete record of proposal quality can omit information needed to determine the final result. Our theory constructs searches with identically distributed complete proposal-quality sequences but sharply different returns under one selection rule. The searches differ in which candidates appear together and win the score comparisons, exposing information absent from quality records. With earlier history fixed, we also characterize when raising a last-step proposal's score without lowering its quality cannot reduce final quality (Section 5). Together, these results motivate tracking which alternatives feedback makes competitive and which improvements selection preserves.

2 Design: Separating Feedback, Proposals and Selection

Tasks and one search. The LLM agent chooses among 32 configurations, each specifying a classifier trained before search. Its target is one of two CIFAR-100 image conditions (blur or noise), or the emotion labels supplied by one of two annotators in a GoEmotions block (Krizhevsky, 2009; Demszky et al., 2020). Targets are labeled anonymously as A and B. In the original searches, four evaluated configurations provide the starting examples. The agent then makes four sequential proposals, receiving feedback as specified by the condition (Figure 2). Revisits are allowed and produce no new score measurements; a *new visit* selects a configuration not previously tried.

Changing the feedback. We vary which target labels accompany scores, which scores are visible, and how scores are ordered and compared. **Candidate scores and the selection rule stay fixed across these feedback conditions.** Label interventions begin after the four starting examples. Continuation interventions instead branch from a shared history after the first proposal, before its feedback, leaving

three proposals in the original searches. Subsequent choices and histories can differ across conditions, allowing feedback to change which candidates the agent is likely to propose.

Selecting the final candidate. We select the highest-scoring tried configuration, including the starting examples; ties favor the earliest occurrence. Let F denote the feedback sample and $F_z^{\mathrm{can}}(x)$ the fixed, one-decimal score of configuration x for the requested target z, regardless of its displayed label. From the ordered record R of starting examples and proposals, the selector returns

$$h_F(R) = \operatorname*{arg\,max}_{x \in R}^{\text{earliest}} F_z^{\mathrm{can}}(x). \tag{1}$$

The rule uses scores even when hidden from the agent. Its current winner is the *incumbent*. For label interventions, we compute the selected result by applying this rule offline to the recorded search.

Evaluating proposal and final quality. Separate test samples, denoted U, evaluate each proposal and the final choice. **These test outcomes never guide proposal generation or final selection.** Write $Y(x)$ for a configuration's test quality and $H = Y\{h_F(R)\}$ for final quality; Y uses U unless another sample is specified. Each original F/U sample contains 150 images for CIFAR-100 accuracy or 32 comments for GoEmotions micro-F1 against the target annotator's labels.

Paired comparisons. Each original task uses 32 disjoint evaluation groups (*blocks*), each with two targets and two repeated searches per model, target and feedback condition. We average the four paired quality differences within each block, then compute mean effects and nominal t intervals from the 32 block averages. Intervals treat previously selected policies as fixed. Study-specific designs, analysis status and material reuse are detailed in Appendix A.4.

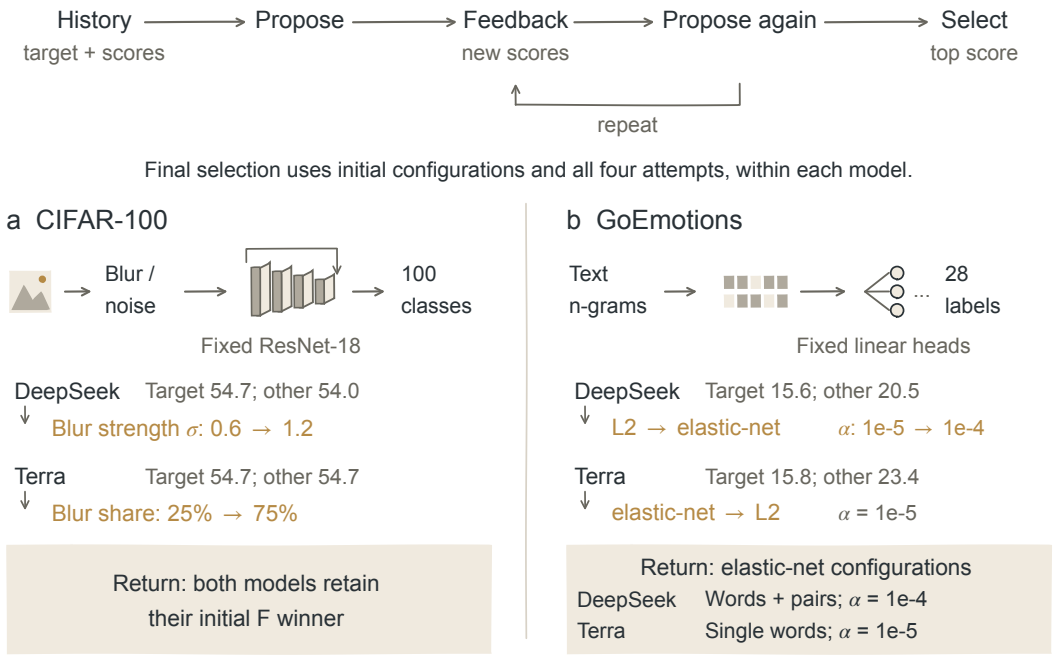


Figure 2: Feedback, configuration changes and retention in recorded searches. Displayed target and other scores are gray; gold text marks configuration changes. CIFAR uses accuracy (%), correct score labels and corruption probability $p = .5$; blur share is conditional on corruption. GoEmotions uses micro-F1 (0–100), partial scores and visit order; α controls regularization. Histories are agent-specific; inputs are paraphrased, and intermediate steps are omitted. F selects among tried configurations with earliest ties (CIFAR: offline replay); hidden tests evaluate returns.

3 SCORE LABELS: PROPOSAL GAINS AND FINAL QUALITY

We ask whether correct score labels improve proposals and the final choice. Each classifier has separate scores for targets A and B. We keep the requested target and score values fixed, and either label the scores correctly or swap their A/B labels. From the same four starting examples, each agent makes four proposals. After search, we select the highest-scoring tried classifier offline, including the starting examples and using the correct target mapping in both conditions. Hidden test data measure two outcomes: the mean quality of the four proposals and the quality of the selected classifier. Table 1 compares the feedback conditions on both outcomes.

Table 1: Effects of correct minus swapped score labels on test quality. The last column subtracts the final-quality effect from the mean-proposal effect. Units are percentage points (pp) of accuracy (CIFAR-100) or micro-F1 (GoEmotions); 128 searches per condition in 32 paired blocks. These post hoc block-t analyses use nominal 95% proposal intervals and simultaneous 95% intervals across the eight final-quality and paired-gap estimates. Test families and corrections are detailed in Appendix B.1.1.

Task	Agent	Mean proposal quality effect nominal 95%	Final selected quality effect simultaneous 95%	Difference between effects simultaneous 95%
CIFAR-100	DeepSeek	$+1.297$ $[0.646, 1.947]$	-0.182 $[-0.873, 0.508]$	$+1.479$ $[0.620, 2.339]$
CIFAR-100	Terra	$+0.857$ $[0.337, 1.377]$	$+0.130$ $[-0.649, 0.910]$	$+0.727$ $[-0.069, 1.522]$
GoEmotions	DeepSeek	$+0.782$ $[0.015, 1.550]$	$+0.116$ $[-0.517, 0.748]$	$+0.667$ $[-0.439, 1.772]$
GoEmotions	Terra	$+0.964$ $[0.252, 1.677]$	-0.169 $[-0.880, 0.542]$	$+1.133$ $[0.010, 2.257]$

Correct labels improve requested-target proposal quality in all four original settings by 0.78–$1.30\,\mathrm{pp}$; all four tests pass Holm correction (Table 1; Appendix B.1.1).

Paired post hoc comparisons support larger effects on proposal quality than on final quality in CIFAR/DeepSeek and, narrowly, GoEmotions/Terra. Final-quality effects remain uncertain in all four settings; the intervals allow small benefits.

Fresh responses on reused materials provide limited confirmation; DeepSeek switches from high to service-default reasoning in this follow-up. Among the four task–agent settings, only DeepSeek on CIFAR supports both a proposal-quality gain and a larger effect on proposal quality than on final quality under the prespecified approximate block-t test. The accompanying bounded analysis leaves all four settings inconclusive (Table 6).

Which proposals can replace the starting winner? To replace the starting winner, a proposal must exceed its fixed canonical F score; we call such proposals *F-qualified*. Correct labels increase their share and reduce the below-threshold group's mean test-quality gap from the starting winner. Changes in this group's frequency and mean quality account for most of the estimated proposal-mean gain in each original setting (Table 5; Appendix B.1.3).

From available to retained improvement. About half the starting histories have no candidate in the fixed set that beats the starting winner in both score and test quality (Table 9). Where such candidates exist, search must find and retain them to improve the final choice. In the exploratory CIFAR/DeepSeek account, correct labels yield larger visited improvements among F-qualified candidates alongside greater loss at final selection. The other three settings leave both changes uncertain (Table 8). **Explaining the effect difference requires tracking which improvements search finds and which it retains.**

4 FROM COMPARISON FEEDBACK TO CANDIDATE COMPETITION

4.1 SCORE COVERAGE CONDITIONS THE RESPONSE TO COMPARISON

We cross comparison format with score coverage on GoEmotions. Partial coverage retains disclosed scores; revisiting can reveal a previously hidden score. Complete coverage supplies requested F for all tried configurations. Visit order lists scores by first occurrence; organized comparison sorts disclosed scores from low to high, places hidden scores last, marks the highest disclosed score, and reports each disclosed score minus it. Coverage also changes prompt length; organization changes order, emphasis and precomputation. Score labels and selection stay fixed (Table 10).

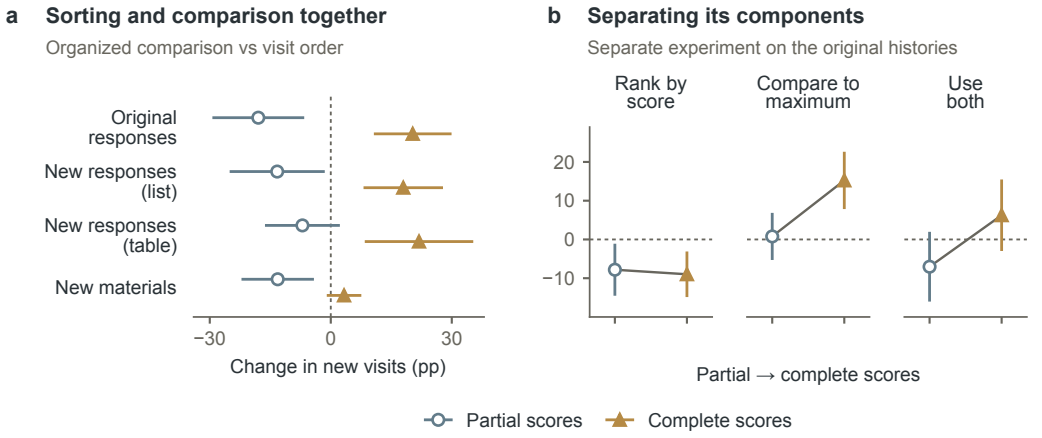


Figure 3: GoEmotions/Terra: effects on first-continuation new visits. (a) Organization effects. Original and fresh-response cohorts: 128 histories in 32 blocks; new materials: 152 histories in 19 blocks. (b) Separate component experiment on 128 histories in 32 blocks. Sorting ranks scores; explicit comparison marks the highest disclosed score, shows differences from it, and explains the comparison. Each component effect averages over the other; their sum gives the effect of applying both. Effects are percentage points, with nominal 95% block-t intervals.

Organization changes direction with coverage. Organized comparison shifts Terra toward revisiting under partial scores and toward new choices under complete scores. In the original exploratory comparison, first-continuation new visits fall by about 18 percentage points under partial coverage and rise by about 20 under complete coverage (Figure 3a). The pattern persists over three proposals and in full-cohort contributions from histories with a higher-F candidate available or with the true starting maximum already displayed. DeepSeek chooses an untried classifier in nearly every first continuation (Appendix B.2).

Does the pattern persist under new responses and materials? Fresh responses on the same materials reproduce both directions of the organized-list reversal. A table display also increases new visits under complete scores, but its weaker partial-score effect leaves the joint criterion for all four list/table directions unmet (Appendix B.2.1).

A prespecified replication with new comments and new pairings of existing raters repeats the shift toward revisiting and meets its joint directional criterion. The increase in new choices is smaller and sensitive to individual blocks (Figure 3a; Appendix B.2.2). The raters and training sources are reused.

4.2 FROM VISIT BEHAVIOR TO CANDIDATE COMPETITION

A continuation F-qualifier is *beneficial* if it improves the starting winner's outcome. A *replacement* beats the current winner's F and can raise or lower its outcome. Retaining the starting winner and F-qualifiers preserves the winner of the ordered search history (Lemma 1). Removing nonqualifiers from feedback could change later proposals.

More visits, competition and uncertain return. In the original searches with complete scores, organized feedback makes Terra try more new configurations and propose more candidates that outscore the starting winner (Figure 4). Replacements also increase (Appendix B.2), while the return effect remains uncertain on U and a second hidden GoEmotions comment panel, W, with 32 new comments per rater.

The same new-visit rate, different competition. Every DeepSeek proposal under complete scores is new, yet organization reduces how often proposals outscore the starting winner (Figure 4). **The new-visit rate does not capture this change in candidate competition.** The two agents thus differ in their competition response; DeepSeek's frequent new visits leave little room to assess an analogous visit reversal.

With new materials, organization again increases Terra's new visits under complete scores across three steps, while the estimated effect on F-qualified share is negative and return remains uncertain (Appendix B.2.2). Across the original coverage and organization comparisons, return effects remain uncertain for both agents on U and W (Table 12).

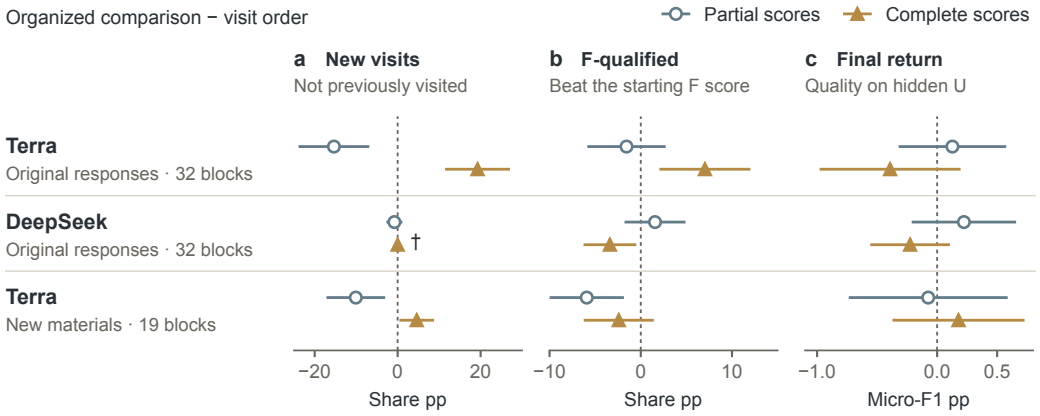


Figure 4: GoEmotions, organized comparison minus visit order: (a) new visits, (b) F-qualified share and (c) U return. Each history contributes three proposals; qualification counts repeats. Original cohorts use 128 histories in 32 blocks; new-material Terra uses 152 histories in 19 blocks. Exploratory paired effects on secondary outcomes, with nominal 95% block-t intervals. † marks zero empirical variance, for which no population interval is reported.

Lower average quality can coexist with better opportunities. In an exploratory DeepSeek comparison, complete scores lower mean proposal quality on U while increasing the best improvement reached among F-qualified candidates. This comparison holds visit order fixed. Proposals below the qualification threshold account for much of the mean decline: their contribution combines how often they appear and how their quality compares with the starting winner. For Terra, the proposal mean instead rises, again largely through below-threshold proposals. The DeepSeek pattern does not repeat on W, and its return effect remains uncertain (Table 13).

A proposal's contribution to return depends on the winner it faces. Let i_t be the winner after x_t. Sums run after the shared prefix through $t = 8$: from $t = 5$ for score-label interventions or $t = 6$ for continuation interventions. Against incumbent i, the outcome change is $k_Y(i, x) = \mathbf{1}\{F(x) > F(i)\}[Y(x) - Y(i)]$. Its positive and negative parts accumulate into gains $P_{\text{repl}} = \sum_t [k_Y(i_{t-1}, x_t)]_+$ and losses $D_{\text{repl}} = \sum_t [-k_Y(i_{t-1}, x_t)]_+$. Let H_{prefix} be the starting winner's outcome and $H_{8,a}$ the final outcome under feedback policy a. Feedback policies a, b induce incumbent–proposal laws $\mu_t^a(i, x)$; Δ denotes $a - b$, giving

$$H_{8,a} - H_{\text{prefix}} = P_{\text{repl},a} - D_{\text{repl},a}, \tag{2}$$

$$\Delta \mathbb{E} H = \sum_{t,i,x} \Delta\mu_t(i, x) k_Y(i, x) = \Delta \mathbb{E} P_{\text{repl}} - \Delta \mathbb{E} D_{\text{repl}}. \tag{3}$$

The cumulative account includes losses recovered by later proposals, which terminal regret omits (Equation 2; Appendix C.2).

4.3 WHICH INFORMATION CHANGES THE RESPONSE?

Sorting and explicit comparison. Which parts of the organized display change the agent's response? A separate Terra experiment crosses coverage with score sorting and explicit comparisons that mark the highest disclosed score, show each disclosed score minus it, and explain the comparison. **The effect of explicit comparisons on first-continuation new visits is larger under complete than partial scores** (Figure 3b).

This coverage-by-comparison interaction passes the prespecified corrected test. The sorting and three-way interactions, the difference between the two component interactions, and the reversal with both components remain unconfirmed in this study. The supported result concerns the three comparison features together (Appendix B.2.3).

Hidden scores and revisiting. Under partial coverage, revisiting could test whether a hidden score exceeds the displayed maximum or reveal its value. Telling the agent that no hidden score exceeds that maximum yields an interaction interval excluding the prespecified 10-point weakening of the organized display's reduction in new visits; smaller effects remain possible (Appendix B.2.4). A separate test finds no support for the prediction that revealing hidden scores on revisits would amplify the organized display's effect on visits to initially hidden-score demonstrations (Appendix B.2.5).

4.4 HOW COMPARISONS AND SAMPLE COUNTS GUIDE CHOICES

From search reuse to controlled transfer. Across the four original settings, continued scores increase reuse of a favored field value in new combinations. Fixed-choice tests ask whether comparisons redirect that preference (Appendix B.3.1).

In two fixed-choice experiments, DeepSeek chooses between hidden-score configurations whose higher-F option either follows or reverses an earlier favored field. A comparison reversing the earlier preference improves reverse-pair accuracy over one preserving it, while same-direction accuracy stays perfect on both tasks. Holding the nearest comparison and global maximum fixed also raises reverse accuracy, but the prespecified joint criterion fails. **The first experiment supports conditional transfer; the stricter test leaves alternative rules unresolved** (Appendices B.3.2–B.3.3).

Sample-count disclosure. With four GoEmotions options, we hold scores fixed while showing or hiding challengers' four- or sixteen-comment counts; the incumbent's 32-comment count stays visible. Four-comment disclosure raises incumbent retention by about 30 percentage points, more than sixteen-comment disclosure. Scores stay fixed within show/hide contrasts but differ across sample sizes. Direct choice bypasses h_F; lower mean U is descriptive, with reliability weighting and incumbent preference unresolved (Appendix B.3.4).

5 WHAT CAN BE INFERRED ABOUT FINAL QUALITY?

5.1 EVEN THE FULL PROPOSAL-QUALITY PROCESS CAN BE INSUFFICIENT

Can complete proposal-quality records alone determine return? **Even the entire proposal-quality process can match while returns differ.** We use abstract F ranks, normalized $Y \in [0, 1]$ and return $H = Y(h_F(R))$.

Theorem 1 (Matched quality, sharply different returns)**.** *For every integer $m \geq 2$, a set of $2m + 1$ candidates admits two search laws A and B with the same prefix, two distinct new proposals, identical candidate marginals at each round and candidate inclusion probabilities, and the same joint law of the entire proposal-Y sequence. Both always visit a candidate with outcome 1, yet*

$$\mathbb{E}H_A = 1, \qquad \mathbb{E}H_B = 1/m.$$

The interval $[1/m, 1]$ is sharp among laws with these matched summaries on this candidate set.

The construction changes which candidates appear together. Place $p, b_1, g_1, \ldots, b_m, g_m$ in increasing F order, with $Y(p) = Y(b_i) = 0$ and $Y(g_i) = 1$. A pairs b_i, g_i; B pairs g_i, b_{i+1} cyclically. Uniform

pair choice and randomized order match every candidate's inclusion and per-round probabilities. Under either law, the ordered outcomes are $(0, 1)$ or $(1, 0)$, each with probability $1/2$.

In A, the outcome-1 candidate always wins. In B, the outcome-0 partner wins except at wraparound, so the return is 1 for only one of the m pairs. Every matching law includes global F winner g_m with probability $1/m$, giving a lower bound of $1/m$ on expected return; mixtures of A and B attain the entire interval up to 1 (Appendix C.3).

The constructions differ in their joint (F,Y) paths: proposal-quality marginals alone do not determine return. Exact matching uses outcome ties. For $0 < \varepsilon < 1/2$, distinct outcomes can give quality sequences coupled within ε at every round, yet a return gap of $(1 - 1/m)(1 - \varepsilon)$. Thus, near-matching proposal quality can coexist with sharply different returns (Appendix C.3). In the observed GoEmotions paths, exact-quality matching leaves return ambiguous in only 1–5 of 128 U histories per agent and condition (Appendix B.5.3).

5.2 When a better proposal preserves return

When does improving a final proposal preserve return? Raising both F and outcome can still displace a better incumbent. Here F abbreviates canonical requested F, and $\mathcal{Z}$ is the set of possible earlier winners.

Proposition 1 (When a proposal improvement transfers)**.** *Replace a final proposal x by x', with $F(x') > F(x)$ and $Y(x') \geq Y(x)$. Return is nondecreasing for every earlier winner $z \in \mathcal{Z}$ if and only if*

$$F(x) \leq F(z) < F(x') \implies Y(z) \leq Y(x') \qquad (z \in \mathcal{Z}).$$

Fix the earlier winner z. If its score is below $F(x)$, both proposals win and outcome cannot fall; at or above $F(x')$, z wins both comparisons. Only $F(x) \leq F(z) < F(x')$ switches return from z to x', making $Y(x') - Y(z) \geq 0$ necessary and sufficient there. Earliest ties fix the interval endpoints.

In a GoEmotions/DeepSeek example, changing one field improves $(F, U) = (26.9, 22.73)$ to $(31.2, 23.88)$, crossing prefix winner $(29.6, 28.89)$. Ending the offline evaluation after this first continuation gives a 5.01 pp return loss: the improved proposal displaces a better incumbent. This is an offline, outcome-informed substitution using fixed candidate scores and test outcomes (Appendix B.5.5).

F-order outcome inversions bound return's shortfall from the best outcome among the starting winner and visited qualifiers. On support $\mathcal{S}$ containing the prefix winner and higher-F candidates, define $D(\mathcal{S}) = \max_{x,y \in \mathcal{S}: F(x) \leq F(y)} [Y(x) - Y(y)]_+$. The best-outcome candidate among visited qualifiers and the starting winner cannot outrank the F winner, so its outcome advantage is at most D. Two proposal slots can attain D: propose a maximizing pair's lower-Y member before its partner; F and earliest ties retain the lower-Y member. For $D = 0$, repeat the prefix winner. Using complete qualifying supports gives negative mean return-effect lower bounds in all evaluated contrasts, so no positive gain is certified (Appendix B.5.4).

6 Does more evaluation information improve final quality?

Two new GoEmotions panels use the same raters. V has 24 comments per rater and is used either to reselect among tried candidates or to update feedback before further proposals; hidden W has 32 comments per rater and evaluates both studies.

More information for selection. For the candidates tried with continued or withheld scores, V informs only selection; we pool F+V counts before computing micro-F1. This raises observed mean W return in all four agent–feedback conditions. Its effect on the advantage of continued scores remains uncertain for both agents (Table 24).

Updating feedback with expanded scores. A separate experiment supplies subsequent feedback using either original F scores or expanded scores from 32 F plus 24 V comments; the original F selector returns the final candidate. Its generation comparison and evaluation budget differ from the selection-only study. Earlier F history and fixed-score instructions remain in place, although the

updated scores can differ; the messages name no new panel or objective. On the previously inspected W panel, the primary outcome measures the increase in best visited test quality over the best test quality in the starting history; final selected quality is secondary (Appendix B.4.2).

Expanded feedback's local Terra return gain is chiefly accounted for by lower cumulative replacement losses. In the original responses, the secondary W-return gain is about $0.6\,\mathrm{pp}$. The primary increase in best visited W quality is unconfirmed for both agents (Appendix B.4.2). Fresh responses on the same materials leave Terra's return gain unconfirmed, and DeepSeek's return effects remain uncertain (Table 27).

What did further search add? We replay both selectors on this separate feedback cohort. **A higher final return can come from reselecting a solution already available.** Let $H^p_{g,s}$ be final quality on outcome panel p under feedback g and selector s, B^p_s the selector's return on the starting history, and $G^p_{g,s} = H^p_{g,s} - B^p_s$ the gain from continuation. For both g and s, F uses original scores and FV uses pooled F+V scores. Comparing expanded versus original feedback and selection gives

$$H^p_{FV,FV} - H^p_{F,F} = (B^p_{FV} - B^p_F) + G^p_{FV,FV} - G^p_{F,F}. \tag{4}$$

Only the last two terms concern further search: the first precedes any new proposal. Each generation–selector combination's continuation interval allows both signs (Table 26).

7 Related work

Feedback as search information. Feedback guides self-refinement and agent memory (Madaan et al., 2023; Shinn et al., 2023), score-conditioned prompt and program search (Yang et al., 2024; Romera-Paredes et al., 2024), and text optimization (Yuksekgonul et al., 2025; Agrawal et al., 2026b; Lee et al., 2026); code repair provides another setting (Olausson et al., 2024). Prior comparisons vary feedback form and fidelity under best-so-far selection (Chakraborty et al., 2025), feedback sensitivity (Gupta et al., 2025), frozen-state components (Chong et al., 2026), diagnostic detail (Agrawal et al., 2026a) and success versus failure feedback (Liu et al., 2026). Our GoEmotions/Terra contrast holds candidate scores and selection fixed while score coverage reverses the new-visit response to one display.

From availability to return. Prior work selects initial states for execution (Heuillet & Peddiraju, 2026), diagnoses search and final-selection losses (Toledo et al., 2025), and isolates answer recognition with fixed-candidate replay (Ji et al., 2026). We trace challengers and replacements; Theorem 1 shows matched proposal-quality sequences can yield different returns under one selector.

Adaptive evaluation. Proxy-score optimization can reduce true reward (Gao et al., 2023; Huang et al., 2025); adaptive-evaluation work addresses feedback reuse (Dwork et al., 2015; Blum & Hardt, 2015; Russo & Zou, 2020; Bertran et al., 2026). We ask when score–outcome disagreements affect return through the candidates that compete.

8 Discussion and conclusion

Correct score labels improve proposal quality in four original settings; post hoc comparisons support larger effects on proposal quality than on final quality in two. Comparison effects depend on coverage; competition can change even when new visits do not. Expanded feedback's local Terra gain is chiefly accounted for by lower cumulative replacement losses and remains unconfirmed by fresh responses.

Implications for evaluating feedback. Evaluate final quality with paired return effects alongside proposal gains. Record feedback, proposals, F, winner updates and hidden outcomes to identify challengers and replacement gains/losses (Table 30). Use selector-specific eligibility and starting returns to distinguish reselection from further search (Equation 4).

Scope and future work. These controlled comparisons use fixed candidate sets; effects in open-ended generation remain to be established. Fresh responses provide limited confirmation of proposal–return gaps in the score-label study. Future work can tailor feedback to changing incumbents and evaluation uncertainty, balancing exploration, comparison and retention.

AI USE STATEMENT

We used generative AI tools, including ChatGPT and Codex-based coding agents, to help refine hypotheses and experimental designs, write and refactor code, and run and monitor experiments. These tools also helped us develop audit checks, analyze results, prepare figures, search the literature, and draft, translate, and edit the manuscript.

We defined the research questions and made the final decisions about methods and interpretation. We approved and supervised each experiment, inspected the raw outputs, and checked the code, references, figures, and claims in the paper. All reported measurements came from the deterministic evaluators and intervention pipelines described in the paper, not from language-model judgments. We reviewed all AI-assisted work and take responsibility for the final paper.

ETHICS STATEMENT

The study uses existing public benchmark data and involves no newly recruited human participants. Raw comments, item-level annotations, images, target identities, and held-out outcomes were excluded from agent inputs. External model APIs received candidate configurations, remaining budget, requested anonymous target, and prescribed feedback.

REPRODUCIBILITY STATEMENT

Appendix A specifies task construction, classifier fitting, search and inference; Appendix B gives the experimental results and study-specific settings; Appendix C gives derivations and proofs. Appendix D states the agent settings and common feedback messages.

REFERENCES

Lakshya A. Agrawal, Donghyun Lee, Shangyin Tan, Wenjie Ma, Karim Elmaaroufi, Rohit Sandadi, Sanjit A. Seshia, Koushik Sen, Dan Klein, Ion Stoica, Joseph E. Gonzalez, Omar Khattab, Alexandros G. Dimakis, and Matei Zaharia. optimize_anything: Unified text optimization can outperform specialized systems. In *Proceedings of the ACM Conference on AI and Agentic Systems*, pp. 1–16. ACM, 2026a. doi: 10.1145/3786335.3813167.

Lakshya A. Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziems, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J. Ryan, Meng Jiang, Christopher Potts, Koushik Sen, Alex Dimakis, Ion Stoica, Dan Klein, Matei Zaharia, and Omar Khattab. GEPA: Reflective prompt evolution can outperform reinforcement learning. In *The Fourteenth International Conference on Learning Representations*, 2026b.

Martin Andres Bertran, Aaron Roth, and Zhiwei Steven Wu. What fits (into few tokens) doesn't overfit: Compression and generalization in ML research agents. arXiv preprint arXiv:2606.11045v1, 2026.

Avrim Blum and Moritz Hardt. The ladder: A reliable leaderboard for machine learning competitions. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1006–1014. PMLR, 2015.

Souradip Chakraborty, Mohammadreza Pourreza, Ruoxi Sun, Yiwen Song, Nino Scherrer, Furong Huang, Amrit Singh Bedi, Ahmad Beirami, Jindong Gu, Hamid Palangi, and Tomas Pfister. On the role of feedback in test-time scaling of agentic AI workflows. arXiv preprint arXiv:2504.01931v4, 2025.

Yee Hin Chong, Jiaming Wu, Youhui Zhang, and Peng Qu. Towards feedback-to-plan decisions for self-evolving LLM agents in CUDA kernel generation. In *International Conference on Machine Learning*, 2026.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4040–4054. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.372.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. Preserving statistical validity in adaptive data analysis. In *Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing*, STOC '15, pp. 117–126. Association for Computing Machinery, 2015. doi: 10.1145/2746539.2746580.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10835–10866, 2023.

Rushil Gupta, Jason Hartford, and Bang Liu. LLMs for Bayesian optimization in scientific domains: Are we there yet? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 15482–15510, 2025. doi: 10.18653/v1/2025.findings-emnlp.838.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Maxime Heuillet and Sharadind Peddiraju. CIPHER: A decoupled exploration-selection framework for test-time scaling of data science agents. arXiv preprint arXiv:2607.14386v1, 2026.

Audrey Huang, Adam Block, Qinghua Liu, Nan Jiang, Akshay Krishnamurthy, and Dylan J. Foster. Is best-of-N the best of them? coverage, scaling, and optimality in inference-time alignment. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 25075–25126. PMLR, 2025.

Jia-Hao Ji, Sijie Li, Jiabei Cheng, Zixi She, Jin-Tai Yu, and Zhiyuan Yuan. Candidate supply and answer selection shape the value of LLM judging in multi-agent systems. arXiv preprint arXiv:2608.25937v2, 2026.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Yoonho Lee, Joseph Boen, and Chelsea Finn. Feedback Descent: Open-ended text optimization via pairwise comparison. In *ICLR Workshop on Memory for LLM-Based Agentic Systems (MemAgents)*, 2026.

Yuxuan Liu, Zhaochen Su, Yuhao Zhang, Jiahe Guo, Zhongwei Xie, Huihao Jing, Lingyun Xie, Qing Zong, Yauwai Yim, Zhixiong Zhang, Haoran Li, and Yangqiu Song. Rethinking self-evolving agent skills: Feedback dynamics over multiple rounds. arXiv preprint arXiv:2608.02636v1, 2026.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-Refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-2019.

Theo X. Olausson, Jeevana Priya Inala, Chenglong Wang, Jianfeng Gao, and Armando Solar-Lezama. Is self-repair a silver bullet for code generation? In *The Twelfth International Conference on Learning Representations*, 2024.

Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M. Pawan Kumar, Emilien Dupont, Francisco J. R. Ruiz, Jordan S. Ellenberg, Pengming Wang, Omar Fawzi, Pushmeet Kohli, and Alhussein Fawzi. Mathematical discoveries from program search with large language models. *Nature*, 625:468–475, 2024. doi: 10.1038/s41586-023-06924-6.

Daniel Russo and James Zou. How much does your data exploration overfit? controlling bias via information usage. *IEEE Transactions on Information Theory*, 66(1):302–323, 2020. doi: 10.1109/TIT.2019.2945779.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-0377.

Edan Toledo, Karen Hambardzumyan, Martin Josifoski, Rishi Hazra, Nicolas Baldwin, Alexis Audran-Reiss, Michael Kuchnik, Despoina Magka, Minqi Jiang, Alisia Lupidi, Andrei Lupu, Roberta Raileanu, Tatiana Shavrina, Kelvin Niu, Jean-Christophe Gagnon-Audet, Michael Shvartsman, Shagun Sodhani, Alexander Miller, Abhishek Charnalia, Derek Dunfield, Carole-Jean Wu, Pontus Lars Erik Saito Stenetorp, Nicola Cancedda, Jakob Foerster, and Yoram Bachrach. AI research agents for machine learning: Search, exploration, and generalization in MLE-bench. In *Advances in Neural Information Processing Systems*, volume 38, pp. 39717–39756, 2025. doi: 10.52202/085713-1186.

Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1):1–27, 2024. doi: 10.1093/jrsssb/qkad009.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, and Xinyun Chen. Large language models as optimizers. In *The Twelfth International Conference on Learning Representations*, 2024.

Mert Yuksekgonul, Federico Bianchi, Joseph Boen, Sheng Liu, Pan Lu, Zhi Huang, Carlos Guestrin, and James Zou. Optimizing generative AI by backpropagating language model feedback. *Nature*, 639(8055):609–616, 2025. doi: 10.1038/s41586-025-08661-4.

GUIDE TO THE APPENDICES

The map below follows the paper's questions. Shared methods are in Appendix A; agent settings and common feedback messages are in Appendix D.

Question or material	Appendix
Score labels, proposal quality and final selection	B.1
Score coverage, comparison format and follow-up interventions	B.2
Continued feedback, transfer and sample-count choices	B.3
Additional information at generation and selection	B.4
Replacement accounts, measurement checks and recorded examples	B.5
Mathematical derivations and complete proofs	C
Agent settings and feedback-message formats	D

A EXPERIMENTAL SETUP AND INFERENCE

This section expands the design in Section 2: task construction and evaluation panels, the fixed candidate sets, the common search protocol, and the units used for inference. Study-specific changes are described with their results in Appendix B.

A.1 TASKS AND EVALUATION PANELS

Evaluation panels and information boundaries. F supplies feedback and the canonical requested-target selection score; U is hidden from generation and selection. Each F/U panel has 150 images on CIFAR-100 or 32 comments on GoEmotions. GoEmotions additionally allocates 24 V comments per rater for feedback or selection and 32 hidden W comments for evaluation. U comments are pair-shared; V/W comments are rater-specific, exclude earlier allocations and reuse the same raters. Appendix B.4.1 describes the allocation. All task outcomes retain their own scale: accuracy or per-rater micro-F1. Offline analyses use hidden test outcomes without exposing them to the agent or selector.

A.1.1 CIFAR-100 CONSTRUCTION

The task construction was specified before task-specific outcomes. A shared training pool contains 470 images per fine class (47,000 total); eight calibration blocks use 2,400 additional training-split images. The study uses 9,600 test-split base images in 32 sample-disjoint blocks, with paired blur/noise views and disjoint feedback and outcome panels. Each 150-image panel includes every fine class once and a rotating 50 classes twice. These allocations define held-out performance under the specified sampling and corruption protocol.

Five binary configuration fields give 32 ResNet-18 configurations (He et al., 2016) (Appendix A.2). The agent sees pooled scores for four initial configurations and, before its first proposal, the target and scores with condition-specific labels for the last demonstration. Later proposals receive their predecessor's scores. Four proposals in eight trajectories spanning target, score-label condition and draw per block give 1,024 proposal slots per agent and task. Separate feasibility calibration is excluded; agent settings are given in Appendix D.2.

A.1.2 GOEMOTIONS CONSTRUCTION

The task uses all 28 labels (27 emotions plus neutral), a fixed decision threshold of 0.5, and one global micro-F1 over each panel's comment–label decisions. Each block contains a distinct pair of raters who labeled the same comments, 32 shared feedback comments, and 32 separate shared outcome comments. Training, feedback, and outcome comments are disjoint. A pooled score averages the two rater-specific F1 values arithmetically.

Training uses 20,000 additional comments with consensus labels from at least two clear annotations. The 32 configurations cross n-grams, regularization and class weighting; four demonstration configurations rotate by block (Appendix A.2). The numbers of blocks, targets, draws and proposals match

CIFAR-100; all proposals select classifiers trained before search. Agents receive configurations and anonymous scores, without raw comments, labels, rater identities or outcomes.

GoEmotions block allocation. Raw annotations are grouped by comment ID, excluding `example_very_unclear=True` and previously allocated evaluation comments. Eligible pairs share at least 64 clear-rated comments; maximum-cardinality matching with seeded weights selects disjoint raters, retaining the first 32 pairs in seeded order. Maximum flow assigns 64 globally distinct comments per pair, split in seeded order into 32 F and 32 U comments; seeded orientation maps raters to A/B. Each of the 64 evaluation raters appears in one block. From remaining comments with at least one label supported by two clear raters, seeded order supplies 20,000 training and 5,000 development comments, excluding text identical to evaluation comments. Development remains separate from fitting and F/U evaluation.

A.2 Fixed candidate sets

Configuration fields and training. Table 2 lists the allowed values. IDs enumerate the Cartesian product, last field fastest: $1 + 16i_1 + 8i_2 + 4i_3 + 2i_4 + i_5$ for CIFAR, $1 + 8i_1 + 4i_2 + 2i_3 + i_4$ for GoEmotions. Indices are zero-based; GoEmotions's first has four levels, all others two. CIFAR demonstration order is fixed; GoEmotions rotates its four demonstration IDs left by $(b-1) \bmod 4$ in block b.

Table 2: Allowed configuration values and initial configuration IDs. Each row's levels are ordered as printed; JSON `null` means no class weighting.

Task	Field	Ordered levels
CIFAR-100	`blur_share`	0.25, 0.75
	`corruption_probability`	0.25, 0.5
	`blur_sigma`	0.6, 1.2
	`noise_sigma`	0.04, 0.1
	`label_smoothing`	0.0, 0.1
	Demonstration IDs	C01, C16, C22, C27
GoEmotions	`analyzer_ngrams`	`word:1-1`, `word:1-2`, `char:3-5`, `char:4-6`
	`alpha`	10^{-5}, 10^{-4}
	`penalty`	`l2`, `elasticnet`
	`class_weight`	`null`, `balanced`
	Demonstration IDs	C01, C12, C22, C31

CIFAR-100 uses torchvision ResNet-18 initialized without pretrained weights, with a 3×3 stride-1 input convolution and no initial max-pooling. Training uses 12 epochs, batches of 256, Nesterov stochastic gradient descent (SGD; learning rate 0.1, momentum 0.9, weight decay 5×10^{-4}), and an epoch-wise cosine schedule. Random crop to 32×32 with padding 4 and horizontal flip precede corruption. A corruption is applied with probability `corruption_probability`; conditional on corruption, it is 5×5 Gaussian blur with probability `blur_share` and additive Gaussian noise otherwise. Noise is clipped to $[0, 1]$. Inputs are normalized by mean $(0.5071, 0.4867, 0.4408)$ and standard deviation $(0.2675, 0.2565, 0.2761)$. Cross-entropy uses the configuration's label smoothing; CUDA fitting uses bfloat16 autocasting. Evaluation uses blur kernel 5 and $\sigma = 1.0$, or sample-seeded noise with $\sigma = 0.08$, followed by the same normalization and top-1 accuracy. The allocation seed is 20260831; training seeds vary by block and are shared across that block's 32 configurations.

GoEmotions uses `HashingVectorizer` with 2^{18} features, nonnegative hashing (`alternate_sign=False`), float32 values, and its default text preprocessing and normalization. The `analyzer_ngrams` field supplies the analyzer and inclusive n-gram range. Each of 28 one-versus-rest classifiers is an `SGDClassifier` with logistic loss, the listed regularization and weighting, 10 iterations, `tol=None`, and elastic-net ratio 0.15 when applicable; other estimator arguments use library defaults. A label is predicted when its probability is at least 0.5. Micro-F1 uses $2\mathrm{TP}/(2\mathrm{TP} + \mathrm{FP} + \mathrm{FN})$ over all comment–label decisions, with TP, FP and FN denoting true positives, false positives and false negatives; a zero denominator returns zero. The allocation seed is 20260904; all 32 configurations in a block share its training seed, and both agents use the same trained classifiers.

Original-material interventions reuse the same fitted classifiers within each block; the new-material replication in Appendix B.2.2 instead fits 608 candidates with new block training seeds.

A.3 Search protocol and selection

The original search agents are DeepSeek-V4-Flash-0731 (DeepSeek) and GPT-5.6-Terra (Terra). Appendix D.2 gives the generation settings.

Each original task has 32 sample-disjoint evaluation blocks, two anonymous targets per block and two draws per target. Four demonstrations precede four agent proposals. Score-label interventions start after the demonstrations; continuation experiments instead fork the saved history after the initial proposal, before its feedback. The new-material replication uses 19 blocks and four draws per target (Appendix B.2.2).

Messages and proposal tool. System prompts name the task model, anonymous A/B targets and displayed-history objective, requiring one `propose_candidate` call per round. Before the first proposal, the agent receives four demonstration configurations and pooled scores, the target, the last demonstration's A/B scores, and the remaining budget. Later rounds repeat the latest-candidate body. Scores have one decimal and accuracy/micro-F1 headings; configurations use compact, alphabetized JSON. The score-label intervention swaps score labels; the withheld-score control removes new scores. Continuation branches also receive the selector and fixed-score instructions below. Later interventions specify their changes separately.

Sessions retained user messages, assistant calls and tool replies without compaction. The sole tool required every configuration field and forbade extra properties. CIFAR fields and GoEmotions `alpha` were numeric; other GoEmotions fields were strings except nullable `class_weight`. A successful call ended the round with "Candidate recorded." Repeats were legal. Common message templates appear in Appendix D.1.

Continuation messages state that all demonstration and proposed candidates remain eligible, the last proposal is not a final submission, and the selector uses canonical one-decimal requested F with earliest ties even when scores are hidden from the agent. Scores are fixed; repeats produce no new measurement. Shared-history forks preserve the initial assistant response and score-free tool reply before arm-specific feedback (Appendix D.1).

Each continuation round records one valid proposal. Tool-argument validation errors are returned to the agent, allowing up to three corrections within that round; failed submissions add no candidate or evaluation. Fresh score-label searches instead retain invalid responses as assigned slots (Appendix B.1.2).

Execution order. Score-label trajectories balance target, arm and draw through cyclic order: the eight identities rotate one position per block, each occupying each position four times over 32 blocks. The four rounds within a trajectory are sequential. Analysis pairs the complete matched trajectories within blocks.

A.4 Estimation and uncertainty

The inferential target depends on what is held fixed. Block-t intervals address variation across blocks under a working population model; bounded-mean intervals address generated responses on fixed contexts. The score-label proposal Holm summary and requested-F return/gap replays are post hoc, with separate four- and eight-test families (Appendix B.1.1). Original visit and path diagnostics are exploratory. Component and controlled-choice primary contrasts precede collection; directional replication and query predictions use intersection–union tests. Criteria and failures are reported per study. There is no paper-wide correction, and reused materials create dependence across studies. Unless otherwise stated, paired block-t confidence intervals (CIs) are nominal.

Block-level inference. Average targets and draws within each block. For 32 contrasts $\widehat{\theta}_b$, report $\overline{\widehat{\theta}} \pm t_{31,0.975} s_{\widehat{\theta}}/\sqrt{32}$, with sample SD s. The working block-population model conditions on training data, task construction, fixed policies and agent configuration; its variance combines material and

generation variation. On the fixed cohort, the estimand instead averages expected generation responses over the saved histories and panels: this target calls for uncertainty over repeated generation on the same contexts. Rounds, draws and U/W remain nested. Earlier policy and diagnostic selection remains outside the uncertainty calculation, including for fresh responses collected on reused materials. The score-label replay and post hoc intervals are exploratory.

Inference on fixed contexts. The controlled-choice studies also use bounded-mean inference for the expected response averaged over their fixed contexts. Each study specifies its unit count, contrast bounds and multiplicity family. Independent generation across those units and a stable agent configuration are the working assumptions. Appendix C.6 gives the construction, endpoint conventions and interval inversion. Zero empirical variance is reported descriptively.

B Complete experimental results

The studies follow the questions in Sections 3, 4 and 6: score labels, coverage and comparison format, controlled choices, and additional evaluation information. Each group gives its comparison and reported outcomes; cross-study replacement and measurement diagnostics follow in Appendix B.5. Common agent and message settings appear in Appendix D.

B.1 Score labels and proposal–return effects

This study supports Section 3. It compares correct and swapped score labels on the same four task–agent settings, then follows the changes from proposal quality to the selected return.

B.1.1 Original cohort

The proposal and return comparison in Table 1 uses test quality $Y(x)$ on the requested target, in percentage points. Requested-F selection (h_F) was added as an offline replay after inspecting the earlier analysis of test performance on the requested target minus performance on the other target. It uses canonical, one-decimal scores for all prefix and suffix occurrences and earliest ties (Equation 1). This post hoc replay measures the outcome of applying a fixed requested-F selector to the recorded candidates.

Table 3: Original score-label conditions: mean proposal quality P and final requested-F return H on hidden U. Units are accuracy pp (CIFAR-100) or micro-F1 pp (GoEmotions). Each condition retains 128 searches in 32 blocks. The levels provide context for the direct paired effects and intervals in Table 1.

		P: mean proposal		H: final return	
Task	Agent	Swapped	Correct	Swapped	Correct
CIFAR-100	DeepSeek	57.630	58.927	60.453	60.271
CIFAR-100	Terra	58.289	59.146	60.474	60.604
GoEmotions	DeepSeek	30.663	31.445	31.798	31.913
GoEmotions	Terra	29.990	30.955	31.749	31.580

Table 4: Other-target and pooled proposal responses to correct minus swapped score labels. Task-specific pp; nominal 95% paired intervals over 32 blocks. Requested-target proposal and return effects appear in Table 1.

	CIFAR-100		GoEmotions	
Quantity	DeepSeek	Terra	DeepSeek	Terra
Non-requested response	-1.221 $[-1.701, -0.742]$	-0.967 $[-1.620, -0.315]$	-0.694 $[-1.384, -0.004]$	-0.604 $[-1.367, 0.159]$
Pooled response	0.038 $[-0.207, 0.282]$	-0.055 $[-0.393, 0.282]$	0.044 $[-0.371, 0.459]$	0.180 $[-0.219, 0.579]$

Table 5: Score-label effects (correct minus swapped; 128 trajectories per arm, 32 blocks). For starting winner i_0, split $\frac{1}{4}\sum_{t=5}^{8}[U(x_t) - U(i_0)]$ at $F(x_t) > F(i_0)$, zero outside each group. All proposals and starting histories remain in the denominator. Contributions sum to the proposal effect, combining frequency and quality; gain minus loss recovers return. Units: accuracy pp (CIFAR-100), micro-F1 pp (GoEmotions). Exploratory nominal 95% paired block-t intervals, unlike Table 1's simultaneous intervals.

Task / agent	Unqualified contribution	Qualified contribution	Replacement gain	Replacement loss	U return
CIFAR-100	$+1.306$	-0.009	$+0.375$	$+0.557$	-0.182
DeepSeek	$[0.713, 1.899]$	$[-0.223, 0.204]$	$[0.166, 0.584]$	$[0.154, 0.961]$	$[-0.662, 0.298]$
CIFAR-100	$+0.750$	$+0.107$	$+0.177$	$+0.047$	$+0.130$
Terra	$[0.296, 1.204]$	$[-0.210, 0.423]$	$[-0.377, 0.731]$	$[-0.178, 0.271]$	$[-0.412, 0.672]$
GoEmotions	$+0.644$	$+0.138$	$+0.118$	$+0.003$	$+0.116$
DeepSeek	$[-0.082, 1.370]$	$[-0.106, 0.382]$	$[-0.169, 0.406]$	$[-0.319, 0.325]$	$[-0.324, 0.555]$
GoEmotions	$+1.035$	-0.071	$+0.072$	$+0.241$	-0.169
Terra	$[0.350, 1.720]$	$[-0.331, 0.189]$	$[-0.206, 0.349]$	$[-0.172, 0.654]$	$[-0.663, 0.325]$

Paired proposal–return effects. For block b, ΔP_b is the correct-minus-swapped effect on mean requested U over four proposals, ΔH_b the effect on final requested-F return, and $G_b = \Delta P_b - \Delta H_b$. Computing G within each block retains the covariance of the proposal and return effects. The four original proposal-effect tests form a separate two-sided Holm family, with adjusted p values $.00122, .00622, .04604, .01923$ in Table 1 order. The post hoc family of four G and four H tests uses two-sided Holm correction and t_{31} intervals at $\alpha = .05/8$. In the same order, the Holm-adjusted G p values are $.0001, .0702, .4334, .0410$; all four adjusted H p values equal 1. The original requested-F replay followed the between-target analysis described above.

The simultaneous H upper endpoints $.508, .910, .748, .542$ pp exclude one-point gains on these task-specific scales; no equivalence margin was prespecified. G compares two intervention effects. Selection loss instead compares the chosen candidate with the best outcome in a specified visited set. For example, even with F=U, proposals improving from .4 to .6 cannot displace a .9 prefix winner. Replaying the 2,048 original score-label and coverage–format paths with unrounded F changes no final winner, so the score-label P/H/G estimates and tests stay unchanged.

B.1.2 Fresh responses on the same materials

A prespecified follow-up retains the 32 original blocks and two targets per task, with one new paired draw per context and four slots per arm: 2,048 slots across Terra and DeepSeek. DeepSeek uses service-default reasoning with an 8,192-token output cap; Terra retains high reasoning (Appendix D.2). Proposal utility P averages all four assigned slots; invalid responses score zero and add no selectable candidate. H is computed with the common requested-F, earliest-tie selector on demonstrations and valid proposals. All starting histories remain, including those with no higher-F candidate available.

Table 6 reports the complete four-setting results. Two target contrasts are averaged within each block for approximate t_{31} inference. The four prespecified conjunctions $\Delta P > 0$ and $G > 0$ use the larger one-sided component p value, followed by Holm correction. The 12 P/H/G intervals use Bonferroni adjustment. The bounded sensitivity uses the same 12 parameters and 64 context contrasts, with one common logical range per task and metric (Appendix C.6; inversion threshold $12/.05 = 240$). It assumes independent generation across contexts conditional on materials and stable agent configuration. One draw per context does not identify within-context generation variance; block-t uncertainty includes block heterogeneity.

Only DeepSeek on CIFAR meets the approximate conjunction; none meets the bounded criterion. Its joint $p = .01605$ and G interval spanning zero refer to different families: four one-sided conjunctions versus 12 simultaneous two-sided intervals. Support is limited, depends on the inferential method, and comes from reused materials.

Table 6: Fresh-response score-label effects: correct minus swapped score labels. Each setting has 64 original block–target contexts, one fresh paired draw per context, and 256 slots per arm. Units are accuracy pp (CIFAR-100) or micro-F1 pp (GoEmotions). $G = \Delta P - \Delta H$. Approximate and bounded columns each report the prespecified 12-parameter simultaneous 95% intervals. Joint p tests $\Delta P > 0$ and $G > 0$ with one-sided block-t components and Holm correction across four settings. All four bounded joint Holm values are 1.

Setting	Effect	Estimate	Approximate CI	Bounded CI	Joint p_{Holm}
CIFAR-100 / DeepSeek	ΔP	$+1.544$	$[0.268, 2.821]$	$[-7.736, 10.446]$	0.01605
	ΔH	$+0.292$	$[-0.515, 1.099]$	$[-1.688, 2.294]$	
	G	$+1.253$	$[-0.115, 2.620]$	$[-8.418, 10.615]$	
CIFAR-100 / Terra	ΔP	$+0.794$	$[-0.109, 1.698]$	$[-8.366, 9.757]$	0.09974
	ΔH	$+0.156$	$[-0.782, 1.095]$	$[-1.877, 2.245]$	
	G	$+0.638$	$[-0.400, 1.676]$	$[-8.922, 10.039]$	
GoEmotions / DeepSeek	ΔP	$+0.972$	$[-0.582, 2.526]$	$[-5.892, 7.606]$	0.09974
	ΔH	$+0.254$	$[-0.395, 0.902]$	$[-2.668, 3.114]$	
	G	$+0.718$	$[-0.778, 2.215]$	$[-7.557, 8.823]$	
GoEmotions / Terra	ΔP	$+0.741$	$[-0.493, 1.975]$	$[-6.050, 7.350]$	0.09974
	ΔH	-0.125	$[-0.743, 0.492]$	$[-2.981, 2.758]$	
	G	$+0.866$	$[-0.410, 2.142]$	$[-7.383, 8.901]$	

DeepSeek produces ten invalid slots: correct/swapped score labels yield 0/1 on CIFAR and 5/4 on GoEmotions, out of 256 slots per arm; all Terra slots are valid. Averaging per-trajectory valid-only means gives descriptive DeepSeek contrasts of $1.309\,\mathrm{pp}$ and $1.156\,\mathrm{pp}$, respectively, with treatment-dependent denominators. Allowing invalid-slot utilities anywhere in $[0, 100]$ gives proposal-effect ranges $[1.154, 1.544]$ and $[-0.591, 2.925]\,\mathrm{pp}$. These descriptive ranges vary invalid-slot utilities while keeping subsequent histories fixed and leaving sampling uncertainty unmodeled; primary inference retains all assigned slots.

B.1.3 Qualification frequency and conditional composition

An arm's proposal mean reflects both how often a qualification group occurs and the quality of its members. Center each proposal on the shared prefix F winner i_0: $z = U(x) - U(i_0)$. If a group occurs with share p and its mean z is μ, its contribution to the full-cohort mean is $C = p\mu$. This zero-filled contribution differs from the group's conditional mean. The symmetric arithmetic split

$$\Delta C = \Delta p \, \frac{\mu_a + \mu_s}{2} + \Delta \mu \, \frac{p_a + p_s}{2}$$

uses correct/swapped arms a/s. Group membership and prefix weights can change, so this allocation combines frequency and within-group composition changes. Table 7 uses 20,000 paired-block bootstrap resamples (seed 20260924) and nominal 95% percentile intervals, without new tests or multiplicity correction. Empty groups stay undefined; none occurs in these resamples.

Table 7: Composition of the score-label proposal effects. Top: swapped → correct; q is the qualified share of 512 occurrences per arm, and μ is mean proposal U minus starting-winner U within the indicated group (task-specific pp). Bottom: the symmetric frequency and conditional-composition parts of the unqualified contribution, with exploratory nominal 95% paired-block bootstrap intervals. Their sum equals the unqualified contribution in Table 5.

Task	Agent	q (%)	$\mu_{\text{unqualified}}$	$\mu_{\text{qualified}}$
CIFAR-100	DeepSeek	$9.96 \to 25.78$	$-3.313 \to -2.260$	$-0.850 \to -0.364$
CIFAR-100	Terra	$12.50 \to 21.48$	$-2.598 \to -1.940$	$-1.083 \to -0.133$
GoEmotions	DeepSeek	$18.75 \to 30.08$	$-1.097 \to -0.353$	$0.244 \to 0.610$
GoEmotions	Terra	$15.62 \to 23.83$	$-1.953 \to -0.804$	$0.835 \to 0.249$

Task	Agent	Frequency part	Conditional-composition part
CIFAR-100	DeepSeek	$+0.441$ $[0.307, 0.594]$	$+0.865$ $[0.358, 1.386]$
CIFAR-100	Terra	$+0.204$ $[0.082, 0.348]$	$+0.546$ $[0.140, 0.959]$
GoEmotions	DeepSeek	$+0.082$ $[-0.049, 0.239]$	$+0.562$ $[-0.155, 1.326]$
GoEmotions	Terra	$+0.113$ $[0.018, 0.262]$	$+0.922$ $[0.318, 1.621]$

B.1.4 Available, visited and retained improvement

Of 64 block–target prefixes per task, CIFAR has 6 without an F challenger, 26 with challengers but none improving U, and 32 with beneficial challengers; GoEmotions has 14, 15 and 35. These counts use all 32 candidates. Mean challenger counts are 3.766/3.344; mean maximum positive U improvement among F-qualified candidates is 1.885 accuracy/2.423 micro-F1 pp. Agents and draws share these 64 contexts within 32 blocks.

For the visited candidates, $A = \max(\{0\} \cup \{U(x) - U(i_0) : F(x) > F(i_0),\ x \text{ visited}\})$ and $L = U(i_0) + A - H$. With $Z = H - U(i_0)$, the path identity is $L = (A - Z_+) + Z_-$; hence $\mathbb{E}L = g_+ + \mathbb{E}Z_-$ (Appendix C.1). Table 8 uses exploratory nominal 95% paired block-t intervals, with $\Delta H = \Delta A - \Delta L$.

Table 8 compares this account with discovery Q and regret r over all visited candidates (Appendix C.1) on the same trajectories. The two accounts separate discovery from retention using different candidate sets and baselines: qualification measures opportunity relative to the starting F winner. In CIFAR/DeepSeek, both show greater mean discovery and loss, with uncertain return. The intervals quantify uncertainty in each account separately. Availability in the complete candidate set additionally uses scores for unvisited candidates; the replacement account in Table 5 instead follows ordered incumbent histories.

Table 8: Two endpoint accounts of the same original score-label trajectories. Correct minus swapped, task-specific pp; all 128 trajectories per arm and 32 paired blocks. Q is improvement over the best prefix U; r is selection loss over all visited candidates. A is best positive U improvement among F qualifiers relative to the prefix F winner, and L its selection loss. Both $\Delta Q - \Delta r$ and $\Delta A - \Delta L$ equal the return effect. Brackets: exploratory nominal 95% block-t intervals for each reported effect separately. Table 5 reports pathwise replacements for the same cohorts.

Task / agent	Discovery ΔQ	Loss Δr	Qualified ΔA	Loss ΔL
CIFAR-100	$+0.250$	$+0.432$	$+0.411$	$+0.594$
DeepSeek	$[0.012, 0.488]$	$[-0.008, 0.873]$	$[0.109, 0.714]$	$[0.108, 1.080]$
CIFAR-100	$+0.234$	$+0.104$	$+0.214$	$+0.083$
Terra	$[-0.141, 0.609]$	$[-0.347, 0.555]$	$[-0.310, 0.737]$	$[-0.151, 0.317]$
GoEmotions	-0.014	-0.130	$+0.112$	-0.003
DeepSeek	$[-0.444, 0.415]$	$[-0.656, 0.396]$	$[-0.125, 0.350]$	$[-0.316, 0.310]$
GoEmotions	-0.122	$+0.047$	$+0.118$	$+0.287$
Terra	$[-0.416, 0.172]$	$[-0.449, 0.543]$	$[-0.166, 0.401]$	$[-0.125, 0.698]$

Among all 128 trajectories per arm, beneficial-challenger visits change from swapped to correct labels as 11 to 25, 11 to 13, 39 to 53 and 31 to 36 in Table 1 order. Final improvement counts are 10 to 20, 11 to 12, 37 to 50 and 31 to 34. Frequencies omit gain/loss magnitudes and therefore need not track mean return. All paths remain in every denominator, including those with no higher-F candidate available or no replacements.

Table 9 reports G contributions from the three starting conditions described above. For metric $M \in \{P, H, A, L\}$ and condition k, its contribution is $C_{k,M} = 32^{-1} \sum_b 4^{-1} \sum_{r \in b} \mathbf{1}\{r \in k\}(M_{r,a} - M_{r,s})$. Thus $\sum_k C_{k,M} = \Delta M$, $C_{k,G} = C_{k,P} - C_{k,H}$ and $C_{k,H} = C_{k,A} - C_{k,L}$. Missing categories contribute zero within a block. Without an F challenger, H cannot change; with F challengers but no beneficial one, both arms can incur losses of different sizes. The categories precede treatment, but this analysis is post hoc and depends on fixed-panel U. Path states such as discovery without positive return describe outcomes under each treatment and can differ across treatments.

Table 9: Score-label gap contributions and stability (task-specific pp). The three starting conditions contribute to $G = \Delta P - \Delta H$ using the full-cohort denominator. Counts are 6/26/32 of 64 CIFAR starts and 14/15/35 GoEmotions starts. Contribution intervals are descriptive nominal 95% block-t, with all 32 blocks retained. Draws share materials; deletion ranges span 32 leave-one-block-out point estimates. Full-cohort simultaneous inference remains in Table 1.

Task	Agent	No F challenger	F only	F and U	Draw 1 / 2	Delete-one range
CIFAR-100	DeepSeek	$+0.108$ $[-0.028, 0.244]$	$+0.691$ $[0.288, 1.095]$	$+0.680$ $[0.190, 1.169]$	$1.161/1.797$	$[1.360, 1.577]$
CIFAR-100	Terra	$+0.207$ $[-0.001, 0.415]$	$+0.210$ $[-0.133, 0.553]$	$+0.310$ $[-0.100, 0.720]$	$1.010/0.443$	$[0.626, 0.827]$
GoEmotions	DeepSeek	$+0.287$ $[-0.105, 0.679]$	$+0.197$ $[-0.136, 0.529]$	$+0.183$ $[-0.386, 0.752]$	$0.664/0.669$	$[0.471, 0.833]$
GoEmotions	Terra	$+0.193$ $[-0.092, 0.477]$	$+0.611$ $[0.021, 1.201]$	$+0.330$ $[-0.105, 0.764]$	$0.801/1.465$	$[0.914, 1.258]$

The final columns of Table 9 retain positive G point estimates in both draws and every block deletion for all four settings. The full-data intervals and multiplicity corrections remain those in Table 1; these sensitivity checks reuse the same data under the original, post hoc analysis choices.

A sensitivity comparator randomly ends some CIFAR/DeepSeek searches before continuation to match replacement frequencies across conditions; stop probabilities are fitted on the opposite block fold. Its opportunity contrast is $+0.057$ pp with a nominal interval spanning zero; adjusted return is $+0.025$ $[-0.320, 0.408]$ pp. Because stopping changes expected compute, this is a comparison under a different budget. It weakens selective-improvement explanations beyond competition frequency.

B.2 Score coverage and comparison organization

The original experiment compares organization under partial and complete scores. Follow-ups test whether the response persists with fresh decisions and materials, then separate the display's components. Here I denotes score coverage and O denotes display organization; 0 and 1 index partial/complete coverage or visit-order/organized display, respectively. Each cohort retains its own test family, with separate criteria for component interactions and reversal under combined sorting and comparison.

Original policies. Policies preserve the selector and eligibility rules of Appendix A.3. Across both agents, four three-step branches from each agent's 128 saved histories yield 1,024 continuations and 3,072 proposals, with four cyclic orders balanced within blocks. Partial coverage lists disclosed requested F; complete coverage supplies all eligible F. Visit order uses first-eligible order; organized comparison sorts disclosed scores in ascending order, places hidden scores last, marks the highest disclosed score, and reports each disclosed score minus it. Each distinct configuration appears once, at its first eligible occurrence; earliest occurrence breaks score ties. The marked maximum uses only displayed scores. A/B feedback, configurations and disclosure for the other target are common.

Table 10: A recorded Terra first-continuation display, selected after inspection. Feature names label four initial configurations; the first proposal repeats Words+pairs. Partial coverage retains disclosed scores ("?" is hidden); organized comparison sorts disclosed scores, marks their maximum, and shows each disclosed score minus that maximum in parentheses. Other input is common; feature labels are paraphrases.

	Visit order	Organized comparison
Partial	Words: ?; Words+pairs: 15.8; Chars 3–5: ?; Chars 4–6: 0.0.	Chars 4–6: 0.0 (-15.8); Words+pairs: 15.8 (0.0); then Words: ?, Chars 3–5: ?. Highest disclosed: Words+pairs.
Complete	Words: 0.0; Words+pairs: 15.8; Chars 3–5: 15.3; Chars 4–6: 0.0.	Words: 0.0 (-15.8); Chars 4–6: 0.0 (-15.8); Chars 3–5: 15.3 (-0.5); Words+pairs: 15.8 (0.0). Highest disclosed: Words+pairs.

All branches begin from the same saved history and evolve separately; the partial/visit-order reference is also freshly generated. Candidates enter before feedback and remain eligible. The three-proposal budget and cached candidate scores are fixed. Coverage changes how much is shown, and organization changes ordering and explicit comparisons; prompt length and inference cost vary.

Endpoints include the first-continuation proposal's F score and U/W outcomes, final returns, and gains relative to each starting history; we also provide a uniform-draw reference. U/W reuse inspected comments and unrounded micro-F1; summaries retain four target–draw starting histories per block. An exact reference appends three uniform draws with replacement from all 32 configurations, using canonical F and earliest ties. Its mean U continuation gains are $-0.209/-0.104\,\mathrm{pp}$ for DeepSeek/Terra; W gains are $+0.013/-0.005\,\mathrm{pp}$. All 16 observed-minus-reference intervals include zero. The reference samples uniformly from the finite candidate set, while the agent branches follow adaptive policies.

Table 11: Original first-continuation new visits: 128 histories, 32 blocks per agent. Organized-minus-visit-order effects use probability pp and nominal 95% block-t_{31} intervals. ^{z} denotes zero empirical variance; the sample contrast is reported descriptively.

Agent	Coverage	Visit order	Organized	Difference [95% CI]
DeepSeek	Partial	128/128	125/128	$-2.344\ [-5.013, 0.326]$
DeepSeek	Complete	128/128	128/128	0.000^{z}
Terra	Partial	113/128	90/128	$-17.969\ [-29.004, -6.933]$
Terra	Complete	93/128	119/128	$+20.312\ [11.030, 29.595]$

Original return contrasts. For return H_{ij}, coverage $i = 0/1$ means partial/complete scores and format $j = 0/1$ means visit order/organized comparison. Then $\tau_O(i) = H_{i1} - H_{i0}$, $\tau_I(j) = H_{1j} - H_{0j}$ and $\eta = \tau_O(1) - \tau_O(0)$. Table 12 gives all planned U/W return contrasts; Table 14 gives the three-step qualification contrasts. Figure 4 keeps agents and cohorts separate.

Both agents' original U organization interactions remain unconfirmed, and the W effects do not always have the same direction as the U effects.

Qualification and available score improvement. Table 14 reports all original simple behavior contrasts, retaining 128 starting histories and 384 proposal occurrences per arm. Qualification uses the starting F winner, including repeats and zero-qualified paths. Terra's new-visit effects per three-proposal continuation are $-0.461\ [-0.707, -0.215]$ visits at partial coverage and $+0.578\ [0.353, 0.803]$ at complete coverage. Dividing by three and multiplying by 100 gives the table's probability-pp scale.

In 93 Terra starting histories already displaying the true eligible F maximum, the first-continuation organization effects under partial/complete coverage contribute $-13.281\ [-20.869, -5.693]$ / $+12.500\ [5.632, 19.368]$ points on the full 128-history denominator. In the 88 starting histories with a higher-F candidate available, contributions are $-14.844\ [-24.244, -5.443]$ /

Table 12: Coverage and organization return effects: requested-outcome pp, 95% paired block intervals, and positive/negative/zero blocks (32 total). U/W evaluate identical paths and F winners.

Agent / contrast	U: ΔH [95% CI]; signs	W: ΔH [95% CI]; signs
DeepSeek $\tau_O(0)$	$+0.223$ $[-0.200, +0.646]$; 10/9/13	-0.060 $[-0.448, +0.327]$; 8/8/16
DeepSeek $\tau_O(1)$	-0.225 $[-0.546, +0.095]$; 7/8/17	$+0.310$ $[-0.136, +0.755]$; 10/4/18
DeepSeek $\tau_I(0)$	$+0.150$ $[-0.284, +0.585]$; 10/11/11	-0.093 $[-0.409, +0.223]$; 7/12/13
DeepSeek $\tau_I(1)$	-0.298 $[-0.815, +0.219]$; 9/13/10	$+0.277$ $[-0.324, +0.878]$; 9/11/12
DeepSeek η	-0.448 $[-0.992, +0.095]$; 10/13/9	$+0.370$ $[-0.356, +1.096]$; 12/8/12
Terra $\tau_O(0)$	$+0.127$ $[-0.309, +0.563]$; 8/9/15	$+0.036$ $[-0.398, +0.469]$; 10/7/15
Terra $\tau_O(1)$	-0.392 $[-0.967, +0.184]$; 6/10/16	$+0.002$ $[-0.541, +0.545]$; 6/10/16
Terra $\tau_I(0)$	$+0.142$ $[-0.336, +0.620]$; 12/11/9	$+0.239$ $[-0.182, +0.660]$; 10/13/9
Terra $\tau_I(1)$	-0.377 $[-1.101, +0.347]$; 11/10/11	$+0.205$ $[-0.385, +0.796]$; 9/12/11
Terra η	-0.519 $[-1.456, +0.418]$; 10/11/11	-0.034 $[-0.429, +0.362]$; 8/12/12

$+12.500$ $[3.934, 21.066]$. These exploratory pre-treatment strata retain the same 32 blocks. Displaying the maximum value does not ensure that the agent knows it is maximal.

For disclosure with visit-order displays, split prefix-centered proposal outcome by the fixed qualification threshold and divide both contributions by all three proposals. Qualified/unqualified U changes are $+0.034/-0.742$ pp for DeepSeek and $+0.046/+0.871$ pp for Terra; on W they are $-0.051/+0.162$ and $+0.086/+0.269$ pp. The zero-filled unqualified terms account for most of the opposite U means, mixing frequency and conditional composition; these directions do not repeat on W.

Table 13 uses A, Z, g_+ and Z_- from Appendix C.1. Each agent/panel retains every original starting history, including those with no higher-F candidate available. U-beneficial candidate–prefix pairs often fail to remain beneficial on W; this decomposition compares outcome panels on the same recorded paths and previously inspected comments.

Table 13: Score disclosure under visit order: the same paths connect opportunity to return. Complete minus partial scores, micro-F1 pp with 95% paired block intervals; A is the largest visited qualified improvement, g_+ the mean gap between available and retained positive improvement, Z_- ending below the start. Each column retains all 128 starting histories and 32 blocks. $\Delta\mathbb{E}Z = \Delta\mathbb{E}A - \Delta g_+ - \Delta\mathbb{E}Z_-$. Opportunity components are exploratory; returns are the design's prespecified endpoints. Z_- differs from cumulative replacement loss.

	DeepSeek/U	DeepSeek/W	Terra/U	Terra/W
Opportunity $\Delta\mathbb{E}A$	0.357 $[0.046, 0.668]$	0.185 $[-0.032, 0.402]$	-0.178 $[-0.490, 0.133]$	0.075 $[-0.258, 0.407]$
Gap Δg_+	0.026 $[-0.076, 0.127]$	0.140 $[-0.030, 0.311]$	-0.067 $[-0.164, 0.031]$	-0.077 $[-0.197, 0.043]$
Endpoint loss $\Delta\mathbb{E}Z_-$	0.181 $[-0.017, 0.378]$	0.138 $[-0.076, 0.351]$	-0.254 $[-0.631, 0.124]$	-0.088 $[-0.289, 0.113]$
Return $\Delta\mathbb{E}Z$	0.150 $[-0.284, 0.585]$	-0.093 $[-0.409, 0.223]$	0.142 $[-0.336, 0.620]$	0.239 $[-0.182, 0.660]$

Table 14: All original simple behavior contrasts. Organization is organized comparison minus visit order; disclosure is complete minus partial scores, at the stated level of the other factor. Each arm has 384 proposal occurrences from 128 starting histories in 32 blocks. Visits and qualification use probability pp; mean U/W use micro-F1 pp. Reported intervals are nominal 95% block intervals; [d] denotes a descriptive estimate without an interval; [z] denotes zero empirical variance, reported descriptively without a population interval.

Agent	Contrast	New visits	F-qualified	Mean U	Mean W
DeepSeek	Organization / partial	-0.78 $[-2.37, 0.81]$	$+1.56$ $[-1.62, 4.75]$	$+0.07^{d}$	$+0.23^{d}$
DeepSeek	Organization / complete	$+0.00^{z}$	-3.39 $[-6.12, -0.65]$	-0.25^{d}	$+0.07^{d}$
DeepSeek	Disclosure / visit order	$+0.52$ $[-0.22, 1.26]$	$+5.21$ $[1.90, 8.51]$	-0.71 $[-1.35, -0.07]$	$+0.11$ $[-0.50, 0.72]$
DeepSeek	Disclosure / organized	$+1.30$ $[-0.04, 2.65]$	$+0.26^{d}$	-1.03^{d}	-0.05^{d}
Terra	Organization / partial	-15.36 $[-23.55, -7.18]$	-1.56 $[-5.70, 2.58]$	-0.28^{d}	-0.13^{d}
Terra	Organization / complete	$+19.27$ $[11.78, 26.77]$	$+7.03$ $[2.20, 11.86]$	-0.75^{d}	-0.27^{d}
Terra	Disclosure / visit order	-14.84 $[-23.37, -6.32]$	-2.08 $[-7.79, 3.63]$	$+0.92$ $[0.08, 1.76]$	$+0.36$ $[-0.27, 0.98]$
Terra	Disclosure / organized	$+19.79$ $[11.87, 27.71]$	$+6.51^{d}$	$+0.45^{d}$	$+0.21^{d}$

Actual replacements. Under complete coverage, mean replacements per three-proposal continuation change from visit order to organized comparison as 0.445 to 0.586 for Terra and 0.656 to 0.602 for DeepSeek. The paired effects are $+0.141$ $[0.029, 0.253]$ and -0.055 $[-0.122, 0.013]$, respectively, with nominal 95% intervals over the same 32 blocks. These counts compare each proposal with the current winner; U/W share the same F-selected replacements.

B.2.1 Fresh first decisions

Terra's 128 saved starting histories start six fresh first-continuation arms: partial/complete scores crossed with visit order, organized list a and table b (768 proposals). Table b puts the same candidates, scores, ordering, highest disclosed score, differences from it, and coverage text in three columns; differences for hidden scores are blank. System, history, tools and "3 proposals remain" are preserved despite single-continuation execution. Cyclic randomization (seed 20260914) gives 21/22 placements per position. The 24-proposal pilot is included; the decision to continue did not depend on outcomes. Scores and starting histories are reused; all continuation responses are newly generated.

With N indicating a candidate outside the starting eligible set, let $\tau_j(i) = \mathbb{E}[N_{i,1j} - N_{i,0}]$. Prespecified confirmation requires $\tau_{a,b}(0) < 0$ and $\tau_{a,b}(1) > 0$: four one-sided paired block-t_{31} tests at $\alpha = 0.05$, whose maximum p is the intersection–union p. Table 15 reports the condition counts and all four primary directional contrasts. The list supports both directions of the reversal; the weaker partial-score table response leaves the four-direction criterion unmet. The within-template intersection–union tests give $p = 0.012178$ for list a and $p = 0.059339$ for table b; the latter is also the four-direction joint p. All four Holm-adjusted p values for the prespecified opposite directions equal 1.

The list/table coverage interactions are 31.250 $[14.426, 48.074]$ / 28.906 $[13.360, 44.452]$ pp, with 20/5/7 and 23/4/5 positive/negative/zero block contrasts.

Table 15: Fresh first-continuation decisions on the original Terra histories. A: new-candidate counts out of 128 per condition. B: format minus visit order, in probability pp, with nominal two-sided 95% paired block-t_{31} intervals and prespecified directional tests. Signs count positive/negative/zero block contrasts.

A. New-candidate counts

Coverage	Visit order	List a	Table b
Partial	103	86	94
Complete	80	103	108

B. Directional contrasts

Format	Coverage	Effect [95% CI]	Directional p	Block signs
List a	Partial	-13.281 $[-24.725, -1.838]$	0.012178	5/16/11
List a	Complete	$+17.969$ $[8.464, 27.473]$	0.000272	15/2/15
Table b	Partial	-7.031 $[-15.967, 1.905]$	0.059339	8/13/11
Table b	Complete	$+21.875$ $[8.773, 34.977]$	0.000923	19/5/8

B.2.2 New materials and new starting proposals

Seeded matching and distinct-comment allocation use all 19 feasible vertex-disjoint new rater pairs: 32 F and 32 U comments per block, 1,216 in total. Previously allocated IDs, normalized-text duplicates and original pair identities are excluded. The 38 raters and original 20,000 training/5,000 development comments remain shared. Each block trains classifiers for the same 32 configurations with new training seeds (608 fits). The study thus uses new comments, rater pairings and fits while retaining the raters, training sources, agent and configuration fields.

Each target receives four fresh unscreened initial proposals. The resulting 152 histories each branch into four coverage/format policies with three continuations: 1,976 new proposals in total. Repeats, histories with no higher-F candidate available, and the operational pilot are included. The preregistered criterion is $D_{\text{partial}} < 0$ and $D_{\text{complete}} > 0$ for first-continuation new visits. Average eight paired starting histories within each of 19 blocks; joint p is the larger one-sided t_{18} value. Two-sided 95% intervals and the interaction are reported separately.

Table 16: New-material first-continuation organization effects and their interaction, in probability pp. The joint directional $p = 0.048078$ meets the prespecified criterion. The complete-coverage effect's two-sided interval still includes zero; the joint test combines the two directional effects, with the interaction reported separately.

Contrast	Estimate [95% CI]	p	Test
D_{partial}	-13.158 $[-21.790, -4.526]$	0.002467	One-sided, negative
D_{complete}	$+3.289$ $[-0.647, 7.226]$	0.048078	One-sided, positive
J_{IO}	$+16.447$ $[8.656, 24.239]$	0.000320	Two-sided

The positive complete-coverage direction depends on a few blocks: five are positive, one negative and thirteen zero. Leaving out any positive block raises the one-sided p to .0816–.0936; enumerating signs of the six nonzero blocks gives .09375. The latter assumes sign exchangeability conditional on magnitudes. These post hoc sensitivities show the fragility of the positive direction alongside the passed primary criterion. First-continuation new counts are 135/115/142/147 out of 152 in partial/visit-order, partial/organized, complete/visit-order and complete/organized order.

Across three steps, new-visit-share effects are -10.088 $[-16.826, -3.350]$ / $+4.605$ $[0.766, 8.445]$ points at partial/complete coverage. F-qualified-share effects are -5.921 $[-9.843, -2.000]$ / -2.412 $[-6.098, 1.274]$; U-return effects are -0.074 $[-0.725, 0.576]$ / $+0.178$ $[-0.361, 0.716]$ pp. These nominal secondary intervals do not confirm a competition or return gain under complete coverage.

B.2.3 ORGANIZATION COMPONENTS

This study crosses coverage (I), sorting (S) and comparisons (P) on 128 original Terra histories (32 blocks, two targets and two draws). Each history starts eight branches with three new proposals per branch (3,072 total). Coverage retains the original two levels. S controls ascending score order, breaking ties by first occurrence and placing unknown scores last; unsorted displays retain first eligible order. P marks the highest disclosed score, reports each disclosed score minus it, and explains that comparison. Raw requested scores and unknown markers match within I; omitting/including both components reproduces visit-order/organized messages. The allowed configurations, F selector, high-reasoning agent and saved starting histories remain fixed. U stays hidden. All assigned starting histories, including those with no higher-F candidate available, are retained, and these responses are analyzed separately from earlier response cohorts.

The contrasts were specified before collecting responses. The primary outcome indicates whether the first continuation chooses outside the starting eligible set. Paired contrasts within each starting history are averaged equally within each block, yielding 32 block means. J_{IS} and J_{IP} compare the complete versus partial coverage sorting and P effects, respectively, each averaged over the other component; J_{ISP} compares the S×P interaction across I. Two-sided block-t_{31} tests receive Holm correction over these three primary interactions. All 128 starting histories are retained.

Table 17: Component interactions. First-continuation new visits are primary probability pp; only these three tests receive Holm correction. Final U returns are secondary micro-F1 pp. Both use nominal 95% paired block-t_{31} intervals on all 128 starting histories in 32 blocks.

Contrast	First-step new [95% CI]	Raw p	Holm p	Final U [95% CI]
J_{IS}	-1.172 $[-9.917, 7.573]$	0.78643	1.00000	$+0.173$ $[-0.168, 0.514]$
J_{IP}	$+14.453$ $[4.937, 23.969]$	0.00412	0.01237	$+0.199$ $[-0.207, 0.604]$
J_{ISP}	$+2.344$ $[-14.536, 19.223]$	0.77892	1.00000	$+0.443$ $[-0.244, 1.131]$

P's effect on new visits depends on coverage: J_{IP} passes the primary family, while its superiority over the sorting interaction remains unconfirmed (Table 17). At each coverage i, averaged first-step visit effects satisfy $\tau_O(i) = \tau_S(i) + \tau_P(i)$, giving combined S+P effects $-7.03/ + 6.25$ points. Intervals here are nominal.

Table 18: Prespecified secondary reversal checks. Each requires a negative partial-coverage effect and positive complete-coverage effect; the joint p is the larger one-sided p, with Holm correction across these three checks. Effects are probability pp with nominal 95% intervals.

Contrast	Partial effect [95% CI]	Complete effect [95% CI]	Joint p	Holm p
Combined S+P	-7.031 $[-15.669, 1.606]$	$+6.250$ $[-2.617, 15.117]$	0.0803	0.2409
Sorting, average P	-7.812 $[-14.160, -1.465]$	-8.984 $[-14.502, -3.467]$	0.9988	1.0000
Comparisons	$+0.781$ $[-4.935, 6.498]$	$+15.234$ $[8.203, 22.266]$	0.6089	1.0000

Sorting reduces mean new visits at both coverage levels; P's mean effect is near zero under partial coverage and positive under complete coverage. The reversal under combined S+P remains unconfirmed in this study (Table 18). P changes the maximum marker, score differences and explanation together, so the supported interaction concerns these features as a group.

For combined S+P, U-return effects are -0.293 $[-0.845, 0.259]$ pp with partial scores and $+0.079$ $[-0.218, 0.376]$ pp with complete scores. Both intervals include zero.

B.2.4 MAXIMUM-SCORE GUARANTEE

On each of 93 Terra histories displaying the true eligible F maximum, four fresh first-continuation conditions cross the presence or absence of a guarantee with visit order or organized comparison (372 responses). These pre-treatment histories span 31 blocks and include 28 histories with no

higher-F candidate available. Scores remain partially disclosed. The guarantee states that no hidden eligible score exceeds the displayed maximum; the control makes no such statement. It supplies no hidden value and makes no claim about unvisited configurations. Prompts retain "3 proposals remain" although only the first continuation is collected.

For starting-history contrast z_{br}, block size m_b, $N = 93$ and $B = 31$, use

$$\widehat{J}_G = N^{-1} \sum_{b,r} z_{br}, \qquad \widehat{\mathrm{SE}}^2 = \frac{B}{B-1} N^{-2} \sum_{b} \left(\sum_{r} z_{br} - m_b \widehat{J}_G \right)^2 .$$

The prespecified test uses t_{30} and a $+10$-point attenuation reference. Visit-order/organized new counts are 84/78 without and 89/79 with the guarantee. Their organization effects are -6.452 $[-15.045, 2.141]$ and -10.753 $[-21.426, -0.080]$ points. The primary difference is $J_G = -4.301$ $[-16.432, 7.830]$, two-sided $p = .4746$: its interval excludes the specified large attenuation, but not smaller positive effects.

B.2.5 Revisits that reveal hidden scores

Organization and revisit disclosure are crossed on all 128 original Terra histories, yielding 1,536 continuation proposals. Scores are initially partial. One rule announces that revisits reveal hidden target scores before the next proposal; the other announces that they do not. New candidates reveal both scores under either rule. All displayed scores persist; summaries follow each branch's disclosure history. The saved starting history, score-free acknowledgement and canonical-F selector are unchanged.

The primary outcome is a first-continuation visit to the fixed set H_0 of demonstration candidates with initially hidden scores (excluding the last demonstration and initial proposal). All starting histories are retained; empty sets would receive zero. Let D_q be the organization contrast, with $q = 0/1$ according to whether revisiting discloses a hidden score, and $J_Q = D_1 - D_0$. Confirmation requires $D_1 > 0$ and $J_Q > 0$, using the larger one-sided t_{31} p-value on 32 paired block means. Visit-order/organized counts are 5/16 without and 44/50 with new-score disclosure. Effects are $D_0 = 8.594$ $[2.708, 14.480]$, $D_1 = 4.688$ $[-6.157, 15.532]$ and $J_Q = -3.906$ $[-14.173, 6.360]$ points. The prespecified joint criterion is unmet: the primary one-sided values $.1924/.7782$ give joint $p = .7782$. Later-step counts distinguish fixed H_0 membership from scores still hidden before the proposal. Both use 128 decisions per arm and round, or 384 proposals per arm overall.

B.3 Continued feedback and controlled choices

These studies expand Section 4.4. In the continued-score experiment, agents propose configurations from shared histories. The subsequent DeepSeek experiments hold the available options fixed and vary the comparisons or sample counts shown to the agent. Their separate constructions and inference units are specified below. The GoEmotions fixed-choice messages call the configurations "TF-IDF and SGD" in every condition; the evaluated classifiers use the hashing representation in Appendix A.2.

B.3.1 Continued feedback: local reuse and returned value

The continued-score comparison reuses all 32 blocks per task and 128 starting histories per agent. Each saved history is copied before its score feedback into two fresh three-step branches: continued canonical A/B scores or an explicit notice that no new feedback is provided. Candidate identities, target, previous history, remaining budget and fixed-selector instructions remain. This contrast and requested-F selection were prespecified. Each branch comprises three adaptive continuations from a common prefix.

To measure reuse of a local comparison, Table 19 pairs the last demonstration with the initial proposal when they differ in one field and F. Demonstration scores are visible; the intervention reveals or withholds proposal scores. Adoption means that the next proposal uses the field value of the higher-scoring candidate while at least one other field differs from the source pair. Previously visited configurations meeting this definition also count. Across the four settings, such comparisons occur in 145 histories across 17/15/16/14 blocks, respectively; means retain all 128 histories per setting, including repeats and zeros for histories without such a comparison. Adoption increases in all four settings, but the exploratory comparison combines field preference with context choice.

Table 19: Exploratory reuse of the field value favored by an earlier comparison. Adoption counts are continued/withheld scores at the first continuation; comparable histories contain a one-field, non-tied comparison of the last demonstration and initial proposal. Differences use all 128 starting histories and 32 blocks per setting, with nominal descriptive 95% block intervals and positive/negative/zero block counts.

Task / agent	Comparable histories	Adoption	Difference (points)	$+/-/0$
CIFAR / DeepSeek	37	37/19	$+14.06$ $[6.85, 21.28]$	13/0/19
CIFAR / Terra	25	24/14	$+7.81$ $[2.47, 13.15]$	10/1/21
GoEmotions / DeepSeek	45	44/32	$+9.38$ $[1.53, 17.22]$	7/0/25
GoEmotions / Terra	38	37/29	$+6.25$ $[0.64, 11.86]$	6/0/26

Discovery and return. Contrasts in discovery Q, selection loss r, and return H average four paired starting histories within each of 32 blocks (Appendix A.4). Each setting retains all 128 starting histories. The common-prefix identity is $\Delta H = \Delta Q - \Delta r$. In each of the four settings, both the discovery and requested-F return intervals include zero; the complete primary contrasts follow.

Table 20: Continued-feedback contrasts in requested-target pp. Continued minus withheld scores, with 95% paired block-t intervals; $\Delta r(h_F) = \Delta Q - \Delta H(h_F)$. h_G is the pooled-F comparator.

Study	ΔQ	$\Delta r(h_F)$	$\Delta H(h_F)$	$\Delta H(h_G)$
CIFAR-100 DeepSeek	0.307 $[-0.029, 0.644]$	0.255 $[-0.162, 0.672]$	0.052 $[-0.430, 0.534]$	0.401 $[-0.227, 1.030]$
CIFAR-100 Terra	0.094 $[-0.174, 0.361]$	0.172 $[-0.146, 0.490]$	-0.078 $[-0.316, 0.160]$	0.047 $[-0.366, 0.460]$
GoEmotions DeepSeek	0.021 $[-0.183, 0.225]$	0.072 $[-0.292, 0.436]$	-0.050 $[-0.369, 0.268]$	0.011 $[-0.486, 0.508]$
GoEmotions Terra	-0.048 $[-0.332, 0.236]$	-0.105 $[-0.650, 0.440]$	0.058 $[-0.551, 0.667]$	0.104 $[-0.281, 0.489]$

B.3.2 Comparison evidence for transfer to unshown configurations

This fixed-choice experiment asks whether a comparison that reverses a field preference changes transfer to a specified, unshown configuration. It uses DeepSeek on the existing candidate sets for both tasks, with no new fits. Construction uses only canonical requested F and configuration structure, never U or old agent effects. Four distinct fields denote choice j, moderator k, control context l and held-out context n; remaining fields are fixed. Five pairs vary the same two j values at $(k, l, n) = (0, 0, 0), (1, 0, 0), (0, 1, 0), (0, 1, 1), (1, 1, 1)$. Their strict F directions relative to the first pair are respectively $+, -, +, +, -$. Here 0/1 index actual field values. Deterministic sampling selects one eligible construction per block–target: 64 on CIFAR and 62 on GoEmotions, excluding two unsupported block–target combinations. All 32 blocks per task remain; inference conditions on constructions selected for the specified interaction pattern.

Both messages show the source pair. R additionally shows the moderator-change pair; C shows the ordinary control-context pair. Each contains four real scores in the same format. The same and reverse target pairs are distinct from all shown configurations; their scores, direction labels and recommended answers are withheld. In each of four independent calls, the agent chooses between the fixed target pair. R/C feedback messages jointly change score values and their configuration locations. C is closer to the same-direction target; the two added comparisons are equally distant from the reverse target. Inferring unshown scores from these comparisons requires structural assumptions.

All 504 allocated choices were valid. With correctness $C_{p,d}$ under message p and direction d, the prespecified contrasts are $D_{\text{rev}} = C_{R,\text{rev}} - C_{C,\text{rev}}$ and $J = D_{\text{rev}} + C_{R,\text{same}} - C_{C,\text{same}}$. The first measures the effect on reverse-target accuracy; J is twice the change in balanced accuracy across the two directions. We average available targets within each block, then 32 blocks per task, without zero-filling a singleton block's missing target. GoEmotions' reverse effect is therefore 98.44 pp, rather than the unweighted $61/62$ rate.

Table 21: Fixed-choice transfer experiment with DeepSeek. Top: correct choices per eligible target; R is the interaction comparison, C the ordinary comparison. Bottom: equal-block effects in pp, simultaneous 95% family intervals, and four-test Holm p. $J = D_{\mathrm{rev}} + D_{\mathrm{same}}$ is twice the balanced-accuracy gain; its possible range is $[-200, 200]$ pp.

Task	Same C	Same R	Reverse C	Reverse R
CIFAR-100	64/64	64/64	0/64	61/64
GoEmotions	62/62	62/62	0/62	61/62

Task	Contrast	Effect	Simultaneous interval	Holm p
CIFAR-100	D_{rev}	$+95.31$	$[55.57, 99.82]$	1.10×10^{-7}
CIFAR-100	J	$+95.31$	$[35.60, 116.95]$	1.88×10^{-4}
GoEmotions	D_{rev}	$+98.44$	$[58.34, 100.00]$	8.57×10^{-8}
GoEmotions	J	$+98.44$	$[38.27, 119.08]$	1.88×10^{-4}

Both positive contrasts are supported in each task (Table 21). Ordinary comparisons preserve the source choice even at reversed targets; the interaction comparison usually changes that choice while retaining same-direction accuracy. For DeepSeek, the evidence supports conditional transfer across the specified fixed-choice pairs on both tasks.

The four contrasts use the bounded-mean method in Appendix C.6, with simultaneous intervals at $\alpha = .05/4$, bounds $[-1, 1]$ for D and $[-2, 2]$ for J, and Holm correction.

B.3.3 Transfer with conflicting comparisons

This construction tests transfer when the nearest demonstrated pair favors the source winner and both feedback messages contain the same global score maximum. It uses new fixed-choice instances on the same candidate sets as the 504-choice study. F-only construction fixes 56 block–target units in 31 CIFAR blocks and 57 in 32 GoEmotions blocks; unsupported targets are not zero-filled.

Write S for the source pair, D for the nearest pair, R for the interaction-revealing pair and C for the control pair. Both feedback messages show S+D, with R or C as the additional pair: six truthful scores each. S/D contain the common global maximum, exceeding all scores in R/C. S, D and C favor source value a; R favors b. Two unshown target pairs favor a in the same-direction context and b in the reverse context. For the reverse target, non-choice-field Hamming distances to D/R/C/S are respectively 1/2/2/3. Anonymous candidate IDs, option order, record positions and request order are fixed in advance. Target scores and direction labels are never shown; unshown directions are not logically determined without structural assumptions.

For correctness $c_{p,d}$ under message p and target direction d, define

$$\begin{aligned} D_{\mathrm{rev}} &= c_{R,\mathrm{rev}} - c_{C,\mathrm{rev}}, \\ J &= D_{\mathrm{rev}} + c_{R,\mathrm{same}} - c_{C,\mathrm{same}}, \\ T_R &= c_{R,\mathrm{rev}} + c_{R,\mathrm{same}} - 1. \end{aligned}$$

The third contrast distinguishes target-dependent transfer from always choosing b under R. Such a constant rule can produce positive D_{rev} and J if C chooses a at reverse targets and b at same targets, yet gives $T_R = 0$. The first two contrasts alone therefore cannot exclude this explanation.

We average targets within supported evaluation blocks, then weight the 31/32 blocks equally. The two tasks' six tests share Holm correction and simultaneous intervals at $\alpha = .05/6$, with bounds $[-1, 1]$, $[-2, 2]$ and $[-1, 1]$. Confirmation requires all three contrasts to be positive within a task. Invalid choices count as incorrect and retain their slots; a/b/invalid counts are reported for every cell. F/U consequences among valid paired choices are descriptive with explicit denominators. A positive conjunction would limit the specified nearest-comparison, global-maximum and target-independent rules as complete explanations on these constructions, while leaving other mechanisms and search-policy performance open.

Choices with conflicting comparisons. All 452 assigned choice slots across 113 units yielded responses. One GoEmotions reverse/R response ended without a valid choice and retains its slot

as incorrect. Under C, all 56/57 same targets choose a correctly and all reverse targets choose a incorrectly. Under R, same-target choices remain perfect, while 12/56 CIFAR and 17/57 GoEmotions reverse targets choose b correctly; the latter also includes 39 a choices and one invalid response.

The three sample effects coincide within each task: $D_{\text{rev}} = J = T_R = 20.97$ pp on CIFAR and 31.25 pp on GoEmotions. Their bounds differ because J ranges over $[-2, 2]$. The prespecified conjunction is unmet in both tasks (Table 22), leaving the nearest-comparison, common-maximum and target-independent accounts unresolved. Materials and feedback messages differ from the earlier 504-choice study, so the accuracy gap combines these design changes.

Table 22: All six conflict contrasts. Effects and simultaneous 95% family intervals are in probability pp, averaging supported targets within evaluation blocks. J is twice the balanced accuracy change.

Task	Contrast	Estimate	Simultaneous interval	Holm p
CIFAR-100	D_{rev}	20.97	[-7.51, 43.46]	0.2361
CIFAR-100	J	20.97	[-28.58, 61.81]	0.3316
CIFAR-100	T_R	20.97	[-7.51, 43.46]	0.2361
GoEmotions	D_{rev}	31.25	[-0.31, 58.44]	0.0543
GoEmotions	J	31.25	[-20.16, 70.73]	0.2507
GoEmotions	T_R	31.25	[-0.31, 58.44]	0.0543

B.3.4 Evaluation-count disclosure at fixed displayed scores

This study isolates the effect of showing sample counts within a given score table. GoEmotions has the required stored per-comment prediction counts. Fix 32 blocks, two targets and two predetermined subset draws, giving 128 contexts. The incumbent is scored on all 32 F comments with its count always shown; three challengers use common nested 4- and 16-comment subsets. Within each subset size, show/hide changes only the challengers' count fields; scores, configurations and candidate order are identical. Across sizes the scores can change. Messages omit full-panel F scores for challengers, U outcomes, raw comments and annotations.

Let $K_{n,v}$ indicate a valid incumbent choice, where $n \in \{4, 16\}$ and v is show or hide. The primary contrasts are

$$D_4 = \mathbb{E}(K_{4,\text{show}} - K_{4,\text{hide}}), \qquad J_K = D_4 - \mathbb{E}(K_{16,\text{show}} - K_{16,\text{hide}}).$$

The hypothesis requires positive support for both. The two two-sided tests share Holm correction and simultaneous intervals at $\alpha = .025$, using bounds $[-1, 1]$ and $[-2, 2]$. Inference targets generated responses on 128 fixed contexts, assuming independent generation and a stable configuration. The 32-block summaries are descriptive. All intervals use the bounded-mean method of Appendix C.6, for repeated generation on these fixed contexts.

Invalid choices occupy their slots, give $K = 0$ and fall back to the incumbent for final quality. Scores on the full F sample provide a reused finite-sample benchmark. Because scores differ between subset sizes, the interaction compares disclosure effects at different score vectors. It cannot distinguish reliability weighting from a general shift toward the incumbent.

Effect of disclosing sample counts. All 512 assigned choices were valid. Table 23 reports every condition and both primary contrasts. Within four-comment tables, all 38 changed retention indicators move toward the incumbent. At sixteen comments, every paired candidate choice is unchanged.

Table 23: Sample-count disclosure and incumbent choice. Each show/hide cell contains 128 fixed contexts. Effects are probability pp with simultaneous 95% intervals for the two primary contrasts; the sixteen-comment contrast has zero empirical variance and is descriptive.

Contrast	Hide: retained	Show: retained	Effect [family interval]	Holm p
Four comments, D_4	26/128	64/128	29.69 [17.98, 44.32]	7.21×10^{-10}
Sixteen comments	26/128	26/128	0 (descriptive)	—
Interaction, J_K	—	—	29.69 [15.17, 43.53]	9.25×10^{-6}

Disclosure increases incumbent retention more at four than at sixteen comments. The primary tests support both $D_4 > 0$ and $J_K > 0$.

Mean full-sample F scores and U test quality of the final choices fall descriptively by 1.643 and 1.247 pp in the four-comment condition; the prespecified tests concern incumbent choice.

B.4 Additional evaluation information

Section 6 distinguishes two uses of additional evaluations: selecting among a fixed set of tried candidates and changing the feedback that generates new candidates. We report selection validation first, then expanded-feedback search, the generation–selection cross, and new search continuations on the same starting histories.

B.4.1 New-comment selection validation

This experiment asks whether additional selection information changes the return advantage of continued feedback on fixed sets of tried candidates. It retains the original 32 GoEmotions blocks, 64 raters, both agents, and trajectories with or without continued scores. For each rater, we allocate 24 previously annotated comments to selection panel V and 32 to outcome panel W, shared across draws, agents, arms, and selectors. Allocation, selection, contrasts and sample size were specified before scoring. Earlier training, development and evaluation comments and normalized duplicate text are excluded. Maximum flow assigns 56 globally distinct clear-annotated comments per rater; seeded shuffling splits them into 1,536 V and 2,048 W comments, disjoint across raters and panels. Original F/U are pair-shared; V/W are rater-specific, with the same raters. Shared-comment eligibility and per-rater allocation need not induce the same comment distribution: U/W are distinct finite-panel outcomes. F+V changes information amount and composition; no cross-rater margins are computed on V/W.

Classifiers refitted with the original training comments, configurations, seeds, and fitting settings matched the original F/U scores to numerical precision. New comments enter prediction only; selection is recomputed on the original search histories.

The expanded-score selector h_{FV} pools TP/FP/FN over the original 32 F plus 24 V comments, then recomputes one global micro-F1. It retains requested-rater maximization, one-decimal precision, earliest-occurrence ties, and the same eligible candidates; W does not enter selection. Scoring adds $24|\operatorname{unique}(R)|$ candidate–comment evaluations per trajectory, and reuses the same 1,536 V annotations.

Let B/N index continued/withheld scores. Writing $G_a = H_a^W(h_{FV}) - H_a^W(h_F)$, the primary paired interaction is

$$I = \mathbb{E}(G_B - G_N). \tag{5}$$

Targets and draws are averaged within 32 blocks separately for each agent; the selection gain on the starting history is counted once per starting history. Table 24 retains all 512 arm trajectories. For DeepSeek, selection changes in 51 continued-score and 55 withheld-score trajectories, each out of 128; the corresponding Terra counts are 48 and 52. All four mean selection gains are positive. The interaction estimates have opposite signs, but neither interval excludes zero.

Table 24: New-comment primary W comparisons: requested-rater micro-F1 pp with 95% intervals over all 32 block means per agent.

Quantity	DeepSeek	Terra
Continued scores, F selection	$32.756\ [29.889, 35.622]$	$32.661\ [29.742, 35.579]$
Withheld scores, F selection	$32.863\ [29.824, 35.901]$	$32.760\ [29.784, 35.736]$
Continued scores, F+V selection	$33.547\ [30.423, 36.671]$	$33.566\ [30.370, 36.762]$
Withheld scores, F+V selection	$33.917\ [30.721, 37.112]$	$33.207\ [30.011, 36.403]$
Difference, F selection	$-0.107\ [-0.613, 0.399]$	$-0.099\ [-0.583, 0.385]$
Difference, F+V selection	$-0.370\ [-0.919, 0.180]$	$0.359\ [-0.027, 0.745]$
Interaction I	$-0.263\ [-0.721, 0.195]$	$0.458\ [-0.154, 1.071]$

B.4.2 FEEDBACK-PANEL CONTINUATION AND RETURNED CANDIDATES

Giving additional evaluations to the agent. Each agent's 128 original starting histories (32 blocks, two targets, two draws) start two fresh three-step branches: original (F) or expanded (FV) feedback. Across both agents, the branches yield 1,536 proposals. Branch order is balanced 64/64 across each agent's starting histories; the reference is freshly generated.

F uses 32 pair-shared comments. FV pools their TP/FP/FN with each rater's 24 V comments (Appendix B.4.1): $100 \cdot 2(TP_F + TP_V)/[2(TP_F + TP_V) + FP_F + FP_V + FN_F + FN_V]$, zero for a zero denominator. Pooling the counts changes both sample size and composition. U/W evaluate the same raters on pair-shared/rater-specific comments; W was previously inspected.

Before the first continuation, feedback re-reports the last demonstration's scores and reports the initial proposal's scores for the first time. Both arms begin with the same instruction: "For the remaining proposals, use the following current evaluation scores. The earlier history is unchanged." Later continuations report the latest candidate. Target, format, allowed configurations, and budget stay fixed; earlier F history remains verbatim without adding earlier demonstrations' hidden requested scores. Expanded feedback changes the last demonstration's score in 124/128 starting histories per agent, by a mean absolute $4.913\,\text{pp}$; the initial proposal has no earlier displayed score. The fixed-score instruction remains, and the update neither names panels nor redefines "canonical"; selection uses original F.

Endpoints and inference. For prefix P, suffix S_g and $\mathcal{R}_g = P + S_g$, $g \in \{F, FV\}$, selector h_s maximizes the requested-target score on panel $s \in \{F, FV\}$ after rounding micro-F1 percentages to one decimal, with earliest ties. Test quality $Y^p(x)$ on hidden panel $p \in \{U, W\}$ remains unrounded. During generation the selector remains h_F; h_{FV} is applied offline to the ordered search record. The discovery, return and regret quantities Q, H, r (Appendix C.1) use $\mathcal{R}_g$ and prefix P; continuation is $G^p_{g,s} = H^p_{g,s} - Y^p\{h_s(P)\}$. W discovery is the primary endpoint; fixed-F W return is secondary, and U is auxiliary. All 128 starting histories use the block inference in Appendix A.4. Across four planned W contrasts, Holm-adjusted two-sided p values are $0.399/1.000$ for DeepSeek discovery/return and $1.000/0.032$ for Terra.

Table 25: Original/expanded-feedback means and paired differences with 95% block intervals, requested-target micro-F1 pp. H_F uses canonical F; the shared prefix gives $\Delta G_F = \Delta H_F$.

Agent	Panel	Endpoint	F	FV	FV−F [95% CI]
DeepSeek	W	Q	0.810	0.911	$0.101\ [-0.032, 0.234]$
	W	H_F	32.963	33.026	$0.063\ [-0.275, 0.401]$
	U	Q	0.882	1.005	$0.123\ [-0.087, 0.332]$
	U	H_F	31.775	31.797	$0.022\ [-0.401, 0.445]$
Terra	W	Q	0.549	0.607	$0.058\ [-0.158, 0.274]$
	W	H_F	32.749	33.344	$0.594\ [0.166, 1.023]$
	U	Q	0.837	0.560	$-0.277\ [-0.465, -0.088]$
	U	H_F	31.576	31.490	$-0.087\ [-0.475, 0.301]$

B.4.3 GENERATION, SELECTION, AND EACH SELECTOR'S STARTING RETURN

An improved final return can come from reselecting an existing candidate or from further search. Equation 4 separates the selector-specific prefix return B_s from continuation $G_{g,s} = H_{g,s} - B_s$; Table 26 reports both on the same cohort. Changing the selector holds generated candidates fixed; changing feedback generates new continuations. Giving V to the agent changes its displayed scores; giving V only to the selector leaves agent feedback unchanged. The two uses require different numbers of candidate evaluations, so their costs are not matched.

Table 26: Original/expanded-feedback generation g crossed with offline selector s on the same cohort, all 32 paired blocks per agent. A: final $H_{g,s}$ and selector-specific prefix B_s means. B: continuation $G_{g,s} = H_{g,s} - B_s$. Units are micro-F1 pp; intervals in B are exploratory nominal 95% paired block intervals. W/U measure the same paths; the replication cohort remains separate.

A. Final and prefix returns

Agent	Panel	$H_{F,F}$	$H_{F,FV}$	$H_{FV,F}$	$H_{FV,FV}$	B_F	B_{FV}
DeepSeek	W	32.963	33.326	33.026	33.652	32.985	33.562
Terra	W	32.749	33.249	33.344	33.792	33.082	33.369
DeepSeek	U	31.775	31.851	31.797	31.868	31.693	32.126
Terra	U	31.576	31.641	31.490	31.312	31.593	31.384

B. Continuing from each selector's prefix

Agent / panel	$G_{F,F}$	$G_{F,FV}$	$G_{FV,F}$	$G_{FV,FV}$
DeepSeek / W	-0.021 $[-0.812, 0.770]$	-0.236 $[-1.116, 0.644]$	0.042 $[-0.780, 0.864]$	0.090 $[-0.625, 0.805]$
Terra / W	-0.333 $[-1.174, 0.508]$	-0.119 $[-0.596, 0.358]$	0.262 $[-0.487, 1.010]$	0.423 $[-0.067, 0.914]$
DeepSeek / U	0.082 $[-0.555, 0.719]$	-0.275 $[-1.164, 0.614]$	0.104 $[-0.643, 0.852]$	-0.258 $[-0.990, 0.474]$
Terra / U	-0.016 $[-0.925, 0.892]$	0.257 $[-0.071, 0.585]$	-0.103 $[-0.890, 0.683]$	-0.072 $[-0.420, 0.276]$

B.4.4 New search continuations under original and expanded feedback

New continuations leave the return gain unconfirmed for both agents and both panels (Table 27). This preregistered replication generated two additional three-step continuations per starting history in each feedback condition (1,536 proposal slots per agent). Each target and starting-history draw is equally weighted within its block; repeated continuations are averaged within the starting history. The 95% intervals use 32 paired blocks. Terra/W's original positive estimate attenuates in each fresh repetition ($+0.168$ and $+0.089$ pp). The table separates fresh-only and original-plus-fresh estimates, testing generation variability on the same materials.

Table 27: Expanded-minus-original feedback returns (micro-F1 pp, 95% paired-block intervals). Each agent has 128 original starting histories per arm; fresh-only averages two new continuations per starting history, and combined averages all three.

Agent	Panel	Original	Fresh-only	Combined
Terra	U	-0.087 $[-0.475, 0.301]$	0.153 $[-0.128, 0.435]$	0.073 $[-0.202, 0.349]$
	W	0.594 $[0.166, 1.023]$	0.128 $[-0.236, 0.492]$	0.284 $[-0.037, 0.604]$
DeepSeek	U	0.022 $[-0.401, 0.445]$	0.078 $[-0.269, 0.425]$	0.059 $[-0.263, 0.382]$
	W	0.063 $[-0.275, 0.401]$	0.126 $[-0.395, 0.646]$	0.105 $[-0.276, 0.486]$

B.5 Replacement and measurement diagnostics

These analyses connect the recorded paths to the selection identities and examine the role of score measurement. They retain each study's candidates, outcomes and analysis cohort.

B.5.1 Replacement gains, losses, and terminal improvement

Gains and losses sum positive and negative selected-outcome increments (Equations 11 and 2). They include losses later recovered. Each contrast retains all 128 starting histories, averaging targets and draws within 32 blocks. Score-label interventions start from the four demonstrations; continued-score and expanded-feedback comparisons start after the initial agent proposal. The expanded-feedback experiment generates its own reference. Table 5 gives score-label results; the following table keeps the other cohorts and panels separate.

Across eight score-label/continued-score U comparisons, gain and loss means both increase; only three retain both increases in both draw splits.

Table 28: Replacement gain and loss contrasts, task-specific pp with nominal 95% paired block intervals. All 128 starting histories are retained per setting. The return effect equals the gain contrast minus the loss contrast; U/W measure the same paths.

Study / panel	ΔP_{repl}	ΔD_{repl}
Withheld → continued scores; three continuation proposals		
CIFAR-100	0.245	0.193
DeepSeek / U	$[-0.134, 0.624]$	$[-0.153, 0.538]$
CIFAR-100	0.026	0.104
Terra / U	$[-0.157, 0.209]$	$[-0.128, 0.336]$
GoEmotions	0.039	0.090
DeepSeek / U	$[-0.202, 0.281]$	$[-0.109, 0.288]$
GoEmotions	0.098	0.205
DeepSeek / W	$[-0.249, 0.445]$	$[-0.138, 0.547]$
GoEmotions	0.108	0.051
Terra / U	$[-0.232, 0.449]$	$[-0.403, 0.505]$
GoEmotions	0.276	0.375
Terra / W	$[0.012, 0.539]$	$[0.010, 0.739]$
Original → expanded feedback; three continuation proposals		
GoEmotions	0.135	0.113
DeepSeek / U	$[-0.227, 0.498]$	$[-0.089, 0.315]$
GoEmotions	0.029	-0.034
DeepSeek / W	$[-0.146, 0.203]$	$[-0.304, 0.236]$
GoEmotions	-0.197	-0.110
Terra / U	$[-0.421, 0.027]$	$[-0.464, 0.244]$
GoEmotions	0.074	-0.521
Terra / W	$[-0.194, 0.342]$	$[-0.885, -0.157]$

Draw sensitivity. Both draw splits and every single-block deletion retain positive score-label effects on proposal quality in all four settings. Only three of eight score-label/continued-score U comparisons retain positive replacement-gain and replacement-loss contrasts in both draws. These draws provide within-material sensitivity checks on shared panels; two draws and unset agent sampling seeds cannot separate generation, training-seed and material variance.

B.5.2 Score measurement on fixed search paths

On 2,048 original score-label and organization paths, unrounded F changes no winner. Uniform ties among distinct F-maximizers change 232 expected returns, but the 12 paired effect changes range from -0.0629 to $+0.0651$ pp. Both post hoc rules share 24 Holm tests (all adjusted $p = 1$); zero-variance contrasts are descriptive.

GoEmotions resamples 32 F comments with replacement 5,000 times (seed 20260923), independently across blocks, sharing weights within blocks across raters, candidates, agents and arms. Pooling TP/FP/FN yields displayed micro-F1 for reselection from the full starting and final search records, with U fixed. Starting winners change in 31.2% of path–replicates. Table 29 applies the same weights to score-label gaps and qualification. Since proposal U is fixed, $G^* = \Delta P - \Delta H^*$; range endpoints reverse. Qualification uses each replicate's full-prefix winner. Its mean contrast decreases, while central ranges remain positive for qualification and G. These conditional measurement ranges quantify F-panel resampling variation with proposal sequences and U outcomes held fixed. Per-image CIFAR perturbation is unavailable.

Table 29: Score-label sensitivity to whole-comment F resampling on fixed GoEmotions paths. Means and central 95% ranges over 5,000 replicates. Gap: micro-F1 pp; qualified share: percentage points. U and proposals stay fixed; each replicate reselects the full prefix before defining qualification.

Agent	Effect	Original	Perturbed mean	Measurement range
DeepSeek	Gap G	$+0.667$	$+0.703$	$[0.245, 1.134]$
DeepSeek	Qualified share	$+11.328$	$+5.825$	$[2.148, 9.375]$
Terra	Gap G	$+1.133$	$+1.047$	$[0.565, 1.503]$
Terra	Qualified share	$+8.203$	$+4.158$	$[0.391, 7.812]$

B.5.3 Outcome ties and return compatibility

Theorem 1 constructs search laws with matched quality processes and different returns. The following checks describe outcome ties and compatible returns in the candidate score tables and observed paths.

Using exact outcome fractions, all 64 block–target candidate sets per task/panel have ties; equal-Y candidates differ in canonical F in 64/64 CIFAR/U, 64/64 GoEmotions/U and 63/64 GoEmotions/W sets. These checks over complete candidate sets document outcome ties; matched search laws are supplied by the theoretical construction.

A compatibility check asks how often outcome ties leave return ambiguous in the observed search records. For each original GoEmotions coverage–organization trajectory, it fixes the prefix, F selector and three ordered exact outcome values, then retains every candidate sequence matching those values. Returned quality remains ambiguous for only 1–5 of 128 histories per agent–arm on U and 6–12 on W. Mean range widths are 0.012–$0.080\,\mathrm{pp}$ on U and 0.186–$0.420\,\mathrm{pp}$ on W. Adding each proposal's qualification relative to initial incumbent F narrows the ranges without specifying its identity. Observed-path ambiguity is therefore confined to a small fraction of these histories. These structural ranges enumerate returns compatible with each fixed observed quality sequence.

B.5.4 Empirical inversion bounds

The inversion bound and its sharpness are proved in Appendix C.5. Here we evaluate the resulting return-effect lower bounds on the observed candidate sets.

For each paired starting history, use the complete qualifying support $\mathcal{S}$ defined in Appendix C.5 and the sample lower bound $\Delta Z \geq \Delta A + \ell_0 - 2\eta^*(\mathcal{S}) = A_1 - Z_0 - 2\eta^*(\mathcal{S})$, where 0/1 denote reference/treatment. After averaging targets and draws within each block and weighting blocks equally, all 20 organization and expanded-feedback bounds remain negative (-3.386 to $-2.436\,\mathrm{pp}$). These inequalities bound return effects conditional on each fixed candidate set and its scores. Changing the selector requires its own prefix winner and qualifying set.

B.5.5 Worked trajectories and a final-step substitution

Table 30 distinguishes qualification from replacement in one recorded search. The third proposal exceeds the starting winner's F score but lowers U; repeating it qualifies against the starting winner without replacing the current winner. Increments total $58.000 - 59.333 = -1.333\,\mathrm{pp}$.

Table 30: A recorded CIFAR-100/DeepSeek path with correct score labels. The initial winner has $F = 63.3$ and $U = 59.33$. Qualification compares with the initial winner's F score; replacement depends on the current incumbent. This example illustrates harmful replacement and qualification without replacement.

Proposal	Visit	F	U	Qualified	Incumbent	ΔU
First	New	60.7	59.33	no	Initial→Initial	$+0.00$
Second	New	61.3	59.33	no	Initial→Initial	$+0.00$
Third	New	64.0	58.00	yes	Initial→Third	-1.33
Fourth	Repeat third	64.0	58.00	yes	Third→Third	$+0.00$

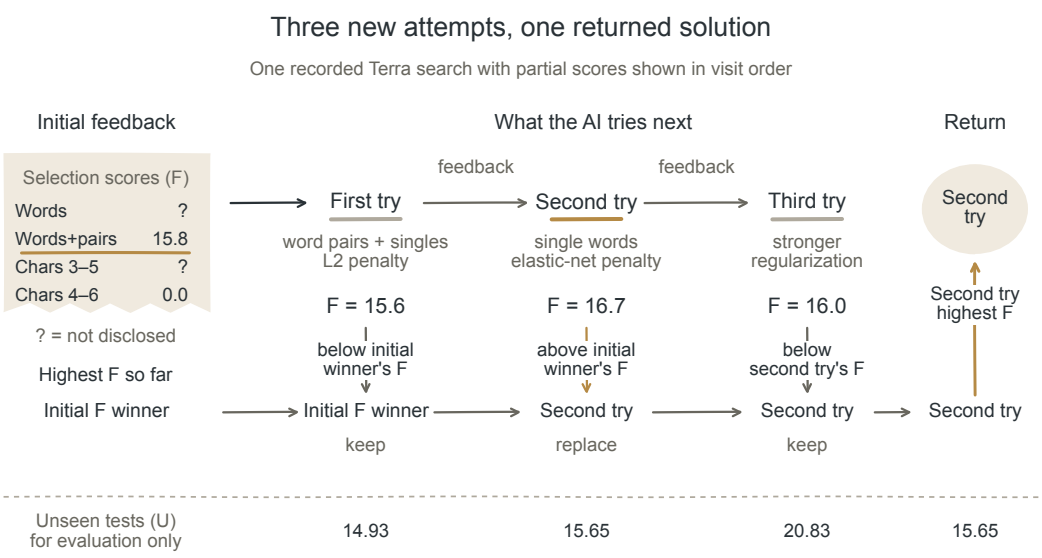


Figure 5: A better candidate can remain unselected. F selects the second try ($16.7 > 16.0$), while hidden U favors the third ($20.83 > 15.65$). Only F guides feedback and selection. The figure follows the first recorded Terra path.

Recorded substitution example. The example in Section 5.2 is a GoEmotions/DeepSeek first continuation with continued scores. The one-field substitution is evaluated offline using the fixed candidate scores and held-out outcomes.

C Derivations and proofs

We first derive the opportunity and replacement accounts used in Sections 3 and 4.2, then prove the quality-matching and return-preservation results of Section 5. The results condition on fixed scores and retained candidates. Empirical checks on the study's candidate sets appear in Appendix B.5.

C.1 Qualifying opportunities and returned improvement

We compare improvement available among qualifying proposals with improvement actually returned, fixing the block, target and outcome panel. Requested outcome $Y(x)$ is in pp; F is the fixed, one-decimal selection score. Let P be the nonempty ordered prefix, $S = (c_1, \dots, c_L)$ the suffix, and $i_j = h_F(P + (c_1, \dots, c_j))$, where $+$ concatenates ordered records. The starting winner has scores $f_0 = F(i_0)$ and $y_0 = Y(i_0)$; returned improvement is $Z = Y(i_L) - y_0$. Prefixes for score-label interventions contain four demonstrations; continuation prefixes also contain the initial proposal. The baseline is the prefix's highest-F candidate, with earliest ties. Let $\mathcal{E}_0 = (c_j : F(c_j) > f_0)$ retain qualifying suffix occurrences in increasing index order, including repeats. Define

$$A = \max\big(\{0\} \cup \{Y(c) - y_0 : c \in \mathcal{E}_0\}\big), \qquad A_{\text{inc}} = \max_{0 \le j \le L} \{Y(i_j) - y_0\}. \tag{6}$$

Here A is the largest positive improvement available among qualifiers, whereas A_{inc} records improvement reached by an actual incumbent. A qualifier beats the starting F score, but may lose to the stronger incumbent present on arrival. All visited candidates remain eligible for return. Write $Z_+ = \max(Z, 0)$ and $Z_- = \max(-Z, 0)$. Probabilities and expectations below concern the search law on this fixed candidate set and prefix.

Lemma 1 (Qualifying opportunities and final improvement)**.** *The reduced ordered record $\widetilde{R} = (i_0) + \mathcal{E}_0$ has the same F winner as $P + S$, and $0 \le Z_+ \le A_{\text{inc}} \le A$. Thus for every $t \ge 0$,*

$$\Pr(A > t) - \Pr(Z > t) = \Pr(A > t, Z \le t). \tag{7}$$

Let $\mathcal{G}_t$ retain, in order, the occurrences $c \in \mathcal{E}_0$ with $Y(c) - y_0 > t$. If it is nonempty, let q_t be its highest-F, earliest occurrence. A miss $A > t, Z \le t$ occurs exactly when a suffix occurrence with improvement at most t has higher F than q_t, or equal F and earlier occurrence.

Proof. No removed occurrence can beat i_0, including on ties. Every incumbent is therefore either i_0 or a member of $\mathcal{E}_0$, which proves winner preservation and the stated inequalities. The event inclusion $\{Z > t\} \subseteq \{A > t\}$ gives the probability identity. If a miss occurs, the final winner has improvement at most t and outranks q_t. Conversely, any such competitor outranks every member of $\mathcal{G}_t$, so none can be returned. □

Opportunity-to-return account. Integrating the threshold misses gives the expected improvement not retained in the final choice g_+. With bounded outcomes, the lemma gives

$$g_+ = \mathbb{E}(A - Z_+) = \int_0^\infty \Pr(A > t, Z \le t)\, dt, \qquad \mathbb{E}Z = \mathbb{E}A - g_+ - \mathbb{E}Z_-. \tag{8}$$

The first equality integrates the threshold identity; the second uses $Z = Z_+ - Z_-$. Endpoint loss Z_- excludes losses recovered before termination, unlike cumulative replacement loss.

Relation to accounting over all visited candidates. For the full ordered search record $R = P + S$, write $b_Y = \max_R Y$ and define

$$w = \max_P Y, \qquad Q = b_Y - w, \qquad r = b_Y - Y(i_L) \ge 0, \qquad r_0 = w - y_0. \tag{9}$$

Here w is the best prefix outcome, Q its improvement through discovery, and r the final shortfall from the best visited outcome. Thus $H = w + Q - r$ and, for two policies sharing the prefix, $\Delta H = \Delta Q - \Delta r$.

The loss on the reduced record is $\ell = A - Z \ge 0$, denoted L in Appendix B.1.4. The two accounts satisfy

$$\begin{aligned} r &= [b_Y - (y_0 + A)] + \ell, \qquad b_Y - (y_0 + A) = r_0 + Q - A \ge 0, \\ Z &= A - \ell = Q - (r - r_0). \end{aligned} \tag{10}$$

The bracket compares outcome maxima over two groups of visited candidates: another prefix candidate or a suffix candidate below the qualifying F threshold can contribute to b_Y. The accounts therefore use different candidate sets and baselines. Both describe the candidates reached under each policy.

C.2 Identifying changes in candidate competition

We use the fixed, rounded requested-target score F and hidden test quality Y from Section 2. Earliest occurrences break ties. All candidates remain eligible, and repeated candidates retain the same score.

For a single trajectory, let $R_t = R_{t-1} + x_t$ and $i_t = h_F(R_t)$: i_{t-1} is the winner before proposal x_t. Earliest ties imply the exact update

$$Y(i_t) - Y(i_{t-1}) = \mathbf{1}\{F(x_t) > F(i_{t-1})\}\,[Y(x_t) - Y(i_{t-1})]. \tag{11}$$

For policies sharing the same prefix, let $k_Y(i, x) = \mathbf{1}\{F(x) > F(i)\}[Y(x) - Y(i)]$ and let $\mu_t^a(i, x)$ be the policy-specific joint law of (i_{t-1}, x_t). The expectation identity is

$$\Delta\mathbb{E}H = \sum_{t,i,x} \Delta\mu_t(i, x) k_Y(i, x) = \Delta\mathbb{E}P_{\text{repl}} - \Delta\mathbb{E}D_{\text{repl}}. \tag{12}$$

Summing Equation 11 along each path and taking expectations under the history-induced incumbent–proposal law proves Equation 12. No independence between rounds or sufficiency of the incumbent as a state is needed. The identity decomposes return under a given incumbent–proposal law. Visited maxima are monotone within a path, not across non-nested arms.

C.3 Proof of Theorem 1

Construction. Order scores as $F(p) < F(b_1) < F(g_1) < \cdots < F(b_m) < F(g_m)$, with prefix p, $Y(g_i) = 1$ and $Y(b_i) = Y(p) = 0$. Choose i uniformly. A visits $\{b_i, g_i\}$; B visits $\{g_i, b_{i+1}\}$, cyclically; both randomize pair order.

Matching properties. Each non-prefix candidate has inclusion probability $1/m$ and per-round probability $1/(2m)$. Under either law, the ordered outcomes are $(0,1)$ or $(1,0)$, each with probability $1/2$, and always include 1.

Returns and attainable range. In A every g_i wins; in B it wins only for $i = m$, giving expected returns 1 and $1/m$.

Every matching law includes global F winner g_m with probability $1/m$, forcing $1/m \le \mathbb{E}H \le 1$; mixtures of A and B attain the interval. The construction matches the entire outcome process, but not joint (F, Y) paths, which determine return. Extra slots may repeat the prefix.

Injective outcomes and approximate matching. If Y is injective on the candidate set, its complete sequence identifies the candidate sequence and hence return. More generally, F being a function of Y suffices for identification under a fixed prefix and earliest ties. Exact nonidentification therefore requires ambiguity within outcome levels. Appendix B.5.3 examines such ambiguity in the recorded searches.

The identification is not uniformly stable to approximate matching. Keep the construction's scores and candidate laws, but for $0 < \varepsilon < 1/2$ set

$$Y(p) = 0, \quad Y(b_i) = \frac{\varepsilon i}{m+1}, \quad Y(g_i) = 1 - \frac{\varepsilon i}{m+1}.$$

All outcomes are distinct. Coupling the same i and low/high order gives $\max_t |Y(X_t^A) - Y(X_t^B)| \le \varepsilon$, while candidate marginals, inclusion and best-visited outcome laws remain identical. A always returns g_i; B returns b_{i+1} for $i < m$ and g_m otherwise. Thus $\mathbb{E}H_A = 1 - \varepsilon/2$, $\mathbb{E}H_B = 1/m + \varepsilon/2 - \varepsilon/m$, and the gap is $(1 - 1/m)(1 - \varepsilon)$. These approximately matched quality processes can therefore conceal a large return difference.

C.4 Proof of Proposition 1

Fix the earlier search history, whose highest-scoring candidate is z. Replace the final proposal x by x', where $F(x') > F(x)$ and $Y(x') \ge Y(x)$. Under earliest ties, the return change is

$$\Delta_z(x, x') = \begin{cases} Y(x') - Y(x), & F(z) < F(x), \\ Y(x') - Y(z), & F(x) \le F(z) < F(x'), \\ 0, & F(z) \ge F(x'). \end{cases} \tag{13}$$

In the first region, the assumed outcome improvement makes the return change nonnegative; in the third, the earlier winner is retained. In the middle region, nondecreasing return is equivalent to $Y(z) \le Y(x')$. Thus the substitution preserves return for every possible earlier winner exactly when all crossed competitors satisfy that inequality; earliest ties give the half-open F interval. Conditional on a shared earlier winner, a coupling supported on such substitutions or unchanged proposals cannot decrease expected return. Unconditional marginal coupling is insufficient, and early interventions can change subsequent competitors.

Appendix B.5.5 locates the recorded final-step substitution used in Section 5.2.

C.5 Inversion bound and sharpness

To bound selection loss, approximate outcome Y by a nondecreasing function of F. On finite support S, the smallest uniform approximation error is

$$\eta^*(S) = \tfrac{1}{2} \max_{x,y \in S:\, F(x) \le F(y)} [Y(x) - Y(y)]_+.$$

An outcome inversion of size d requires $\eta \ge d/2$. On the observed score levels, $\varphi(f) = \max_{F(x) \le f} Y(x) - \eta^*(S)$ is nondecreasing and lies within $\eta^*(S)$ of each outcome, attaining the bound, including ties. For ordered records, set inclusion and η^* refer to their candidate supports.

Theorem 2 (Inversion diameter governs selection loss)**.** *For any finite ordered search record $R = P + S$, let $\ell = A - Z$ be the selection loss on its reduced record $\widetilde{R} = (i_0) + \mathcal{E}_0$. Then*

$$\ell \le 2\eta^*(\widetilde{R}) \le 2\eta^*(R). \tag{14}$$

For a finite candidate set $\mathcal{X}$ containing i_0, define the complete qualifying support, including the prefix winner, by $\mathcal{S} = \{i_0\} \cup \{c \in \mathcal{X} : F(c) > F(i_0)\}$. With at least two proposal slots and repeats allowed over $\mathcal{S}$, the bound is sharp:

$$\max_{(c_1,\dots,c_L)\in\mathcal{S}^L} (A - Z) = 2\eta^*(\mathcal{S}), \qquad L \ge 2. \tag{15}$$

Moreover, $\ell = 0$ whenever $\eta^*(R) = 0$ (F and Y are order-consistent, including ties), and ℓ can be strictly positive only if the record contains an inversion.

Proof. Let x^* maximize Y on $\widetilde{R}$, and let $h = h_F(R) = h_F(\widetilde{R})$. Then $A = Y(x^*) - y_0$, $Z = Y(h) - y_0$, and $F(x^*) \le F(h)$. Both candidates belong to $\widetilde{R} \subseteq R$, so the definition of η^* gives

$$0 \le \ell = Y(x^*) - Y(h) \le 2\eta^*(\widetilde{R}) \le 2\eta^*(R).$$

This also proves the zero-inversion claim. For suffixes drawn from $\mathcal{S}$, both candidates lie in $\mathcal{S}$, giving the same upper bound with $\eta^*(\mathcal{S})$.

For sharpness, take a positive maximizing inversion (x, y) on $\mathcal{S}$, with $F(x) \le F(y)$ and $Y(x) - Y(y) = 2\eta^*(\mathcal{S})$. Necessarily $y \ne i_0$ and $F(y) > f_0$. Also $Y(x) \ge y_0$: otherwise (i_0, y) would have a larger inversion. Propose y first and x second, filling unused slots with repeats of y. The first proposal displaces i_0; y remains the winner by strict F or the earliest-tie rule. The maximum outcome on the reduced record is $Y(x)$, including when $x = i_0$, so the loss attains $2\eta^*(\mathcal{S})$. If $\eta^*(\mathcal{S}) = 0$, repeat i_0 and attain zero. □

For a selector maximizing a fixed total order, the same argument applies after defining qualification and inversions by that order. Proposition 1 identifies a related inversion: a crossed incumbent z with $Y(z) > Y(x')$ loses to x'. The diameter bounds this loss; its existence alone does not imply that a particular substitution lowers return.

The empirical return-effect bounds are reported in Appendix B.5.4.

C.6 Bounded-mean inference at a fixed sample size

The four comparison-transfer primary tests use Holm correction and intervals inverted at $\alpha = .05/4$ from ten fixed bets per direction (Waudby-Smith & Ramdas, 2024). Let $x_b \in [0, 1]$ be the rescaled block contrast, m a candidate mean and m_0 the rescaled null value. Use

$$K(m) = \frac{1}{20} \sum_{f=1}^{10} \left[\prod_{b=1}^{32} \left(1 - \frac{f}{11} + \frac{f x_b}{11m} \right) + \prod_{b=1}^{32} \left(1 - \frac{f}{11} + \frac{f(1 - x_b)}{11(1 - m)} \right) \right].$$

Invert $K(m) < 80$; the two-sided p is $\min\{1, 1/K(m_0)\}$. At $m = 0$ or 1, set $K = 1$ if all observations equal that endpoint and $K = \infty$ otherwise. The transfer contrasts D and J (Appendix B.3.2) have bounds $[-1, 1]$ and $[-2, 2]$ before rescaling.

Expected capital. Let $\mu_b = \mathbb{E}x_b$, $m = 32^{-1} \sum_b \mu_b$ and $\lambda = f/11$. For independent generation across the fixed blocks and $0 < m < 1$,

$$\mathbb{E} \prod_b (1 - \lambda + \lambda x_b/m) = \prod_b (1 - \lambda + \lambda \mu_b/m) \le \left[32^{-1} \sum_b (1 - \lambda + \lambda \mu_b/m) \right]^{32} = 1.$$

The same argument applies to the complementary factors with $1 - x_b$ and $1 - m$. Averaging the twenty products gives $\mathbb{E}K(m) \le 1$ while allowing heterogeneous block means.

Tests and intervals. Markov's inequality gives $\Pr\{K(m) \geq 1/\alpha\} \leq \alpha$. Thus $\min\{1, 1/K(m_0)\}$ is a valid p-value at the prespecified sample size, and inversion at $1/(.05/4) = 80$ gives the four simultaneous intervals by a union bound. Each product expands as a nonnegative sum of powers of $1/m$ or $1/(1-m)$, so it is convex on $(0, 1)$; the sublevel acceptance set is an interval. At an endpoint mean, bounded observations equal that endpoint almost surely, giving the conventions above.

This inference allows heterogeneous block means and evaluates their average at the prespecified sample size, conditional on the constructions, candidate scores and a stable agent configuration. Zero empirical variance is reported descriptively. The conflict and count-disclosure studies use the same construction with their specified unit count, bounds and family size, replacing 32 and the inversion threshold accordingly. The count study conditions on 128 contexts rather than independent material samples.

Logical ranges for fresh score-label responses. For context c, let $[l_{c,M}, u_{c,M}]$, $M \in \{P, H, G\}$, bound the path quantities P, H and $P - H$, respectively, over all allowed four-slot paths. The last quantity has paired contrast $G = \Delta P - \Delta H$. Invalid slots contribute zero to P and add no selectable candidate. Both arms share each range, so its paired contrast lies in $[-b_M, b_M]$, where $b_M = \max_c(u_{c,M} - l_{c,M})$ is shared across contexts within a task. Rescale each contrast d_c as $x_c = (d_c + b_M)/(2b_M)$; transform interval endpoints back by $d = 2b_M x - b_M$. Use 64 factors in K and threshold 240 for the 12 simultaneous intervals. The (b_P, b_H, b_G) values are $(72.000, 14.667, 75.333)$ for CIFAR and $(52.778, 22.608, 64.357)$ for GoEmotions.

D AGENT AND MESSAGE CONFIGURATIONS

Appendix A specifies the candidate sets, evaluation panels and search rules. Appendix B gives each intervention, its analysis unit and its results. This section states the common agent messages and generation settings.

D.1 COMMON MESSAGES AND RESPONSE FORMAT

Base system message. The GoEmotions system message used these four lines:

```
You are optimizing a fixed GoEmotions 28-label multilabel
text classifier.
Evaluator channels A and B are anonymous; their rater
mapping is hidden and may change by block.
Optimize the declared target channel using only the
displayed candidate history and feedback.
Call propose_candidate exactly once per round.
```

For CIFAR-100, the first two lines were "You are optimizing a fixed CIFAR-100 ResNet-18 training recipe." and "Evaluator channels A and B are anonymous; their physical corruption mapping is hidden and may change by block." The last two lines were shared. The `propose_candidate` call required every configuration field in Table 2 and accepted no extra properties. The score-free successful reply was "Candidate recorded."

Feedback template. The latest-candidate feedback used the following form. Bracketed entries are substituted values; `[s]` marks an optional plural suffix. Scores have one decimal. The metric is "validation accuracy" or "feedback-panel micro-F1"; the score-label intervention exchanges the values assigned to A and B.

```
Target channel:  [target]
Latest candidate:  [candidate JSON]
Channel A [metric]:  [score A] percentage points.
Channel B [metric]:  [score B] percentage points.
[remaining] proposal[s] remain including this round.
Propose the next candidate.
```

Continuation instructions. Shared-history continuations added these three system instructions:

```
All warm-start candidates and all candidates you propose
remain eligible for final selection; your last proposal is
not a final submission.
After the last proposal, a separate selector returns the
candidate with the highest canonical feedback score for your
declared target channel, using the displayed one-decimal
precision.  Ties are resolved by the earliest occurrence,
including the warm-start candidates.  The selector can use
these scores even when they are not displayed to you.
Within this search, each candidate has fixed evaluation
scores.  Repeating a candidate is allowed, but does not
produce a new independent measurement.
```

Intervention displays. Table 10 shows the four coverage–organization formats for a shared GoEmotions history. Appendix B.2.3 specifies the sorting and comparison components; Appendices B.2.4 and B.2.5 describe the two hidden-score interventions. The fixed-choice comparisons and count displays are specified in Appendix B.3; the F/FV continuation instruction and score update are in Appendix B.4.2.

Coverage–organization message template. In the four original GoEmotions arms, each continuation message appended this text after the latest-candidate feedback:

```
Requested target:  [target].  Score version:  canonical F.
Scores use one decimal place.
The final return uses the highest requested F among eligible
candidates; ties retain the earliest eligible candidate.
Candidate key (first eligible occurrence order;
configurations already shown in the history):
C[id]:  candidate=[configuration JSON]
Requested F summary (percentage points):
```

The candidate-key line repeats for each distinct eligible configuration. Under visit order, each summary row is `C[id]: [score].` or `C[id]: not disclosed.`; rows follow first-eligible order. Under organized comparison, disclosed rows have the form `C[id]: [score]; gap [signed difference].`, sorted by score ascending, followed by hidden rows. The organized message then says "Leader among disclosed candidates: C[id]." and "Each gap is the candidate's requested F minus the highest disclosed requested F." If any eligible scores remain hidden, the final line is "Some eligible requested F scores are not disclosed; the full eligible winner cannot be determined from disclosed scores." Otherwise it is "All eligible requested F scores are disclosed." The latter can also occur under partial coverage. Scores are rounded to one decimal and tied scores retain first-eligible order.

D.2 Model generation and classifier settings

The agents were DeepSeek-V4-Flash-0731 and GPT-5.6-Terra. Original searches and their continuations used high reasoning. DeepSeek's fixed-choice experiments and fresh score-label searches used the service-default reasoning setting with an 8,192-token output cap; Terra retained high reasoning. Temperature, top-p, top-k and sampling seed were unspecified and took service defaults.

Table 2 gives the 32 candidate configurations per task. Appendix A.2 specifies their preprocessing, fitting and scoring; Appendix A.1 gives the evaluation allocations and their seeds. Within each block, all configurations used the same training seed, and feedback conditions reused the same fitted candidates.