WHICH MESSAGES, AT WHAT BUDGET? A COMPLETE-SUBSET AUDIT OF LEARNED MESSAGE SHARING

Anonymous authors
Paper under double-blind review

ABSTRACT

In multi-agent scientific reasoning, a learned selector decides which of several proposed solutions reach the model that writes the final answer. When that choice leaves accuracy unmoved, accuracy alone cannot tell you why: the developer cannot see whether a winning combination was unavailable, available but not selected, or selected but not repeatable. We separate the three by recording what the answering model does under all 32 subsets of five fixed messages, across 975 scientific questions. The result is a diagnosis rather than a verdict. Winning combinations turn out to be plentiful, as some subset succeeds on 91.2% of biology questions against 70.9% for answering alone, yet neither of two independently seeded selectors converts that headroom into a significant gain on any of five cohorts. Within the selector family we evaluate, the shortfall is in selection rather than in what was available to select. The same tables show that a harmful message rarely forces a rebuild: 245 of 255 affected questions still admit an ordering that is optimal at every budget. But structure is easy to overstate, and our own controls temper it: four of five cohorts match a density-preserving null, and the one exception does not recur on fresh executions. This study therefore suggests a practical order of work: diagnose which component limits the pipeline before increasing its message budget. Absent combinations implicate the candidate generators or the answering model; available but unselected ones implicate the selection policy; unstable ones implicate re-execution variance and receiver alignment. Evaluation should indicate which component to improve next, rather than motivate more messages.

1 INTRODUCTION

A scientific question-answering system can ask several AI agents to propose a derivation, check units, or suggest an alternative explanation before a final model produces the answer. The resulting advice may be useful, redundant, or conflicting. A developer then controls two distinct quantities: how many of those messages are allowed through, and which ones. Only the second concerns the quality of what is shared.

Accuracy alone cannot tell you why. When a selection policy leaves accuracy unchanged, three different failures produce that same reading. No permitted combination would have succeeded. A successful combination existed and was not selected. Or a selected combination succeeded once and did not survive re-execution. Each implicates a different component — the candidate generators, the selection policy, or the stability of the answering model — and an aggregate accuracy number distinguishes none of them.

We study a single sharing stage in which five proposed solutions have already been generated. A *message* is one candidate contribution, the *receiver* is the model that answers the question, and the *selector* chooses which messages it receives. The receiver always sees the question and can answer without shared messages. A *message budget* limits how many contributions reach it. This setting isolates a practical decision inside multi-agent collaboration: given a pool of advice, which parts should the final answerer see? Because the pool is already generated, forwarding fewer messages does not recover its generation cost.

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.

We make these alternatives observable by evaluating every subset of the five messages: 32 separate receiver calls per question. Across 975 questions in five scientific benchmarks, this yields 31,200 outcomes. We train selectors on GPT-OSS mathematics outcomes and evaluate them with a Gemma receiver, testing receiver transfer on mathematics and combined receiver/domain transfer elsewhere. Two independently seeded runs of the primary subset scorer are evaluated on the same cohorts. Neither shows a significant gain over receiver-only answering, yet the complete tables reveal 6.5–20.3 percentage points of recorded best-subset headroom.

The audit then asks how the choices behave as the message budget grows. Adding a message can undo a correct answer. Among 255 questions with a harmful addition, 245 still admit some optimal growing path that matches the best recorded result at every size. We use density-preserving controls and fresh executions to distinguish recorded availability from special path structure and repeatable performance.

Earlier work studies information selection, message contributions, and harmful communication (Parasaram et al., 2024; Xue et al., 2022; Yang et al., 2026; Wynn et al., 2025); we bring the selected combination and its complete alternative set into the same evaluation.

Contributions. (i) A *complete-subset audit*: all 32 subsets of five messages on 975 scientific questions, yielding 31,200 recorded outcomes and making every unselected alternative observable. (ii) A three-part diagnostic separating whether a successful combination was *available*, whether the policy *selected* it, and whether that choice *repeated*, which resolves an unmoved accuracy number into the component where the shortfall appears. (iii) An application to trained selectors in which substantial recorded headroom coexists with no resolved accuracy gain, locating the limit in selection rather than availability. (iv) Density-preserving controls and fresh executions that separate structural findings from what the number of successful subsets alone produces.

2 Related Work

Information selection and multi-agent communication. Prior work already establishes that additional information can help or distract a language model. Fact selection for program repair studies combinations of contextual facts, finds that adding facts can reduce performance, and learns instance-specific selection (Parasaram et al., 2024). Shapley Message Value estimates each message's contribution across groups of messages and learns selective communication in cooperative multi-agent learning (Xue et al., 2022). Methods that remove communication links or adapt agent networks address which agents or connections to retain (Zhang et al., 2025; Liu et al., 2024). Our study connects learned selection to a complete table of receiver outcomes: which answers a policy rescues or harms, which successful alternatives it misses, and whether choices remain compatible across budgets.

Incremental selection across budgets. Incremental maximization compares a solution built by successive additions with a separately optimized solution of each size (Bernstein et al., 2017). This is the same general distinction that motivates our path comparison. Recent work on selecting groups of agents also compares greedy selection with exhaustive search using synthetic reward functions (Eslami, 2026). We instead enumerate binary correctness outcomes from a language-model receiver. Our empirical comparison measures the loss from retaining previously chosen messages, including tables where additions hurt. It concerns measured receiver outcomes rather than a new approximation guarantee. Selection order is also distinct from the order in which an already-selected context is presented (Lu et al., 2022; Liu et al., 2023).

Measuring the effects of collaboration. Wrong but Useful shows why the correctness of a proposed answer does not determine its downstream value, and why helpful and harmful message effects can cancel in aggregate accuracy (Yang et al., 2026). Multi-agent debate studies similarly document harmful correct-to-incorrect transitions (Wynn et al., 2025; Smit et al., 2024). Studies that measure how removing or retaining context changes an answer, including exhaustive document-subset comparisons, provide related measurement approaches (Cohen-Wang et al., 2024; Nematov et al., 2025). BANDMAS studies adaptive communication through replay-based packet valuation and resource-aware scheduling (Dong & Lin, 2026). Its replay-based valuation is related to our add-one diagnostics; our exhaustive tables additionally compare complete selections at each size. Building on

Table 1: **Measurement and learning scope.** Five cached messages give 32 subsets per question. Primary rows use the Gemma receiver; training rows use GPT-OSS. Pairwise models split the 1,154-question pool into 1,009 training and 145 development questions; the 19-question development set is for the subset scorer. Repeats and nested splits are not additional independent questions.

Population	Purpose	Questions	Subset outcomes
Omni-MATH-2	Mathematics evaluation	195	6,240
MaScQA	Materials-science evaluation	200	6,400
JEEBench	Entrance-exam evaluation	199	6,368
SciBench	College-science evaluation	199	6,368
LAB-Bench	Biology evaluation	182	5,824
Primary evaluation	Five expanded cohorts	975	31,200
Source training pool	Selector supervision	1,154	36,928
Source development	Subset-scorer selection	19	608
Source held-out	Separate source evaluation	58	1,856
Biology repeat	Paired second execution	38	1,216

these established message-effect analyses, our measurement object is a policy's selected combination at a receiver message-count budget, linked to every feasible alternative. This complete table exposes multi-message choices that singleton and leave-one-out comparisons cannot observe, supports the growing-path comparison, and enables density controls and fresh-execution tests of the resulting structure.

3 Experimental Setup

3.1 Data and Measurement Protocol

We generate the candidate messages once, then vary which ones reach the final-answer model. Five messages are produced using Diverse Hypothesis Deliberation, a procedure that prompts different reasoning roles to propose solutions (Yang et al., 2026). The selector reads the question and candidate messages and chooses a subset to forward. The receiver reads the question and that subset, then answers once. A separate model, the *evaluator*, compares its answer with the reference answer. The selector cannot see the reference or the receiver's recorded answers when making its choice. Since the candidate messages already exist, forwarding fewer of them does not recover their generation cost.

Every question is evaluated under all 32 subsets, including no messages. Each subset is tested in a separate receiver call with no memory of the other calls. Selected messages always appear in the same fixed relative order, called the canonical order. We write $U_g(S; q) \in \{0, 1\}$ for the recorded correctness of receiver g with subset S on question q, with one denoting a correct answer. A completed response with no extractable final answer is scored zero. A request left incomplete by an operational failure is missing, rather than wrong. A question is included only when all 32 subset outcomes are available; missing outcomes are not filled in. The record contains one outcome per subset, with temperature-zero decoding; repeated execution is examined separately. These 32 calls are an offline evaluation, not 32 turns of a conversation or a deployment-time requirement.

The primary evaluation uses five benchmark-specific groups of questions, called cohorts, totaling 975 questions (Table 1): Omni-MATH-2, a manually revised release of Omni-MATH whose items are checked for solvability and exact-answer verifiability (Ballon et al., 2026), MaScQA (Zaki et al., 2023), JEEBench (Arora et al., 2023), SciBench (Wang et al., 2024) and LAB-Bench (Laurent et al., 2024). The cached messages are generated by GPT-OSS-120B, which also serves as the evaluator without further training (OpenAI, 2025). The receiver is Gemma-4-31B-it (Gemma Team, 2026). Selector training instead uses mathematics outcomes from GPT-OSS as the receiver; we call this the source setting. The larger mathematics evaluation therefore changes the receiver, whereas the other four cohorts change both receiver and domain relative to training. The 195-question mathematics cohort contains the 58 source-held-out question IDs; it is not an additional disjoint test set. Appendix E.9 provides matched-receiver coverage and sampling details.

3.2 Selection Rules and Evaluation

An *optimal growing path* is an ordering that adds messages without removing earlier choices and matches the best recorded result at every budget. We compare three ways to represent choices across message budgets. A *global ordering* uses one ordering for every question in a cohort. A *question-specific ordering* chooses a separate ordering for each question but retains it across budgets. *Budget-specific selection* chooses any feasible subset independently for each question and budget.

To distinguish replacement from incremental selection, the structural comparison uses an exact message count: we write k for the number of messages shared, for $k = 1, \ldots, 5$, and message counts receive equal weight. For example, an ordering A, C, B, D, E gives the selected sets A, AC, ABC, ABCD, ABCDE as the budget increases.

We summarize a choice by how often its selected subsets succeed across the five budgets. The following quantities give the best recorded outcome at size k and the highest average outcome attainable by one path:

$$F_q(k) = \max_{|S|=k} U_g(S; q), \qquad Q_q = \max_{\pi} \frac{1}{5} \sum_{k=1}^{5} U_g(\pi_{1:k}; q). \tag{1}$$

Here $\pi_{1:k}$ is the subset containing the first k messages in ordering π. We enumerate all 120 orderings. A question has an *optimal growing path* when one ordering matches $F_q(k)$ at every size. The path loss is $\frac{1}{5} \sum_k F_q(k) - Q_q$. Every size contributes a nonnegative loss, so a zero total is equivalent to matching the best recorded result at every size. A path may match an unavoidable failure when no subset of that size succeeds. The global-order value optimizes the same average with one ordering shared across questions. All three comparisons choose using the already-observed outcomes. We call them hindsight comparisons; they describe available choices, not a model that can predict those choices.

The exactly-k comparison tests whether choices at different sizes fit one path. Learned selectors instead use a practical spending ceiling: *at most* b messages. They may choose the empty subset, so their accuracy is compared with receiver-only answering and other feasible policies, not subtracted from the exactly-k references above. We report accuracy, mean receiver-visible message count, and corrected answers (rescues) and newly incorrect answers (harms) relative to receiver-only answering. The communication metric is the count reaching the receiver, not the end-to-end cost of generating, selecting, and processing messages. All resampling uses questions as units because subsets of a question share the same task and message pool.

4 Learning and Evaluating Message Selection

4.1 Training Setup

A selector chooses messages before observing the receiver's answer. We compare a subset scorer trained to predict correctness, its untrained initialization, two pairwise selectors, and a message-blind control. All use Qwen3-4B-Instruct-2507 with low-rank adaptation (LoRA), which trains small weight updates (Qwen Team, 2025; Hu et al., 2022). Correctness-only pairwise supervision prefers successful subsets to failures; cost-aware supervision additionally prefers smaller successful subsets. Here "cost-aware" concerns message count, not total system cost. The message-blind control retains the question and subset size but removes the candidate texts.

Training uses source-receiver outcomes only. The subset scorer uses 1,154 training and 19 development questions; pairwise models split the 1,154-question pool into 1,009 training and 145 development questions. We evaluate two independently seeded subset-scorer runs, keeping one checkpoint fixed across all test cohorts for each run. The added run selects the lowest source-development loss reached before its training time limit. The runs have unequal training-search budgets and do not estimate seed variance under a matched recipe. Other configurations retain one run each. Appendix F.1 gives model-specific checkpoint rules and histories; Appendix A.1 gives the two-run results.

Each scorer evaluates all 32 subsets and chooses the highest-scoring feasible one without reference answers or target-receiver outcomes. Ties prefer fewer messages, then canonical subset index.

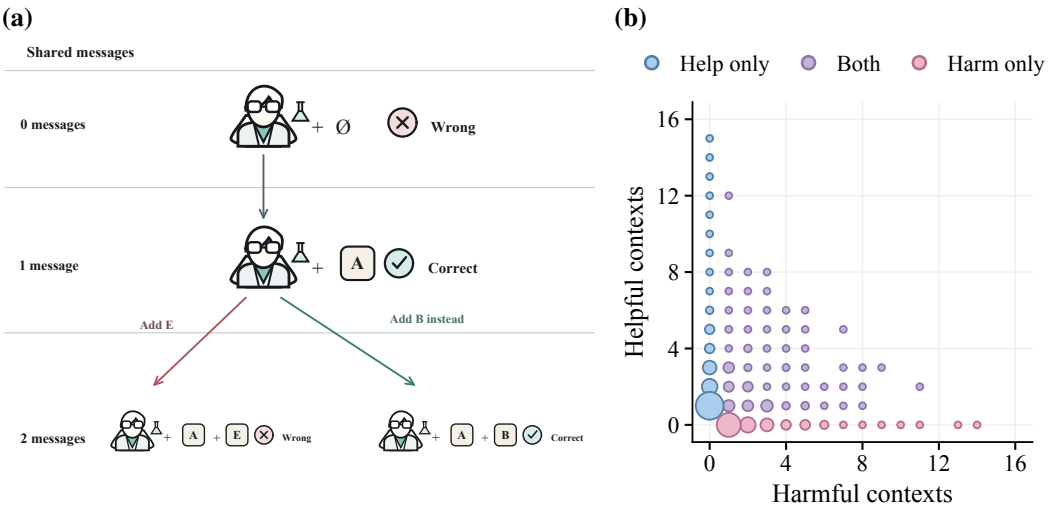

Figure 1: **Message combinations can rescue and then harm a receiver's answer.** (a) Recorded MaScQA thermodynamics example. A–E denote candidate messages, not answer options. Every condition includes the full question; the scientist icon denotes the receiver, and ∅ denotes no shared messages. Arrows compare independently replayed subsets, not successive turns of a conversation. A rescues the answer; adding E makes it wrong, whereas adding B preserves correctness. The real question, message excerpts and numerical answers appear in Appendix C.1. (b) All 975 primary questions, five messages and 16 contexts per message. Bubble area scales with pair count, with a minimum size for visibility; 3,508 no-effect pairs are omitted. Purple denotes the 471 pairs with both helpful and harmful contexts.

The budget is a ceiling: sharing nothing is allowed. Exhaustive scoring uses a pooled median of 32,248 input tokens per question (Appendix F.1); receiver-visible message count is therefore reported separately from selector computation.

4.2 Budget and Selection Performance

A larger allowance does not ensure that a policy uses more messages, or uses them well. Figure 2 follows the original checkpoints as the ceiling rises from one to five messages. The cost-aware policy stays near one message; other selectors generally send more. At budget five on LAB-Bench, receiver-only accuracy is 70.9%, the correctness-only policy reaches 71.4%, and the original subset scorer reaches 65.4%. The complete outcomes are in Appendix Table 5. Equal-step comparisons and a training-free message penalty examine whether reduced communication depends on the training setup (Appendices A.4 and F.2.5).

The added training run tests whether the original subset scorer's result was peculiar to one run. At most five messages, neither seed shows a significant advantage over receiver-only answering in any cohort: all ten question-paired 95% intervals include zero (Appendix Table 3). Point estimates range from -5.5 to $+0.5$ percentage points in the original run and from -7.7 to $+0.5$ in the added run. All 25 selector–cohort intervals in Appendix Table 7 (five selectors on five cohorts) also include zero. The original untrained scorer exceeds its trained counterpart on MaScQA (94.5% versus 93.0%) and SciBench (72.9% versus 72.4%). Appendix A.8 reports the prospective sensitivity of these comparisons to a three-point gain.

4.3 What Accuracy Leaves Unexplained

Rescues and harms show how a policy changes answers before those changes cancel. A rescue turns receiver-only failure into success; a harm turns success into failure. Their difference is exactly the net change in correct answers. On mathematics, the original subset scorer rescues eight questions and harms eight; the added run rescues eight and harms nine. Figure 3 presents the original configurations

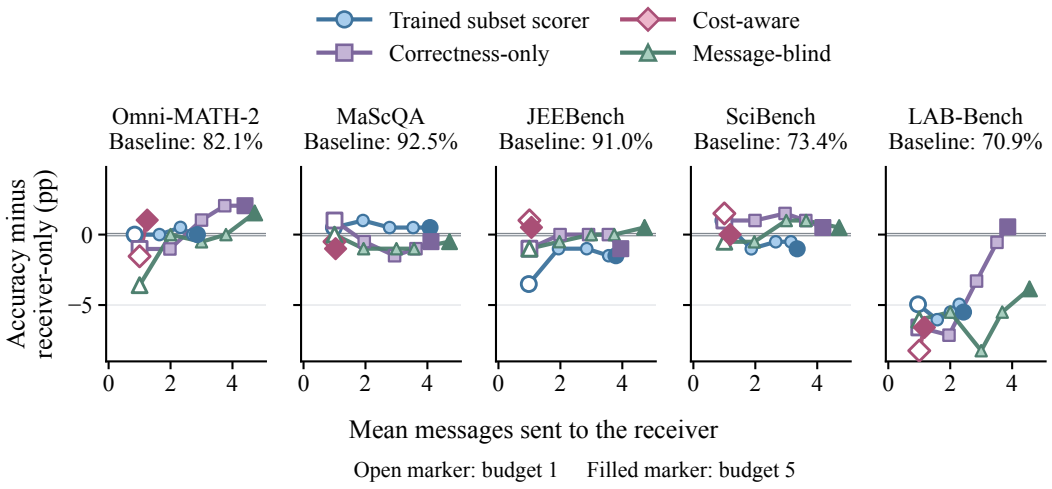

Figure 2: **More allowed messages produce different communication–accuracy trajectories.** Each point represents an at-most budget from one through five; the horizontal coordinate is the mean number of messages actually sent, and the vertical coordinate is accuracy relative to receiver-only answering. Lines connect budgets in increasing order; open markers indicate budget one and filled markers indicate budget five. The four displayed policies share a 4B backbone but use different objectives or decision rules. Points summarize fixed checkpoints and recorded receiver outcomes; paired uncertainty for the budget-five endpoints is reported in Appendix Table 7.



Figure 3: **Similar net accuracy can conceal different rescues, harms, and communication.** At most five messages, including none. Cohort sizes are 195, 200, 199, 199 and 182 in the displayed order; the plot abbreviates Omni-MATH-2 as Omni-MATH. Blue bars show rescues relative to receiver-only answering, pink bars show harms, and diamonds show their net accuracy change, all in percentage points of the full cohort. Numbers at right give mean messages forwarded. "Untrained" and "Subset trained" denote the untrained and trained subset scorers.

alongside their message counts. This separates preserving an already correct answer from finding a helpful alternative.

Different selections can also have identical correctness. On MaScQA, correctness-only and message-blind policies choose different subsets on 145 of 200 questions while sending 4.12 and 4.71 messages on average, and agree on correctness everywhere. Selection therefore changes substantially without moving the aggregate, a difference that accuracy does not record. Most of those differing choices

occur where all subsets have the same outcome, and agreement on the other 23 is consistent with a density-preserving control (Appendix A.1).

Ranking and selection are also distinct. The original trained scorer's success–failure pair concordance is 0.52–0.57 across cohorts, with each interval including the 0.5 chance reference (Appendix Figure 5b). It improves over initialization on some paired comparisons, but deployed accuracy depends on the highest-scored feasible subset, not on the average pair (Appendix Table 45). The receiver-only option is often ranked near the bottom despite its strong accuracy (Appendix Figure 6). We next examine the alternatives the policies leave unselected.

5 Diagnosing Message Selection

The audit turns an aggregate result into three development questions, each with a different remedy: was a successful combination *available* (§5.1), did the policy *select* it (§5.2–5.3), and did that choice *repeat* under another receiver or execution (§5.4)?

5.1 Availability: Were Successful Combinations Recorded?

The recorded alternatives leave room for improvement even when net selection gains are small. Appendix Figure 5 compares the primary trained subset scorer with receiver-only answering and the best recorded subset, allowing up to five messages including none. This *recorded headroom* is 6.5–20.3 percentage points across the full evaluation cohorts (Appendix Figure 5a). LAB-Bench has the largest gap. The best-subset column is a maximum over 32 single temperature-zero replays, so it is an optimistic reading of what was available rather than an expected-performance bound (Appendix B.4). Requiring success in both fresh executions rather than one does not remove the gap. Applying that stricter rule to both sides, so that the best subset and receiver-only answering must each succeed twice, the LAB-Bench headroom on the 75 repeated questions is 26.7 points (89.3% against 62.7%), against 25.3 and 21.3 in the individual runs; MaScQA gives 6.7 points (98.7% against 92.0%), against 5.3 and 6.7. The stricter rule lowers both terms, so the gap does not shrink. Availability is therefore not the binding constraint for the selectors evaluated here. Most of it needs only one message: of the 104 questions that some subset rescues relative to receiver-only answering, 90 are rescued by a single message (Appendix Table 51). On the 137 mathematics questions outside the held-out set, the correctness-only selector shows a $+4.23$-point budget-averaged difference from receiver-only answering that does not survive Holm correction across the fifteen model–population comparisons; we report it as an unconfirmed signal, not a gain (Appendix A.5).

5.2 Selection: Where Choices Change the Answer

Selection matters on a specific part of the recorded population. All 32 subsets succeed on 626 questions and all fail on 70; the remaining 279 have both outcomes. The first group tests whether a policy spends unnecessary communication, the second identifies failures that no recorded selection repairs, and the third permits an accuracy change through selection. Among the 279 mixed-outcome questions, 251 also contain differing outcomes between equal-size subsets. We call these *choice-informative*; they distinguish message identity from message count. Appendix B.2 gives this population breakdown.

A message's effect changes with the surrounding messages. We call those other visible messages its context. An add-one comparison holds that context fixed and measures the change after sharing one additional message. With five candidate messages, each message has 16 possible contexts: the subsets of the other four. Among the 975 primary questions, 255 have at least one harmful addition. A stronger property occurs on 138 questions: the *same* message helps in one context and harms in another. These are distinct counts. Across all 4,875 question–message pairs, 1,367 change correctness in at least one of their 16 contexts, and 471 of those change it in both directions (Figure 1b). Those 471 pairs fall on 138 questions, 3.4 per affected question on average, so reversal is a property of the question rather than of one message. The remaining 3,508 pairs preserve the recorded correct-or-incorrect outcome in every context.

A useful message can rescue an answer, yet another addition can undo that gain. In the MaScQA thermodynamics example in Figure 1a, the receiver alone gives -107.5 kJ, outside the reference

interval $[-168.7, -168.1]$. With message A it gives -168.66, but with A and E it gives $+168.7$. Adding B instead preserves a correct answer; a different pair, D and E, also succeeds. The example separates a harmful addition from a forced replacement: AE fails, but other pairs retaining A also succeed. Appendix C.1 gives the actual question, message excerpts and receiver answers; Appendix D discusses the stricter singleton–pair conflicts and a grading-sensitive case.

5.3 Selection Across Budgets

Harmful additions do not by themselves require replacing earlier choices. In the example above, AE fails but AB preserves A's success. The general question is whether some starting message and addition sequence can match the best recorded result at every size. Here "sequence" means the order in which messages join as the budget grows, which is distinct from their presentation order inside the prompt.

Such an optimal growing path exists for 241 of the 251 choice-informative questions (96.0%). Among the 255 questions with any harmful addition, 245 still admit an optimal growing path, including 141 that succeed at every size. The two populations differ: 251 questions are choice-informative, and 255 contain a harmful addition. Ten questions lack an optimal growing path under each count (Appendix B.2). Harmful branches can therefore coexist with an optimal growing path. Fitting one ordering per question recovers most of the recorded gain over a cohort-wide ordering, while allowing independent choices at every size adds 0.71–1.18 points (Appendix Table 15).

For a developer, path availability tests whether replacement is structurally necessary in the measured pool. We therefore examine both a density control and fresh executions before interpreting that availability.

5.4 Repeatability: Receiver Transfer and Fresh Executions

A policy's value is relative to the model receiving its messages. On the same 58 held-out mathematics questions, the original subset scorer attains 70.7% accuracy with either receiver, yet its gain over answering alone changes from $+3.45$ points with GPT-OSS to -5.17 with Gemma: the baselines are 67.2% and 75.9%. The added seed yields $+1.72$ and -5.17 points, respectively (Appendix Table 4). All four intervals include zero. The matched questions make the receiver comparison explicit, while the changing baseline shows why the signs alone do not establish a transfer penalty. Target-receiver training is not tested here.

A good path may be easy to find simply because many subsets already succeed. We call the proportion of successful subsets the success density. We test this explanation by shuffling outcomes within each question and subset size, preserving how many subsets succeed at every budget. We then recompute the best ordering for each shuffled table. Across 2,000 draws, the observed path loss lies within the central simulation range in four cohorts; biology shows a smaller loss than that range (Figure 4b). In four cohorts, this comparison does not distinguish the observed path loss from what the number of successful choices alone can produce. In the original sample, the LAB-Bench difference also appears within its largest task type, sequence reasoning (133 questions, exploratory $p = 0.001$; Appendix B.1). The smaller task groups do not show a statistically detectable difference, limiting generalization across biology tasks. Applying the same density control separately to two fresh executions of 75 questions per cohort gives no rejection in any of the four tests (Appendix B.1).

Frozen question-specific choices can beat a shared ordering on fresh executions without establishing that sharing beats answering alone. We froze historically fitted choices on 75 LAB-Bench and 75 MaScQA questions before collecting two new executions (9,600 outcomes). Averaging at-most-1–5 budgets, the advantage over the frozen global order is 6.93 percentage points on LAB-Bench (95% question-bootstrap interval $[3.20, 10.93]$), down from 11.20 fitted points; MaScQA gives 0.80 points $[0.00, 2.13]$ (Figure 4c). Relative to receiver-only answering, however, the same choices yield 3.87 points $[-4.93, 12.80]$ and 0.93 points $[-3.73, 5.60]$, respectively. Neither interval excludes zero. Appendix B.3 gives all three paired contrasts and the earlier 38-question study. Receiver-only rescue availability persists for 74–80% of eligible questions across the three admissible repeat sets, yet the union across two runs is larger than either run alone. Availability is therefore partly execution-specific (Appendix B.4).

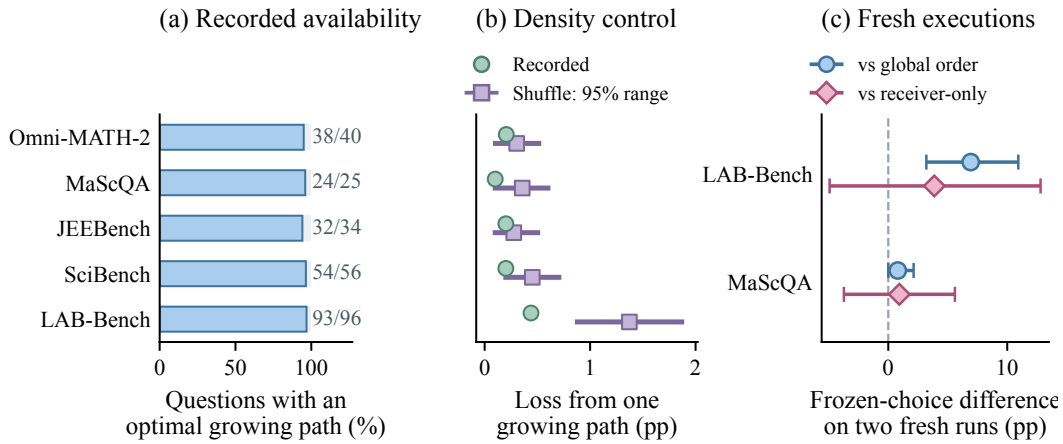

Figure 4: **Path availability, density controls, and fresh-execution performance.** (a) Among questions where equal-size subsets differ, 241/251 admit one growing path that matches the best recorded result at every exact budget. (b) Recorded path loss and the central 95% range of 2,000 within-question, within-size shuffles on the original cohorts; ranges are simulation ranges, not confidence intervals. Only LAB-Bench lies outside its range. (c) Historically fitted question-specific choices evaluated on two fresh executions of 75 questions per cohort. Circles compare with a frozen global order; diamonds compare with receiver-only answering. Bars are 95% question-bootstrap intervals. Thus availability in one table, separation from a density control, and benefit on a fresh execution are distinct claims.

6 Discussion and Conclusion

Complete-subset measurement turns an unmoved accuracy number into a diagnosis. Across five cohorts, some subset of the five available messages succeeds far more often than answering alone — on biology, 91.2% against 70.9% — yet neither independently seeded selector achieves a resolved gain over that baseline, and all twenty-five selector–cohort intervals, five selectors across five cohorts, include zero. Read together, those two results locate the failure rather than merely recording it. Within the selector family we evaluate, the combinations that would have helped were recorded but not converted into a resolved gain, so availability is not the binding constraint. Rescues and harms show the mechanism: on mathematics the primary scorer repairs eight answers and breaks eight, which is how substantially different selections leave aggregate accuracy nearly unchanged. An evaluation reporting only that aggregate would record no effect and give no account of its origin.

These findings suggest a practical order of work for multi-agent communication: diagnose which component limits the pipeline before increasing its message budget. Where no permitted combination succeeds, the limit lies upstream of selection, and the candidate generators or the answering model deserve the effort, since no policy can recover an answer that no combination produces. Where successful combinations exist but go unselected — the regime our experiments occupy — the limit is the selection policy and its fit to the receiver it serves, and a larger message budget does not address it. Where selected combinations succeed once but not under re-execution or a different answering model, the limit is stability, and execution variance should be characterised before a gain is credited. We therefore suggest that evaluations of learned communication report availability, selection and repeatability alongside accuracy, and that message sharing be treated as a diagnostically gated decision rather than a capacity to be scaled.

These conclusions are bounded by the measurement that produced them. The evidence covers fixed five-message pools, two primary-scorer seeds with unequal training budgets, single runs of the remaining configurations, and one main transfer receiver. It does not separate training quality from receiver shift, nor independently regrade the primary outcomes. Exhaustive enumeration costs 2^M receiver calls and suits small candidate pools; larger systems can keep the same three comparisons through targeted subset tests. Within those limits, the audit identifies which component to examine next.

AI USE STATEMENT

Writing and presentation. Generative AI tools assisted with brainstorming, revising prose, checking terminology and cross-references, and preparing figures and tables. The authors reviewed the resulting text and visualizations, approved the final presentation, and take responsibility for every claim in the paper.

Literature discovery. Generative AI tools assisted with locating potentially relevant papers and organizing comparisons with related work. Citations and the claims attributed to them were checked against the cited sources rather than accepted from generated summaries.

Code and analysis support. Generative AI tools assisted with writing, debugging, and reviewing code used for data organization, analysis, and plotting. The authors specified the research questions, experimental comparisons, statistical quantities, and acceptance checks, and reviewed the reported outputs against the retained records. Language models used to generate candidate messages, select subsets, answer questions, or evaluate answers are components of the experiments described in the paper; their outputs are part of the measured data rather than author-written evidence. The authors determined the methodology, interpretation, and conclusions.

ETHICS STATEMENT

This study evaluates language-model communication on scientific question-answering benchmarks. It does not recruit or interact with human participants and does not analyze personal or sensitive user data. The principal practical risk is misinterpreting generated scientific advice as reliable: the recorded examples show that an additional message can both repair and disrupt an answer. The results therefore support evaluation and diagnosis of message-sharing policies, not unsupervised deployment for consequential scientific decisions. The analysis archive is anonymized for double-blind review and excludes endpoint credentials and restricted or non-redistributable source content; Appendix D states the resulting limits on independent regrading and retraining.

REPRODUCIBILITY STATEMENT

The paper reports the receiver, message-generation, selector, and evaluator roles; complete-subset construction; selection rules; estimands; and question-level uncertainty procedures. Appendices A–G provide full robustness analyses, model and checkpoint information, training histories, prompts, evaluation conventions, and detailed results. The supplementary folder includes an anonymized analysis archive and code for reconstructing the documented tables, figures, and statistical comparisons from the included outcomes and scores. Appendix D describes its provenance fields, claim-to-file map, and reproduction scope. The archive also supports new comparisons over the included outcomes and scores, but it does not support independent regrading, retraining, or new receiver calls without the omitted benchmark text, messages, raw responses, and model weights. Repeated-execution results are reported separately from re-exported records, and all stochastic analyses state their sampling unit and random seed.

REFERENCES

Daman Arora, Himanshu Singh, and Mausam. Have LLMs advanced enough? a challenging problem solving benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7527–7543, 2023. URL https://aclanthology.org/2023.emnlp-main.468/.

Marthe Ballon, Andres Algaba, Brecht Verbeken, and Vincent Ginis. Benchmarks saturate when the model gets smarter than the judge, 2026. URL https://arxiv.org/abs/2601.19532.

Aaron Bernstein, Yann Disser, and Martin Groß. General bounds for incremental maximization, 2017. URL https://arxiv.org/abs/1705.10253.

Benjamin Cohen-Wang, Harshay Shah, Kristian Georgiev, and Aleksander Madry. Contextcite: Attributing model generation to context. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2409.00729.

Jiangwen Dong and Wanyu Lin. BANDMAS: Causality-inspired semantic packet scheduling for bandwidth-efficient multi-agent collaboration, 2026. URL https://arxiv.org/abs/2608.00458.

Mojtaba Eslami. Dynamic coalition formation and communication pricing in skill-based agentic ai systems, 2026. URL https://arxiv.org/abs/2608.07532.

Gemma Team. Gemma 4 technical report, 2026. URL https://arxiv.org/abs/2607.02770.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Jon M. Laurent, Joseph D. Janizek, Michael Ruzo, Michaela M. Hinks, Michael J. Hammerling, Siddharth Narayanan, Manvitha Ponnapati, Andrew D. White, and Samuel G. Rodriques. LAB-Bench: Measuring capabilities of language models for biology research, 2024. URL https://arxiv.org/abs/2407.10362.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 2023. URL https://arxiv.org/abs/2307.03172.

Zijun Liu, Yanzhe Zhang, Peng Li, Yang Liu, and Diyi Yang. A dynamic LLM-powered agent network for task-oriented agent collaboration. In *Conference on Language Modeling (COLM)*, 2024. URL https://arxiv.org/abs/2310.02170.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 2022. URL https://arxiv.org/abs/2104.08786.

Ikhtiyor Nematov, Tarik Kalai, Elizaveta Kuzmenko, Gabriele Fugagnoli, Dimitris Sacharidis, Katja Hose, and Tomer Sagi. Source attribution in retrieval-augmented generation, 2025. URL https://arxiv.org/abs/2507.04480.

OpenAI. gpt-oss-120b and gpt-oss-20b model card, August 2025. URL https://openai.com/index/gpt-oss-model-card/.

Nikhil Parasaram, Huijie Yan, Boyu Yang, Zineb Flahy, Abriele Qudsi, Damian Ziaber, Earl T. Barr, and Sergey Mechtaev. The fact selection problem in llm-based program repair, 2024. URL https://arxiv.org/abs/2404.05520.

Qwen Team. Qwen3-4B-Instruct-2507 model card, 2025. URL https://huggingface.co/Qwen/Qwen3-4B-Instruct-2507.

Andries Smit, Paul Duckworth, Nathan Grinsztajn, Thomas D. Barrett, and Arnu Pretorius. Should we be going MAD? a look at multi-agent debate strategies for LLMs, 2024. URL https://arxiv.org/abs/2311.17371.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R. Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. SciBench: Evaluating college-level scientific problem-solving abilities of large language models. In *International Conference on Machine Learning (ICML)*, 2024. URL https://arxiv.org/abs/2307.10635.

Andrea Wynn, Harsh Satija, and Gillian Hadfield. Talk isn't always cheap: Understanding failure modes in multi-agent debate, 2025. URL https://arxiv.org/abs/2509.05396.

Di Xue, Lei Yuan, Zongzhang Zhang, and Yang Yu. Efficient multi-agent communication via shapley message value. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pp. 578–584, 2022. doi: 10.24963/ijcai.2022/82. URL https://www.ijcai.org/proceedings/2022/0082.pdf.

Chih-Hsuan Yang, Anjir Ahmed Chowdhury, Cheng-Hau Yang, Weijian Zheng, Fernando Llorente, Xiaolong Ma, Xinyang Li, Eliu A. Huerta, Ian T. Foster, and Rajeev Thakur. Wrong but useful: Trajectory value beyond answer correctness in multi-agent messages, 2026. URL https://arxiv.org/abs/2608.14375.

Mohd Zaki, Jayadeva, Mausam, and N. M. Anoop Krishnan. MaScQA: A question answering dataset for investigating materials science knowledge of large language models, 2023. URL https://arxiv.org/abs/2308.09115.

Guibin Zhang, Yanwei Yue, Zhixun Li, Sukwon Yun, Guancheng Wan, Kun Wang, Dawei Cheng, Jeffrey Xu Yu, and Tianlong Chen. Cut the crap: An economical communication pipeline for LLM-based multi-agent systems. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=LkzuPorQ5L.

GUIDE TO THE APPENDIX

The appendix begins with evidence most directly bearing on the main conclusions: two-run learned-selection results, receiver comparisons, path robustness, and recorded examples. Detailed measurements, training settings and protocol definitions follow. Table 2 provides direct links to each part.

Table 2: **Appendix evidence guide.** Sections are ordered from the main claims and their checks to the supporting technical details.

Section	Evidence and purpose
A. Learned selection	Two-run results, matched receivers, paired accuracy intervals, equal-step comparisons and receiver-only ranking.
B. Path robustness	Density controls, frozen-choice repeats, outcome agreement, grading checks, and noise sensitivity.
C. Recorded examples	The real question, messages and receiver answers behind Figure 1, plus further cases.
D. Data provenance	Analysis-archive contents, model attribution and grading caveats.
E. Full measurement results	Cohort-level message effects, sampling, attribution and matched receiver comparisons.
F. Training and full results	Objectives, checkpoint histories, ranking diagnostics, communication controls and all-budget tables.
G. Protocol definitions	Outcome, transition and cost conventions used throughout the analyses.

A UNCERTAINTY AND BEHAVIOR OF LEARNED SELECTORS

We first report the two primary-scorer runs and the matched-receiver comparison, followed by the full at-most-five results underlying main-text Figure 2. The uncertainty tables complement the

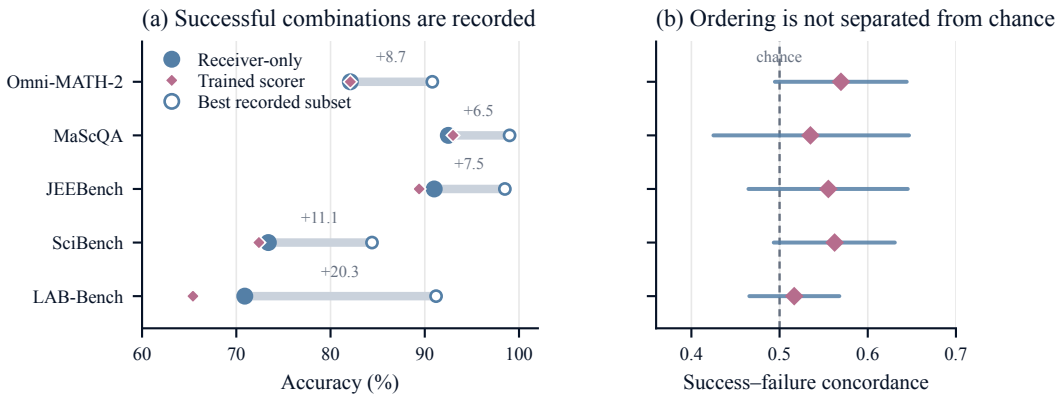

Figure 5: **Successful combinations are recorded, but the scorer does not order them above failures.** (a) At most five messages: receiver-only accuracy, the trained subset scorer, and the best recorded subset on each cohort; the bar spans the recorded headroom, labelled in percentage points. The best-subset column uses outcomes from this execution and is not attainable by a deployed policy. (b) Success–failure concordance for the trained scorer with 95% question-bootstrap intervals; 0.5 is the chance reference. Every interval includes chance, so the ordering quality needed to capture the headroom in (a) is not resolved. Panel (a) plots Table 6 and panel (b) the trained rows of Table 45.

rescue/harm decomposition in main-text Figure 3. Figure 6 describes the subset scorer's preference for sharing messages; rank one is its highest-scored option. Training settings and the complete budget breakdowns appear in Appendix F.

A.1 Two Independently Seeded Subset Scorers

The primary subset scorer is evaluated with seeds 20260908 (original) and 20260925 (added), using the same five held-out Gemma cohorts and one frozen checkpoint per run. Table 3 reports question-paired intervals for each run. They quantify sampling uncertainty conditional on each checkpoint, not training-seed variance. Neither run has an interval wholly above zero in any cohort. The added run uses fewer messages in every cohort, but its net accuracy change is not consistently better.

Table 3: **Two-run primary-scorer comparison.** At most five messages. Accuracy differences from receiver-only answering and 95% paired-question bootstrap intervals are in percentage points (10,000 draws). Rescue/harm entries are question counts; the final column is mean forwarded messages. Cohort sizes are 195, 200, 199, 199 and 182.

Cohort	Run	Difference	95% interval	Rescue/harm	Messages
Omni-MATH-2	Original	0.00	$[-4.10, +4.10]$	8/8	2.87
	Added	-0.51	$[-4.62, +3.59]$	8/9	2.14
MaScQA	Original	$+0.50$	$[-3.50, +4.50]$	8/7	4.08
	Added	0.00	$[-4.00, +4.00]$	8/8	3.05
JEEBench	Original	-1.51	$[-6.03, +3.02]$	9/12	3.79
	Added	$+0.50$	$[-4.02, +5.03]$	10/9	3.10
SciBench	Original	-1.01	$[-6.03, +3.52]$	11/13	3.35
	Added	-2.51	$[-8.04, +2.51]$	12/17	2.60
LAB-Bench	Original	-5.49	$[-12.64, +1.65]$	16/26	2.43
	Added	-7.69	$[-15.38, 0.00]$	18/32	2.14

The added run reaches 11 development evaluations before a 5.5-hour training limit and selects update 2,750, the lowest recorded development BCE reached (0.50097). Its adapter is frozen before test scoring. The original evaluated checkpoint is also update 2,750, but the runs have different search budgets; equal selected steps do not make them matched training replicates. The original

run completes its search through update 5,250 and selects development BCE 0.42764; Appendix F.1 describes both selection conditions. Two runs test whether the observed result recurs, not the full distribution of outcomes over seeds.

Table 4: **The same 58 mathematics questions under two receivers.** Both primary-scorer runs use their fixed predictions and complete outcome tables for each receiver. Differences and question-bootstrap intervals are relative to each receiver's own receiver-only baseline, in percentage points; these are not intervals for the between-receiver difference.

Run	Receiver	Accuracy (%)	Difference	95% interval
Original	GPT-OSS	70.69	$+3.45$	$[-3.45, +10.34]$
Original	Gemma	70.69	-5.17	$[-15.52, +5.17]$
Added	GPT-OSS	68.97	$+1.72$	$[-5.17, +8.62]$
Added	Gemma	70.69	-5.17	$[-15.52, +5.17]$

The receiver-only baselines are 67.24% for GPT-OSS and 75.86% for Gemma. Holding questions and selector fixed removes cross-cohort composition differences, but the signs of these relative gains do not by themselves identify a transfer bottleneck. Neither run is trained on Gemma outcomes. The pairwise models are separate configurations, reported in Appendix A.6.

Conditioning on questions whose outcomes vary. Across the 279 questions whose 32 recorded outcomes are not constant, all 20 selector–cohort intervals against receiver-only answering include zero. The 696 outcome-constant questions contribute exactly zero to every paired correctness difference, so the full-sample difference equals $n_{\text{mixed}}/n_{\text{full}}$ times the conditional difference in every cell. Conditioning therefore rescales the same evidence rather than creating an independent comparison. A separate comparison of correctness-only and message-blind selection shows why accuracy and communication must be reported separately: content-aware selection sends fewer messages in every cohort, but has equal accuracy on MaScQA and SciBench and lower accuracy on JEEBench. On solvable questions, neither policy attains the "minimal" category because the smallest successful choice is usually the empty subset, which neither policy selects at budget five.

Choice agreement under a density control. Of the 145 differing MaScQA choices, 122 occur on outcome-constant questions; their correctness agreement follows by construction. For the remaining 23, we hold both policies' choices fixed and shuffle outcomes within each question and subset size. Across 10,000 draws (seed 20260925), the expected number of agreements is 21.20, with a central 90% range of 19–23. All 23 agree in the observed data; the one-sided Monte Carlo tail probability is 0.1081. Thus the observed agreement is not distinguishable from this density-preserving reference.

Failed selections among solvable questions. On the 166 LAB-Bench questions with at least one successful recorded subset, correctness-only selection fails 36 times and message-blind selection fails 44 times. The paired comparison has 18 correctness-only wins, 10 losses and 138 ties; the failed-selection-rate difference is -4.82 points with a 95% question-bootstrap interval of $[-10.84, +1.20]$. On this fixed population, that difference is exactly the negative of the accuracy difference. It is a diagnostic restatement, not an independent finding or a demonstrated content advantage.

A.2 Uncertainty in Learned Selection

Table 7 gives the uncertainty for the primary comparison: each evaluated scorer versus the receiver answering without shared messages. All 25 intervals include zero. They quantify variation across questions conditional on these checkpoints and recorded answers; they do not estimate variability from retraining or repeated receiver calls.

The larger diagnostic sweep also includes other budgets and configurations. Its 310 reported cells reuse questions, budget rules and overlapping 172- and 182-question LAB-Bench views. None has an interval wholly above zero; 45 have intervals wholly below zero, all in LAB-Bench (29 for the 172-question view and 16 for the 182-question view). These are overlapping diagnostic counts, not 310 independent experiments or the uncertainty summary for Table 5.

Table 5: **Accuracy and communication at most five messages.** Five is a ceiling: choosing fewer messages, including none, is allowed. Each entry is accuracy (%), with mean forwarded messages in parentheses. Selectors trained on GPT-OSS mathematics are evaluated with Gemma; the untrained scorer is a reference. Message counts exclude generation and selector computation. These are single-checkpoint estimates; all 25 scorer-versus-receiver-only paired 95% intervals include zero (Appendix Table 7).

Selector	Omni-MATH-2	MaScQA	JEEBench	SciBench	LAB-Bench
Untrained subset scorer	80.5 (2.53)	94.5 (3.42)	88.4 (2.90)	72.9 (2.68)	63.7 (2.41)
Trained subset scorer	82.1 (2.87)	93.0 (4.08)	89.4 (3.79)	72.4 (3.35)	65.4 (2.43)
Correctness-only	84.1 (4.41)	92.0 (4.12)	89.9 (3.95)	73.9 (4.18)	71.4 (3.86)
Cost-aware	83.1 (1.25)	91.5 (1.03)	91.5 (1.07)	73.4 (1.19)	64.3 (1.16)
Message-blind	83.6 (4.72)	92.0 (4.71)	91.5 (4.72)	73.9 (4.71)	67.0 (4.56)
Receiver only	82.1 (0.00)	92.5 (0.00)	91.0 (0.00)	73.4 (0.00)	70.9 (0.00)
Reveal-all	84.1 (5.00)	92.0 (5.00)	91.5 (5.00)	73.9 (5.00)	67.0 (5.00)

Table 6: **Recorded headroom and the primary trained scorer.** At most five messages on the same five evaluation cohorts; accuracies are percentages. Headroom is best recorded subset minus receiver only, in percentage points (pp). The oracle uses outcomes from this execution; the scorer does not. Its paired differences from receiver only are in Appendix Table 7. The 195-question mathematics cohort contains the 58 source-held-out questions and is not an additional disjoint test set (Appendix A.5).

Cohort	Receiver only	Best subset (oracle)	Headroom (pp)	Trained scorer
Omni-MATH-2	82.1	90.8	8.7	82.1
MaScQA	92.5	99.0	6.5	93.0
JEEBench	91.0	98.5	7.5	89.4
SciBench	73.4	84.4	11.1	72.4
LAB-Bench	70.9	91.2	20.3	65.4

A.3 The Position of the Receiver-Only Option

The subset scorer rarely selects the empty subset on the four expanded transfer domains: $1/200$ materials-QA, $2/199$ engineering, $4/199$ college-science and $6/182$ biology questions at budget one. The option remains available. This is selection of auxiliary information, not abstention from answering. Its rank among all 32 subsets and its score margin against the best singleton describe distinct properties.

Receiver-only answering is correct on 70.9–92.5% of expanded questions, but failure to select it does not necessarily forfeit accuracy: a non-empty subset often also succeeds. Only 4, 2, 2, 1 and 2 questions in biology, mathematics, college science, engineering and materials QA, respectively, succeed exclusively with no messages. We therefore evaluate communication and outcome losses separately.

A.4 Equal-Step Checkpoint Comparison

Comparing both pairwise models at step 125 removes the difference in optimizer-step count for this contrast. It does not equalize supervision content, token exposure, or training-seed variation. We average each question's accuracy over at-most budgets one through five, including the empty subset, before taking paired differences. This differs from the budget-five summary in Table 5.

All five accuracy intervals include zero. The cost-aware checkpoint forwards fewer messages in each cohort, but unresolved accuracy differences do not establish equivalence or rule out an effect of training duration. Within the correctness-only model, changing from step 125 to 375 gives point

Table 7: **Primary selector accuracy differences from receiver-only answering.** At most five messages, the same five Gemma cohorts as Table 5. Entries are paired differences in percentage points, with 95% percentile intervals below (20,000 question-bootstrap draws, seed 20260924). Each interval concerns one fixed checkpoint and recorded execution. All include zero; this does not establish equal performance. These unadjusted intervals differ from the model-versus-model tests later in this appendix. The 195-question mathematics cohort contains the 58 source-held-out questions and is not an additional disjoint test set (Appendix A.5).

Selector	Omni-MATH-2	MaScQA	JEEBench	SciBench	LAB-Bench
Untrained subset scorer	-1.5 $[-6.7, +3.6]$	$+2.0$ $[-2.0, +6.0]$	-2.5 $[-7.5, +2.5]$	-0.5 $[-5.0, +4.0]$	-7.1 $[-14.8, +0.5]$
Trained subset scorer	$+0.0$ $[-4.1, +4.1]$	$+0.5$ $[-3.5, +4.5]$	-1.5 $[-6.0, +3.0]$	-1.0 $[-6.0, +4.0]$	-5.5 $[-12.6, +1.6]$
Correctness-only	$+2.1$ $[-3.1, +7.2]$	-0.5 $[-4.5, +3.5]$	-1.0 $[-5.5, +3.5]$	$+0.5$ $[-4.0, +5.0]$	$+0.5$ $[-6.6, +7.7]$
Cost-aware	$+1.0$ $[-4.1, +6.2]$	-1.0 $[-4.0, +2.0]$	$+0.5$ $[-4.0, +5.0]$	$+0.0$ $[-5.0, +5.0]$	-6.6 $[-13.7, +0.5]$
Message-blind	$+1.5$ $[-3.6, +6.7]$	-0.5 $[-4.5, +3.5]$	$+0.5$ $[-4.0, +5.0]$	$+0.5$ $[-4.0, +5.0]$	-3.8 $[-11.0, +3.3]$

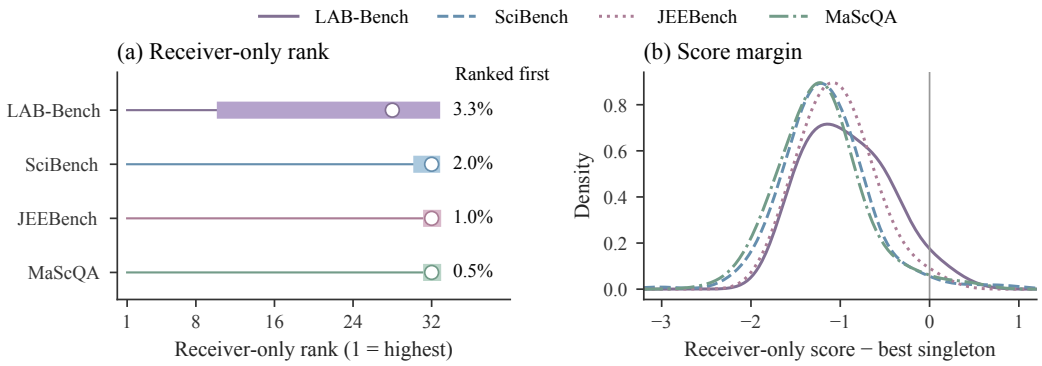

Figure 6: **The subset scorer usually ranks sharing above receiver-only answering.** Four independently reconstructed transfer cohorts: biology, college science, engineering, and materials QA (182, 199, 199, and 200 questions). Panel (a) shows the range, quartiles, and median receiver-only rank among all 32 subsets, plus the share ranked first. Panel (b) shows the score margin between receiver-only answering and the highest-scoring singleton. Scores describe preference; rescues and harms determine whether that preference helps.

differences from -0.20 to $+1.65$ points, again with intervals crossing zero. These comparisons characterize the evaluated checkpoints rather than isolate a causal effect of the training objective.

Completing the two-step comparison. Scoring both objectives at update 375 completes the objective-by-step comparison (Table 9). The cost-aware checkpoint at 375 chooses at most one message on nearly all questions in each cohort. Its lower communication does not imply consistent accuracy improvement: the cost-aware-minus-correctness-only contrast is negative on three cohorts and positive on two. The two pointwise intervals excluding zero are negative. These comparisons hold optimizer steps fixed, not training pairs, token exposure or realized communication. A concentration on small subsets alone does not prove that the policy ignores question content.

A.5 Separating the Mathematics Evaluation Subsets

The 195-question Gemma cohort contains the 58 source-held-out question IDs and 137 additional questions. Identity checks find no overlap between these 137 questions and the selector's training or development manifests. This establishes that they were unseen by selector fitting and selection,

Table 8: **Cost-aware minus correctness-only at step 125.** Accuracy differences and question-bootstrap 95% intervals are in percentage points (10,000 resamples). The final two columns give mean forwarded messages, averaged over budgets. One training seed.

	Accuracy difference		Messages	
Cohort	Estimate	95% interval	Cost-aware	Correctness-only
Omni-MATH-2	-0.10	$[-2.77, +2.56]$	1.20	2.74
MaScQA	-0.40	$[-2.70, +1.80]$	1.03	2.77
JEEBench	$+2.01$	$[-0.80, +4.92]$	1.05	2.73
SciBench	-0.90	$[-4.42, +2.41]$	1.15	2.79
LAB-Bench	-1.87	$[-6.81, +2.97]$	1.13	2.59

Table 9: **Cost-aware minus correctness-only at step 375.** Same budget averaging and paired bootstrap as Table 8; intervals are pointwise, not multiplicity-adjusted. Message columns report means over budgets 1–5.

	Accuracy difference (pp)		Messages	
Cohort	Estimate	95% interval	Cost-aware	Correctness-only
Omni-MATH-2	-3.18	$[-6.46, -0.10]$	0.69	2.83
MaScQA	$+2.00$	$[-0.00, +4.30]$	0.95	2.73
JEEBench	-0.60	$[-3.22, +2.01]$	0.98	2.68
SciBench	$+1.51$	$[-1.21, +4.32]$	0.82	2.75
LAB-Bench	-7.69	$[-13.08, -2.31]$	1.00	2.64

not by pretraining of the underlying language models. Receiver-only accuracy is 75.86% on the 58-question subset and 84.67% on the additional 137; these populations are not interchangeable.

On the 137 additional questions, the correctness-only checkpoint has a $+4.23$-point mean accuracy difference from receiver-only answering, averaged over at-most budgets one through five. Its paired percentile interval is $[0.44, 8.47]$ points. A question-level sign-flip test gives $p = 0.0414$ under the symmetry assumption for paired differences, but $p = 0.6210$ after Holm correction across the 15 model–population comparisons (five models on the 58-, 137-, and 195-question populations). We treat this as an unconfirmed positive signal, not evidence of no effect or a confirmed general accuracy gain. These budget-averaged comparisons do not replace the budget-five intervals in Table 7.

A.6 Matched Questions Across Receivers

A source-receiver comparison separates evaluation on the training receiver from evaluation on another receiver. We use the same 58 held-out mathematics questions, candidate messages and selector scores with recorded GPT-OSS and Gemma outcomes. The correctness-only checkpoints at steps 125 and 375 and the cost-aware checkpoint at step 125 have identified adapter checksums. Selector predictions are newly computed; receiver outcomes are retained executions, not new receiver calls.

Relative to each receiver's own baseline, all three checkpoint point estimates are more favorable on GPT-OSS, with all six intervals including zero. A paired comparison now differences each checkpoint's gain over its own receiver-only baseline across the two receivers. For correctness-only step 375, cost-aware step 125 and correctness-only step 125, the differences are respectively $+11.03$ $[-1.38, 24.48]$, $+3.45$ $[-10.00, 17.59]$ and $+3.45$ $[-8.28, 15.52]$ percentage points (95% question-bootstrap intervals, 10,000 draws). All include zero. Each comparison includes all 58 questions, of which 18, 14 and 19 respectively contribute nonzero differences. The source receiver also has more room for improvement: eight questions are correct only under receiver-only Gemma, versus three correct only under receiver-only GPT-OSS. These results do not isolate a transfer penalty from baseline headroom. The checkpoints share training data and include two steps of one model, so they do not estimate training-seed variability.

Table 10: **Pairwise selection on the same 58 questions with two receivers.** Entries are accuracy differences from each receiver's own receiver-only baseline, in percentage points, with 95% question-bootstrap intervals (10,000 resamples). At-most budgets 1–5 are equally weighted. Fixed selector scores choose the same subsets for both receivers; the final column is mean messages. Receiver-only accuracy is 67.24% for GPT-OSS and 75.86% for Gemma.

Checkpoint	GPT-OSS	Gemma	Messages
Correctness-only, 375	$+3.10$ $[-3.45, +10.00]$	-7.93 $[-19.31, +3.10]$	2.88
Cost-aware, 125	-1.03 $[-9.66, +7.59]$	-4.48 $[-15.86, +6.55]$	1.30
Correctness-only, 125	-2.07 $[-10.00, +5.52]$	-5.52 $[-16.21, +4.83]$	2.70

A.7 A Source-Fitted Message-Slot Baseline

A simple baseline ranks the five message slots by their source-training Shapley values: coalition-averaged marginal correctness contributions computed on the 1,154 training questions. It shares positive-valued slots in descending order, truncated to budget, with no target-outcome tuning. All five fitted values are positive, so it always fills the budget (three messages on average over budgets 1–5). This is a fixed slot ranking, not a question-conditioned communication policy or a reproduction of a learned multi-agent method.

Table 11: **Source-fitted slot ranking.** Mean accuracy and difference from receiver-only answering over at-most budgets 1–5. Intervals are paired-question 95% bootstrap intervals. No interval excludes zero.

Cohort	Accuracy (%)	Difference (pp)	95% interval
Omni-MATH-2	83.28	$+1.23$	$[-3.18, 5.64]$
MaScQA	92.20	-0.30	$[-3.90, 3.30]$
JEEBench	91.26	$+0.30$	$[-3.52, 4.32]$
SciBench	73.87	$+0.50$	$[-3.82, 5.03]$
LAB-Bench	65.05	-5.82	$[-12.09, 0.33]$

A.8 Sensitivity to Small Accuracy Gains

Unresolved differences need not imply that useful small gains are absent. We prospectively calculate the power of a two-sided exact paired McNemar test at level 0.05 for a hypothetical three-percentage-point gain. Each calculation holds the cohort size and its empirical probability of disagreement between policies fixed; it does not estimate power from the observed effect. Table 12 distinguishes receiver-only and initialization references. At a fixed gain and sample size, more discordance creates greater paired-difference variability. Both sample size and disagreement therefore matter. These conditional calculations do not change the primary evaluation or establish equivalence.

A.9 Communication Counts and System Cost

Fewer forwarded messages do not by themselves establish lower total cost. The delivered accounting separates candidate generation, subset scoring, receiver answering and evaluation by model and role. Enumerating 32 scored subsets incurs selector work even when the chosen subset is empty. Run-level role counters are used because concurrent per-cell counter differences can include work from other calls. Without comparable model-specific prices or timing, tokens across different models cannot be added into an end-to-end compute saving. The measured communication reduction therefore supports a claim about receiver-visible message count, not a total-cost advantage in either direction.

Table 12: **Prospective power for a three-point gain.** The primary trained subset scorer is compared with each reference at at-most-five messages. Power uses the corresponding empirical discordance. Sample requirements target 80% power under the same assumptions.

Cohort	Questions	Power	Required questions
Reference: receiver-only answering			
Omni-MATH-2	195	0.231	755
MaScQA	200	0.256	690
JEEBench	199	0.189	969
SciBench	199	0.168	1,101
LAB-Bench	182	0.104	$> 2,000$
Reference: scorer initialization			
Omni-MATH-2	195	0.314	529
MaScQA	200	0.493	309
JEEBench	199	0.357	465
SciBench	199	0.328	518
LAB-Bench	182	0.120	1,539

Receiver input length in the fresh study. Text reconstructed from the manifests gives approximately $1{,}019$ additional characters per message on LAB-Bench and $1{,}129$ on MaScQA, over base prompts of $3{,}516$ and 597 characters. Five messages versus one produces input-length ratios of 1.90 and 3.62, respectively. These are character counts for these message-count conditions, not tokenizer counts or a comparison of the actual cost-aware selections. Per-row token accounting was marked unreliable and is excluded. Generation and the selector's 32-input sweep remain separate costs.

B Path Robustness, Repeatability, and Grading

B.1 Density-Preserving Comparisons

The density control preserves the number of successful subsets at every budget for every question. Each of 2,000 random shuffles rearranges the correct-or-incorrect labels among subsets of the same size, separately for each question, then recomputes the best path. Since each layer retains its maximum, budget-specific best accuracy is unchanged in every draw. Path accuracy and path loss therefore carry equivalent evidence under this control.

The primary two-sided Monte Carlo p-value, computed from these random shuffles, is twice the smaller inclusive tail probability, capped at one, with a plus-one correction in each tail. Integer success counts ensure that tied draws are counted identically for path accuracy and path loss. On the mixed-outcome populations, the values are 0.741629 (mathematics), 0.070965 (materials science), 0.866567 (engineering), 0.142929 (college science), and 0.001000 (biology). These are exploratory, unadjusted comparisons. An alternative two-sided convention counts absolute deviations from the null mean; its values are 0.681659, 0.038981, 0.674163, 0.140430, and 0.001000, respectively. The materials-science result depends on this convention, so we do not treat a threshold crossing there as robust evidence. Failure to reject the control is not evidence of equivalence.

Density controls in fresh executions. We repeat the within-question, within-size control separately on the two fresh executions, using the same primary two-sided rule, seed 20260925 and $2{,}000$ draws. Table 13 reports all four tests. None rejects at 0.05; the smallest raw value becomes 0.2599 under Holm correction within this four-test family. The 75-question samples differ from the original full cohorts, so this comparison changes both sample and execution. Different p-values across runs are not a test of a difference between executions.

Biology task types. The biology comparison remains distinguishable from the density control within SeqQA, the largest sampled task type. Table 14 applies the same exactly-budget path-loss statistic separately to task types. These exploratory tests use 2,000 shuffles and the same inclusive-tail

Table 13: **Density control on each fresh execution.** Exact sizes 1–5; 75 questions per cohort in each run. Entries are primary two-sided Monte Carlo p-values, before correction across the four tests.

Cohort	Run A	Run B
LAB-Bench	0.7016	0.0650
MaScQA	1.0000	1.0000

convention; they are not adjusted for multiple testing. Separate Monte Carlo draws do not replace the primary five-cohort p-values above.

Table 14: **Density control within LAB-Bench task types.** Path loss is the total number of question–budget successes missed by the best paths, across exactly 1–5 messages. All questions within each task type are included.

Task type	Questions	Total path loss	Two-sided p
SeqQA	133	4	0.001
ProtocolQA	40	0	0.201
CloningScenarios	9	0	1.000

This result is not solely a contrast between task types, but it is concentrated in the largest sampled group. Non-rejection in the smaller groups does not establish absence of structure, and differing p-values do not establish a difference between groups.

B.2 Choices That Differ at the Same Budget

Variation across budgets need not imply a meaningful choice within a budget. We call a question choice-informative if at least one subset size from one to four contains both a successful and an unsuccessful subset. Sizes zero and five each contain only one subset. Of the 279 mixed-outcome questions, 28 are constant within every size, including the eleven questions for which only the empty subset succeeds. Removing these automatic path matches leaves 251 questions and 241 optimal paths.

Table 15: **Cross-budget comparisons where equal-size choices differ.** Gains are percentage points, averaged over exactly 1–5 messages. The global order is fitted within each cohort and scope. All choices use recorded outcomes.

Cohort	Questions	Optimal paths	$Q - G$	$F - Q$
Omni-MATH-2	40	38	17.00	1.00
MaScQA	25	24	14.40	0.80
JEEBench	34	32	15.29	1.18
SciBench	56	54	12.86	0.71
LAB-Bench	96	93	22.71	0.83

Here G denotes a shared global ordering, Q a question-specific ordering, and F independent selection at each budget. The pooled path-match fraction changes from 269/279 to 241/251; narrowing the denominator does not remove any of the ten questions with positive path loss. Within-size shuffling also preserves which questions are choice-informative. Every excluded question contributes zero path loss in the observed and shuffled tables. With the same shuffle draws, restricting the scope therefore leaves the path-loss test unchanged; this is a property of the statistic, not additional evidence against the null.

Global-order selection sensitivity. The reported mixed-outcome gaps of 11.61–21.80 percentage points use the global order fitted to each original cohort. In a resampling sensitivity analysis, refitting that global order inside each question-bootstrap draw gives mean gaps of 9.84–20.12 points, a reduction of 7.7–18.4% relative to holding the original order fixed in the corresponding draws. This changes the resampling procedure rather than correcting the original point estimate. Both analyses reuse recorded outcomes; neither measures generalization to another receiver execution.

B.3 Frozen Choices on Repeated Execution

Preregistered two-cohort study. We sampled 75 questions from each retained cohort without conditioning on outcomes (seed 20260925). LAB-Bench sampling preserved subtask proportions: 55 SeqQA, 16 ProtocolQA and four CloningScenarios questions. The sample and choices were frozen before either new execution. Both the global ordering and question-specific orderings were fitted on historical outcomes restricted to these same sampled questions; their budget-specific choices were then held fixed. Selection permits at most b messages, including zero, for each $b = 1, \ldots, 5$, with equal budget weights. Ties prefer fewer messages and then the lower subset index.

All 150 questions have complete 32-subset tables in both new executions: 9,600 outcomes and no missing cells. The primary statistic averages the two new execution gains within each question, then averages questions. Percentile intervals resample questions, keeping their budgets and executions together (10,000 draws, seed 20260925). The target is repeatability on these questions, not generalization to new questions.

Table 16: **Frozen choices under two reference policies.** Paired differences in percentage points on 75 questions per cohort, averaging at-most budgets 1–5 and both new executions. "Question-specific" and "global" use choices frozen from historical outcomes. Receiver-only answers the question without shared messages.

Cohort	Contrast	Difference	95% interval
LAB-Bench	Question-specific − global	$+6.93$	$[3.20, 10.93]$
	Question-specific − receiver-only	$+3.87$	$[-4.93, 12.80]$
	Global − receiver-only	-3.07	$[-12.27, 6.00]$
MaScQA	Question-specific − global	$+0.80$	$[0.00, 2.13]$
	Question-specific − receiver-only	$+0.93$	$[-3.73, 5.60]$
	Global − receiver-only	$+0.13$	$[-4.80, 4.93]$

Relative to global ordering, LAB-Bench gains are 6.40 and 7.47 points in the two executions; MaScQA gives 0.80 in each. The LAB-Bench mean loses 4.27 points relative to the fitted value (38% of that value), yet remains positive within its interval. Different interval exclusions do not establish a between-cohort difference. Cell disagreement describes receiver variability; its ratio across cohorts does not identify a mechanism for path gains. The sample retains the original biology cohort's SeqQA-heavy composition. New-run correctness agreement is 91.75% on LAB-Bench and 99.38% on MaScQA. MaScQA's aggregate ordering advantage is carried by two of 75 questions, with zero advantage at budgets three through five.

Neither frozen policy has a resolved advantage over receiver-only answering. The global order's negative LAB-Bench point estimate contributes to its contrast with question-specific choices, but its interval also includes zero: this does not establish that the global policy is harmful. Repeatability relative to one communication policy and net benefit relative to no communication are separate claims.

Earlier 38-question study. A separate earlier repeat analysis uses 38 biology questions with two complete outcome tables. All 1,216 subset prompts match. Distinct execution identifiers, timestamps about 26 hours apart and 1,214 distinct output hashes establish separate executions under the same code and configuration. Correctness agrees on 1,142 cells. Re-exported records from the same execution are not counted as repetitions.

Here the comparison averages exactly-1–5 budgets. We fit both the global ordering and each question-specific ordering on the first execution, then freeze them for evaluation on the second. The first-execution advantage is 4.74 percentage points; the second-execution advantage is -0.53 points. A question-paired percentile bootstrap with 20,000 resamples gives a 95% interval of $[-3.68, 2.11]$. Averaging uniformly over tied first-execution optima gives -0.05 points with interval $[-2.85, 2.74]$. This study covers one receiver and one benchmark; it does not establish repeatability across all five cohorts.

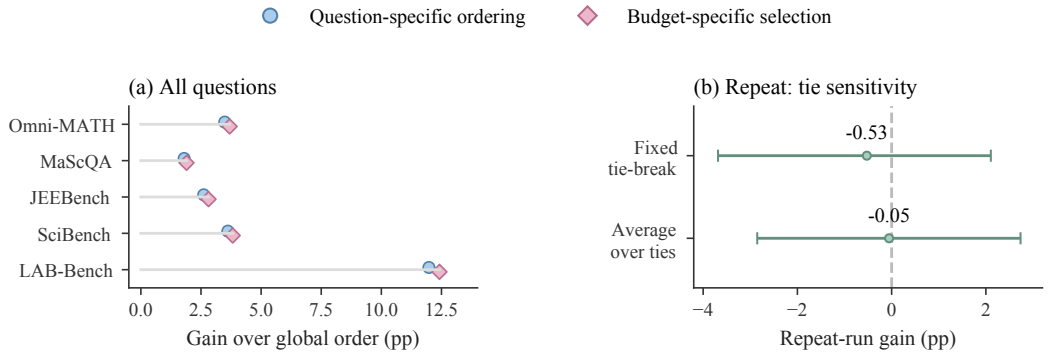

Figure 7: **Full-cohort gains and repeat-run tie sensitivity.** (a) Gains over the best global ordering on all questions in each primary cohort, averaging equally over exactly 1–5 messages. (b) Frozen question-specific minus global-order accuracy on the second execution of 38 biology questions. Error bars are 95% paired-question percentile bootstrap intervals. Fixed tie-breaking and averaging over tied optima lead to the same qualitative conclusion.

B.4 Additional Receiver Repeats and Grading Checks

Receiver-only rescue availability across executions. We call a rescue available when receiver-only answering fails but at least one nonempty subset succeeds. Three repeat sets meet the same-receiver, same-code and same-configuration requirement: 38 earlier LAB-Bench questions and the two fresh 75-question samples. Availability persists from the first execution to the second on 4/5, 14/19 and 3/4 eligible questions, respectively. The union across runs exceeds either run alone in every set: 7 versus 5 and 6, 21 versus 19 and 16, and 6 versus 4 and 5. We therefore report the original 104/975 as a single-execution description and do not extrapolate a repeat-corrected count. A 58-question mathematics pair is excluded because the two executions use different receiver models and code revisions.

Within-receiver full-lattice repeats. The Gemma biology repeat holds receiver, endpoint, evaluator, code and configuration fixed on 38 questions. Mathematics uses two Gemma passes on the 58 held-out questions, 17 days apart, with identical question, candidate-pool and receiver-prompt inputs but different receiver-stage code revisions. Agreement statistics are over questions; cells within a question are correlated.

Table 17 reports agreement and directional changes between repeated full-subset executions.

Table 17: Within-receiver agreement. All 32 subsets per question; Gemma receiver. Intervals are question-bootstrap 95% intervals. Up/down count incorrect-to-correct and correct-to-incorrect cells.

Domain	questions	cells	agreeing	up/down	agreement (95% CI)
LAB-Bench	38	1,216	1,142	44/30	0.9391 [0.9071, 0.9671]
Omni-MATH-2	58	1,856	1,782	32/42	0.9601 [0.9402, 0.9779]

Table 18 reports disagreement by subset size. The apparent gradients do not establish a general subset-size effect. Mathematics has $6/58$ disagreements at the empty set versus $68/1{,}798$ elsewhere;

Table 18: Disagreement between repeated Gemma executions by subset size. Mathematics and biology use 58 and 38 questions, respectively.

$\lvert S\rvert$	0	1	2	3	4	5
Omni-MATH-2	10.3%	4.1%	3.8%	3.8%	3.4%	3.4%
LAB-Bench	7.9%	5.8%	6.1%	6.6%	5.3%	5.3%

a question-cluster-preserving randomization gives $p = 0.0349$. This diagnostic does not survive correction over all six sizes and does not establish a cross-domain difference. The pilot GPT-OSS repeat/order audit is reported separately in Appendix E.6.

Stability of a frozen hindsight choice. On the biology repeat, $33/38$ questions have a successful subset in the first pass. Freezing the smallest such subset gives $30/33$ successes on the second: $26/28$ for empty-set choices and $4/5$ for non-empty choices. All three failures have successful alternatives in the second pass, with 26, 20 and 7 successful subsets respectively. Thus a question remaining solvable does not guarantee that its first-pass optimum repeats.

Cross-receiver agreement. On the same 58 mathematics questions and candidate pools, GPT-OSS and Gemma agree on 89.0% of $1{,}856$ cells, with 130 upward and 75 downward changes. Disagreement is 19.0% at the empty set and 5.2% at the full set. This comparison changes receiver and collection run, so it is not a within-receiver repeatability estimate or a bound on one. The endpoint-solved partition changes from $(7, 44, 7)$ to $(8, 50, 0)$ in (unreachable, endpoint-solved, intermediate-only) counts. These are two receiver-conditioned measurements, not competing estimates of a single noise-free table.

Repeated grading versus answer consistency. Regrading $4{,}032$ byte-identical biology outputs changes one verdict (0.025%). Those cached outputs are not independent receiver replays. A different check groups cells within a question by parsed answer. Across the examined $56{,}512$ cells, $1{,}039$ (1.84%), in 78 answer groups across 53 questions, belong to groups containing both verdicts for the same parsed answer. For example, the evaluator accepts 0.506 as a rounded form of 0.5061 in one cell but rejects it in another. This detects inconsistency, not all grading errors; consistent misgrading remains possible. Numeric answer formats require particular caution. Recorded outcomes are retained rather than selectively removed.

Noise sensitivity of reversal diagnostics. A separate campaign samples 200 questions per domain and repeats $1{,}200$ cells per domain five to ten times. Its six conditions are the full set and five leave-one-out subsets, so it covers sizes four and five only. Pairwise disagreement is 2.23% pooled, ranging from 1.28% on engineering entrance to 4.28% on college science. These are different populations from the full-lattice repeats. Overlap with expanded lattices is partial: $62/182$ biology questions and only $6/195$ mathematics questions, for example.

Calibration and null construction. An independent symmetric flip with probability p gives two-repeat disagreement $d = 2p(1 - p)$. Each domain uses its own measured d; cross-receiver disagreement is not used for calibration. With five to ten draws per cell, the fraction of perfectly stable cells also checks the independent-flip assumption. Observed stable-cell fractions exceed their independent-flip predictions slightly in every domain. The excess of stable cells suggests that noise is somewhat concentrated in a minority of cells. This heterogeneity limits the independent-flip model as a description of the full process.

The additive surrogate addresses a different explanation: context-independent message effects plus noise. For each question, a message's main effect is its mean marginal effect over sixteen contexts. A subset's additive score sums its messages' main effects; a threshold matches that question's observed number of successful subsets. Before noise, this construction has no sign reversals and preserves each question's success count. Noise is then applied at the domain's measured rate. All reported reversal rates use the same denominator, the $5n$ question–message pairs. Conditioning instead on pairs with a nonzero effect would change the denominator across observed and simulated lattices.

Sensitivity to fitting and calibration. Estimating main effects on a random half of each message's contexts and evaluating all 32 cells changes the bands by at most 0.006 and does not change their qualitative comparisons. This is a partial cross-fit because evaluated cells include fitted cells. Restricting evaluation to held-out contexts narrows the bands substantially: for biology, $[0.019, 0.054]$ versus $[0.061, 0.124]$ under the full fit; for materials QA, $[0.001, 0.014]$ versus $[0.007, 0.038]$. Two domains then move above their bands. These analyses change the evaluation design as well as data reuse. We therefore retain the more conservative full-fit and partial-cross-fit comparisons and do not interpret narrower bands as stronger evidence.

Calibration uncertainty also matters. Measured disagreement is 0.0268 $[0.0171, 0.0372]$ for biology and 0.0131 $[0.0065, 0.0207]$ for materials QA. Repeating the null at interval endpoints moves materials QA from above the flip-null band at the point estimate and lower endpoint to inside it at the upper endpoint. At the point estimate its one-sided empirical p-value ranges from 0.013 to 0.028 under different tie conventions. We treat this comparison as unresolved.

What the nulls reproduce. Biology and mathematics exceed both independent-noise and context-independent-effect references under the reported calibration. College science does not; the engineering-entrance result depends on the reference used. Materials QA exceeds only the flip reference, and that result depends on its calibration interval. These are sensitivity findings, not a universal noise correction to structural rates.

The mixed-outcome question fraction supplies a second diagnostic. On biology, the additive surrogate predicts $[0.665, 0.753]$, versus 0.549 observed. The additive discrepancy has a mechanical component: before noise it already matches the observed success counts; independent flips can turn a constant-outcome question into a mixed one, whereas the reverse requires all cells of one outcome to flip. Its excess of mixed questions is therefore not independent evidence for interaction. No examined null is established as an adequate generative model for all five domains.

Agreement among nulls is also dependent. They share the repeat-noise calibration, and the additive surrogate reduces to the flip model's constant base on questions where all 32 outcomes agree. Such questions comprise 45–85% of each domain. These comparisons cannot be counted as independent replications. Permuting outcomes within each question places all five observed reversal rates below the permutation band, but destroys additive structure as well as interactions. It is a reference for disorder, not a specific test of context dependence.

Coverage limits. The calibration campaign covers only subset sizes four and five. Applying its rates to sizes zero through three is a sensitivity assumption. The full-lattice repeats above provide direct checks on two cohorts, but do not supply the same five-domain calibration. The campaign also samples different questions: overlap with the lattices is 34% $(62/182)$ in biology, 33% in college science, 40% in engineering entrance and 3% $(6/195)$ in mathematics. In particular, the mathematics null uses a noise rate measured almost entirely outside its evaluated lattice. Both coverage limits qualify the interpretation of the reversal comparisons.

Frozen-policy gains test whether selected choices remain useful; they do not estimate how often a particular sign reversal recurs. Question-level bootstrap intervals for single-execution reversal rates do not include repeated-call uncertainty. Hindsight optimization over single noisy outcomes can be optimistic; neither a recorded maximum nor its difference from a baseline is a guaranteed bound on expected performance.

B.5 Sensitivity to Deterministic Grading Checks

We retain the original labels as the primary measurement and examine two alternatives. The corrected variant substitutes a deterministic verdict only where the check can decide and disagrees with the original grade. The excluded variant removes those disagreements and retains only questions with all 32 cells remaining. Undecidable cases retain their original labels. These checks do not establish that every retained grade is correct.

Table 19: **Selected headline quantities under grading sensitivity.** Accuracy is in percent. Exclusion changes the question population. "Best singleton (oracle)" selects using recorded outcomes; five messages admit only one subset.

Labels	Questions	Receiver only	Best singleton (oracle)	All five
Original	975	82.15	89.54	81.95
Corrected	975	82.46	89.33	82.26
Excluded	957	82.24	89.45	82.55

The full set remains below the best-singleton oracle in all three variants. Its small difference from receiver-only answering changes sign after exclusion, so that sign is not a robust finding. Within the

choice-informative analysis, JEEBench changes from 34 questions and a 15.29-point $Q - G$ gap to 31 and 12.26 points under correction, or 28 and 10.71 points under exclusion. LAB-Bench and MaScQA are unchanged under these particular checks. This limited invariance is not a general validation of their graders.

B.6 A Common Optimal Path Across Executions

Having an optimal growing path in each execution does not imply that the same path is optimal in both. We enumerate all optimal orderings separately in the two fresh executions, then test whether their sets intersect. This is an exactly-budget comparison over sizes one through five, distinct from the at-most-budget frozen-policy gains in Appendix B.3.

Table 20: **Within-run and shared path compatibility.** Each within-run denominator contains that run's choice-informative questions. The final column restricts to questions choice-informative in both runs and asks whether at least one ordering is optimal in both.

Cohort	Run A	Run B	Shared ordering
LAB-Bench	35/37	31/36	26/35
MaScQA	8/8	7/7	7/7

Selecting an arbitrary representative from tied optimal orderings gives a different question. Its transfer rate can depend on which representative is selected, so a directional difference alone does not establish structural asymmetry. Existence of a shared optimum is symmetric by construction. Only one of the ten original questions without an optimal growing path appears in this repeat sample, so these results do not characterize the repeat behavior of all ten failures.

C Additional Recorded Examples

C.1 A Message Rescues an Answer Before Another Addition Harms It

Figure 8 presents the recorded example behind Figure 1a. The question is MaScQA G-META-16-32, evaluated with the Gemma receiver and five cached GPT-OSS messages. Italic text reproduces the messages' key-idea fields, with mathematical notation and spacing normalized for typesetting. These are model-generated claims, not endorsed explanations. The table reports actual receiver answers from separate subset calls.

The distinction between a message proposal and the receiver's final answer matters here. Messages D and E both propose positive values, yet the receiver given D and E returns a negative value within the reference interval. With A and E, it returns a positive value. Other pairs retaining A—AB, AC and AD—also succeed, so the example demonstrates a harmful addition and an alternative successful pair, not the necessity of replacing A.

C.2 Two Recorded Same-Message, Two-Context Examples

Both examples are Omni-MATH-2 tier-10 source-training questions under the frozen GPT-OSS receiver, five cached GPT-OSS messages and canonical order. The contexts below use one-based message indices. Each row consists of two stateless replays. Examples are selected to illustrate the measured interaction, not randomly sampled to estimate its prevalence or to establish a reasoning mechanism.

Table 21 records the four subset comparisons used in these two examples.

A geometric construction. Problem `t10:3744` asks for the positive thresholds t for which any number of nonoverlapping triangles can be inscribed in a unit circle with each perimeter exceeding t. The fixed message proposes a fan construction. It proposes $0 < t < 4$; the four recorded utilities are sufficient to establish the displayed sign reversal, without assuming that the proposal itself diagnoses why the receiver changes its answer. This question comes from the diagnostic's training-only tier-10 subset.

Recorded example: silicon standard-state conversion

Question (original wording; notation normalized). *The change of standard state from pure liquid to 1 wt.% for Si dissolved in liquid Fe at 1873 K is expressed as Si (liq.) = Si (1 wt.%). Given that the activity coefficient of Si at infinite dilution in Fe is* 10^{-3}*, the standard Gibbs free energy change (in kJ) for this equilibrium is?*

Reference interval: -168.7 to -168.1 kJ.

Candidate messages A–E: key-idea excerpts

A. *The standard-state Gibbs energy change equals* $\mathrm{RT}\ln(\gamma^{\infty}x_{1\mathrm{wt.\%}})$*, because the reference activity for Si in Fe is defined at 1 wt.% (*$a = 1$*) while the actual activity at that composition is* $\gamma^{\infty}x$*.*
Proposed answer: approximately -168 kJ.

B. *The standard-state activity of Si in the 1 wt.% Fe solution is* γX*; thus the equilibrium constant is simply this tiny activity because the activity of pure liquid Si is unity.*
Proposed answer: approximately $+170$ kJ.

C. *The equilibrium constant for the standard-state change is the activity of Si in the 1 wt.% Fe–Si solution, i.e.* $K = \gamma \cdot X_{\mathrm{Si}}$ *(*$\approx 10^{-3} \times 0.02 \approx 2 \times 10^{-5}$*).*
Proposed answer: approximately $+180$ kJ.

D. *Activity of Si in the 1 wt.% Fe solution is* $\gamma \times X = 10^{-3} \times 0.01 = 10^{-5}$*, giving* $K \approx 10^{-5}$*.*
Proposed answer: approximately $+180$ kJ.

E. *The activity of Si in the 1 wt.% Fe melt is* $\gamma^{\infty} \times$ *mole-fraction of Si, giving a very small equilibrium constant and thus a large positive* ΔG°*.*
Proposed answer: approximately $+170$ kJ.

Actual receiver outputs

Shared messages	Answer (kJ)	Reference match
None (question only)	-107.5	✗ Incorrect
A	-168.66	✓ Correct
A and E	$+168.7$	✗ Incorrect
D and E	-168.6	✓ Correct

Figure 8: **Real message excerpts and receiver answers for the thermodynamics example.** Message letters are fixed identities in the cached pool, not answer choices. Proposed message answers are approximate values shown by the source messages; receiver answers are the recorded parsed outputs. Every receiver call includes the complete question and selected messages in canonical order. The displayed grades agree with the explicit reference interval. This is a single recorded comparison, not a repeated success-rate estimate or proof of a textual mechanism.

Table 21: **Four observed outcomes per sign-reversal example.** S contains messages already selected; m is unchanged across the two contexts of a question. Utilities are evaluated binary outcomes.

Problem	S	Added m	$U(S)$	$U(S \cup \{m\})$	Δ
`t10:3744`	$\varnothing$	m_4	0	1	$+1$
`t10:3744`	$\{m_1, m_2\}$	m_4	1	0	-1
`t10:3820`	$\{m_4\}$	m_2	0	1	$+1$
`t10:3820`	$\{m_3\}$	m_2	1	0	-1

An absolute-difference construction. Problem `t10:3820` asks about an anti-Pascal triangle with 2,018 rows, where entries are absolute differences of adjacent entries below and all integers from 1 to 2,037,171 are required to appear. The fixed message begins its key claim with the verbatim excerpt `The set of numbers obtainable from a bottom row` and proposes that the construction exists. Adding it after m_4 helps; adding the identical text after m_3 hurts. This question comes from the source-training population. Neither example licenses a general claim that its proposed mathematical construction is valid.

C.3 THE WORKED EXAMPLE IN FULL

A MESSAGE THAT HELPS ALONE AND HARMS IN COMPANY

Omni-MATH-2 question `omni2:t6:3929`, from the source held-out split, shows the phenomenon in four recorded GPT-OSS receiver calls.

The question asks for the maximum of $a_{2018} - a_{2017}$ in a sequence defined by $a_0 = 0$, $a_1 = 1$, and a_n equal to the mean of some k of its immediate predecessors. The correct answer is $2016/2017^2$. Three of the five candidate messages argue, by visibly different routes, for $1/4034$; none proposes the correct value.

Table 22: Recorded receiver answers for the held-out sequence example. Each row is a separate subset replay.

messages shown	correct	receiver's answer
none	✗	$1/4$
one message (a recursive-structure argument)	✓	$2016/2017^2$
one message (a dynamic-programming argument)	✓	$2016/2017^2$
both of those messages	✗	$1/4034$

Either message alone accompanies a correct receiver answer although neither states it. With both visible, the recorded answer is the incorrect value they share. This illustrates context-dependent effects; the four outcomes alone do not establish why the receiver changed its answer or rule out an ordering that selects a different successful pair.

We report this as an illustration, not as evidence for a mechanism. Each cell is a single replay at temperature zero, so an individual outcome is not separable from receiver variability; the aggregate rates in Section 5 describe prevalence in their stated populations.

D REPRODUCIBILITY AND DATA PROVENANCE

The primary measurement consists of 31,200 valid outcome records for 975 questions, with 32 subsets per question–message-pool–receiver unit. Re-exports are not independent observations: the 172-question biology view overlaps the 182-question evaluation. Execution metadata distinguish the independent repeat study from these duplicated records.

The supplementary analysis archive includes the primary outcomes, selector scores, parsed-answer records and answer keys, together with the added subset-scorer seed and matched 58-question receiver comparison. It includes identifiers, subset masks, measured outcomes, selector scores and relevant provenance fields, with a claim-to-file map and reconstruction commands. It omits benchmark question and message text, full reference solutions, model weights and endpoint credentials. These records support reconstruction of the included analyses and new comparisons using the recorded outcomes and scores. They do not support independent regrading, retraining or new receiver calls without the omitted inputs. Availability of the underlying benchmark inputs remains separate from availability of these measurements.

Checkpoint attribution is specific to each reported model. Table 10 uses newly computed predictions with identified adapters for three pairwise checkpoints against retained source-receiver outcomes. The two primary-scorer source comparisons now have corresponding per-subset score records (Table 4); historical sequential-selector source summaries are separate. Figure 6 uses four transfer cohorts with retained prediction records. The main learning comparisons use the expanded-cohort scores and outcomes; historical source-side summaries are not pooled into those conclusions.

Recorded grades are not assumed infallible. In a strict JEEBench example, the evaluator rejects $11/51$ for a single message but accepts the same fraction for a pair as rounding to the reference 0.22. That case cannot establish a reasoning failure caused by adding messages. The corresponding LAB-Bench case has option letters consistent with its 32 grades: C alone succeeds, every pair retaining C fails, and AE and DE succeed. Both cases succeed without messages. Figure 1 uses a separate MaScQA

rescue-and-harm example. Appendix B.5 reports the aggregate sensitivity to deterministic grading corrections and exclusions.

Scope of the evidence. The experiment varies which of five fixed messages reaches a receiver at one sharing stage. It does not vary message generation, presentation order, network topology or subsequent dialogue. Exhaustive measurement requires 2^M receiver calls for M messages and is therefore suited to small candidate pools. The observed selector results concern the evaluated 4B checkpoints and transfer settings, rather than a capacity limit for learned communication. Training histories document optimization; they do not replace variation across training seeds.

Analysis design and uncertainty. The fresh-repeat study freezes its choices before collecting the new outcomes. The density controls, grading sensitivity and subgroup diagnostics analyze already recorded outcomes; they are not prospective predictions. Bootstrap intervals for individual contrasts are pointwise intervals. The model-comparison tables separately identify the families using Holm correction. Subsets, budgets and overlapping cohort exports are not independent experimental replications.

E Complete Measurement Results

The complete outcome table contains one row for each subset of the five messages. We also call this table a *subset lattice*: its subsets can be connected by adding one message at a time. A successful subset is one for which the receiver's final answer is graded correct. These terms describe recorded outcomes, not repeated success probabilities. We use the same notation as the main text: $U_g(S; q)$ is recorded correctness for receiver g, question q and visible subset S; $\mathcal{M}$ is the full set of five messages. With receiver and question fixed, $U(S)$ abbreviates $U_g(S; q)$.

E.1 Message Effects by Cohort

Figure 9 separates helpful, harmful and sign-reversing effects within each primary cohort. It retains the full question–message denominator while displaying the active pairs.

E.2 Receiver Completion and Terminal Outcomes

The fixed concise prompt uses receiver/evaluator output ceilings of 32,768/4,096 tokens and temperature zero. A ceiling is not actual usage or a deployment's context-window size. Prompt-calibration questions are excluded from training, validation and final tests.

Each scored subset has one terminal outcome: a graded answer or a no-answer. A successful receiver request without a parseable final answer has utility zero, including early stops and cap exhaustion. It remains in accuracy/cost denominators and is not selectively retried. Authentication, rate-limit, server, connection, timeout and evaluator-parse failures instead leave the measurement unresolved; they are not scientific zeros.

Records retain the runner's raw/parsed answer fields, utility, prompt/configuration identities and role accounting. These fields do not establish that every separate API reasoning field was archived; aggregate completion-token counts do not separately measure reasoning tokens. A complete table requires 32 logical receiver calls plus grading calls for answer-bearing states; no-answers need no grading. Thus 32 subsets are not 32 total API requests. Physical attempts and recovered retries are counted separately.

E.3 Diagnostic Definitions and Denominators

A state is one question and one visible subset. An add-one edge is a state plus an available message. For N complete five-message tables there are $32N$ states, $80N$ edges, $5N$ question–message pairs, and N independent questions. A rate over edges is not a rate over questions.

Context dependence. For an add-one comparison, Δ is correctness after adding the message minus correctness before adding it. Helpful, harmful and neutral additions therefore have $\Delta = +1, -1, 0$,

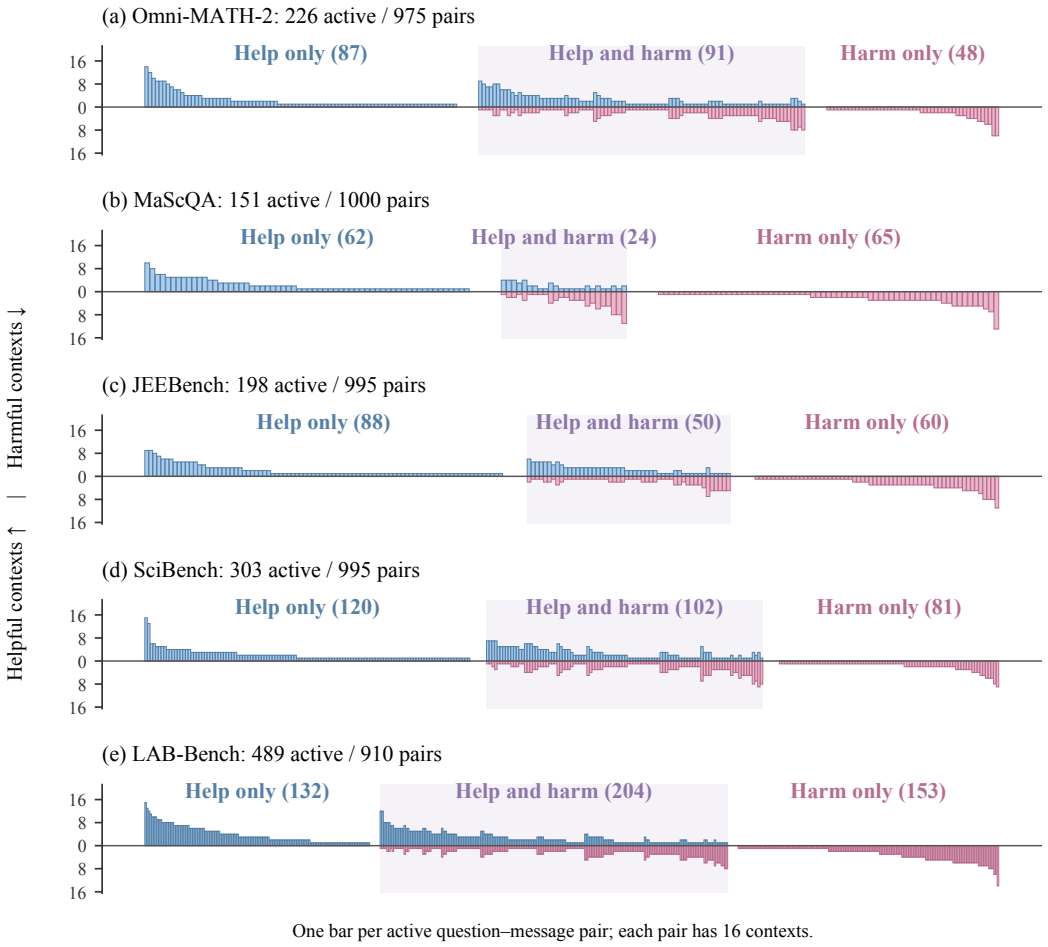

Figure 9: **Helpful and harmful contexts for each active question–message pair.** Primary Gemma-receiver cohorts, totaling 975 questions and 4,875 pairs. Each bar represents one pair; blue counts helpful contexts above zero and pink counts harmful contexts below zero, out of 16 contexts. Pairs with neither effect are omitted: 749, 849, 797, 692 and 421, respectively. Within-cohort sorting makes patterns readable and does not identify a message across questions. Shaded groups contain both signs.

respectively. For one question–message pair, a strict sign reversal requires both a positive and a negative edge across its 16 possible contexts. The main question-level statistic asks whether at least one message reverses sign. A strict ranking reversal requires two messages available in both compared contexts, with one strictly preferred in the first and the other strictly preferred in the second. A transition from a strict preference to a tie is reported separately.

Subset benefit and oracle headroom. An oracle here is a comparison that can inspect recorded answers before choosing; it is not a deployable selector. Its headroom is the improvement that a better choice could achieve over the stated baseline in those records.

A proper-subset rescue occurs when some $S \subsetneq \mathcal{M}$ has $U(S) > U(\mathcal{M})$. This is a hindsight existence statement, not a policy that can identify S without labels. An all-correct or all-wrong table contains no correctness variation even if its answer texts differ. Such cases remain in the population; dropping neutral cases would overstate the prevalence of conditional value.

Monotonicity would require $U(A) \leq U(B)$ for all $A \subseteq B$; a harmful add-one edge violates it. Submodularity would require diminishing marginal returns, $\Delta(m \mid A) \geq \Delta(m \mid B)$ for $A \subseteq B$

Table 23: GPT-OSS difficulty diagnostic. Rates are over questions, each with 32 subset states. Any effect means at least one non-neutral edge; sign reversal and subset rescue are defined in Appendix E.3. Brackets are tier-stratified problem-bootstrap 95% intervals for sign reversal.

Difficulty	Questions	Any effect	Sign reversal	Subset rescue
Tiers 1–4	40	9/40 (22.5%)	6/40 (15.0%) [5.0,25.0]	2/40 (5.0%)
Tiers 5–6	20	10/20 (50.0%)	6/20 (30.0%) [10.0,50.0]	3/20 (15.0%)
Tiers 7–10	40	22/40 (55.0%)	17/40 (42.5%) [27.5,57.5]	8/40 (20.0%)

and $m \notin B$. A violation is an interaction diagnostic, not proof of a cognitive mechanism inside the receiver.

E.4 Omni-MATH-2 Difficulty Diagnostic and Sampling Rationale

Tiers are the difficulty levels carried by Omni-MATH-2 items, from 1 (easiest) to 10 (hardest). Before receiver calls, the diagnostic sampled ten questions without replacement per tier with a fixed seed. Eligibility required authoritative question/reference identities and five renderable cached GPT-OSS messages; questions used for calibration, pilot checks and repeatability tests were excluded. Selection used no receiver outcome or message-quality label.

The 100 questions produced 3,187 graded answers and 13 no-answers, with no unresolved operational errors. Table 23 reports broad bands. Ten questions per tier support exploration, not a precise per-tier prevalence or monotonic difficulty trend.

The hard band contains 242 helpful and 199 harmful additions among 3,200 edges; 2,759 edges (86.2%) are neutral. The corresponding counts are 110 helpful, 85 harmful and 3,005 neutral out of 3,200 easy-band edges, and 82 helpful, 51 harmful and 1,467 neutral out of 1,600 middle-band edges. Thus a substantial fraction of questions can contain a sign reversal even when most individual additions are neutral.

No-message/reveal-all accuracies are 87.5/95.0% in tiers 1–4, 55.0/75.0% in tiers 5–6, and 50.0/65.0% in tiers 7–10. These observations motivated collecting eligible tiers 7–10 and including tier 6 as a medium-difficulty bridge. They do not justify selecting individual questions because their observed effect is large. Tier 6 also avoids treating the small diagnostic as evidence for a sharp difficulty cutoff.

Table 24 separates the source partitions and additional question pools.

Table 24: Frozen question-level split within the eligible tier-6–10 population. The representative test and hard stress are disjoint. Diagnostic questions reused in this population belong to training only.

Population	Train	Validation	Representative test	Hard stress	Total
Tiers 7–10	690	12	36	100	838
Tier 6	464	7	22	0	493
Total	1,154	19	58	100	1,331

Fifty tier-6–10 diagnostic questions are reused in training, leaving 1,281 new questions, collected. All states and receiver variants inherit the question's split. The diagnostic informed the sampling design and is not an independent confirmatory test of that population.

E.5 Expanded Source-Training Results and Action Labels

All results in this section use only the 1,154 core training questions: 1,104 newly collected questions and 50 compatible tier-6–10 diagnostic questions. They contribute 36,928 utilities and 92,320 add-one comparisons. Validation, representative-test and hard-stress outcomes are excluded. The 50 lower-tier diagnostic questions are not part of this analysis. Question-level aggregate intervals use 2,000 bootstrap resamples within tiers (seed 20260908), retaining the observed tier counts. They quantify variation across questions in this population, not repeatability across receiver calls.

Table 25: **Conditional effects by Omni source-training tier.** All rows use the frozen GPT-OSS receiver. Sign reversal and subset rescue count questions with at least one event; helpful and harmful percentages use $80n$ edges, including neutral edges in the denominator. Tier 10 has only 15 training questions. No monotonic trend is asserted.

Tier	n	Sign reversal	Subset rescue	Helpful edges	Harmful edges
6	464	170/464 (36.6%)	80/464 (17.2%)	6.2%	5.2%
7	290	100/290 (34.5%)	52/290 (17.9%)	6.3%	4.8%
8	260	102/260 (39.2%)	52/260 (20.0%)	7.1%	5.8%
9	125	45/125 (36.0%)	33/125 (26.4%)	7.4%	6.6%
10	15	7/15 (46.7%)	6/15 (40.0%)	8.2%	7.2%

Harmful-edge frequency drops from tier 6 to 7 before rising; sign-reversal frequency is also non-monotonic (Table 25). The data support conditional effects throughout the selected tiers, not a claim that every increase in difficulty strengthens them. At the question–message level, 1,546 of 5,770 pairs exhibit a strict sign reversal. At the question–message-pair level, 1,504 of 11,540 pairs exhibit a strict ranking reversal. Multiple qualifying messages can belong to one question, so these count message-level events rather than questions.

Table 26 counts the available training comparisons by subset size.

Table 26: **Add-one effects by selected-set size.** GPT-OSS, 1,154 Omni training questions. Counts include every edge; $k = |S|$ is the number of messages already selected. A helpful edge has $\Delta = 1$, a harmful edge $\Delta = -1$, and a neutral edge $\Delta = 0$.

k	Helpful	Harmful	Neutral	Total	Helpful / total
0	742	421	4,607	5,770	12.9%
1	1,642	1,352	20,086	23,080	7.1%
2	2,116	1,753	30,751	34,620	6.1%
3	1,257	1,257	20,566	23,080	5.4%
4	328	242	5,200	5,770	5.7%
Total	6,085	5,025	81,210	92,320	6.6%

Of the 1,154 complete tables, 320 are correct at every subset, 198 wrong at every subset, and 636 contain both outcomes. Retaining all three classes prevents selecting only questions that display the desired effect.

Immediate action labels. Each question contributes 31 non-full decision states. A positive add-one effect is an acceptable reveal; stopping has immediate advantage zero. Preferring STOP over a neutral reveal is a communication-saving convention, not a uniquely accuracy-optimal choice. At a failing state, it does not imply that no larger subset can succeed. Only $1/31$ of the uniformly enumerated decision states is empty, so action-label accuracy is not a substitute for evaluation of rollouts starting with no messages.

E.6 Receiver Repeatability and Canonical Order

The existing audit sampled 20 pilot questions, four per domain, with a fixed seed (20260824). It compares three canonical-order trials (the pilot and two additional repeats) and three order conditions (canonical and fixed permutations with seeds 11 and 23). Each family adds 1,280 receiver-state replays beyond the shared canonical baseline. Order-related disagreement is compared with disagreement already present under identical repeats, rather than attributing every changed answer to message position.

The main reliability summaries are ordinary state-outcome agreement and best-action agreement. State-outcome agreement averages all three pairwise comparisons among trials, across 32 states per question: 1,920 comparisons, clustered within 20 questions. Best-action agreement requires the action set to match across *all three* trials at each of 31 non-full states: 620 comparisons. The audit uses all tied positive best reveals when any exist, and the singleton STOP otherwise. Thus this

is agreement under the STOP-preferred-at-zero convention, not equality of all utility-maximizing actions including zero-gain reveals. Edge-sign agreement and exact agreement of all 32 states are stricter supporting diagnostics, not new headline metrics.

Canonical order defines the primary replay condition; it does not assert permutation invariance. A random sample supports a repeatability estimate, whereas targeted sign-flip examples illustrate recorded interactions only. Expanded order/repeat studies are robustness extensions, not a reason to reinterpret a single recorded utility as an expected probability. Table 27 distinguishes repeated answering from changes to presentation order.

Table 27: **Existing pilot reliability audit.** GPT-OSS, 20 questions (four each from Omni, JEEBench, SciBench, LAB-Bench and MaScQA). Utility agreement is pairwise; action-set and entire-table agreement require all three trials to match. Counts are correlated within question, not independent sample sizes. These descriptive aggregates are not a formal equivalence test.

Comparison	Utility agreement	Action-set agreement	Entire 32-state table
Identical repeats	1,788/1,920 (93.1%)	538/620 (86.8%)	12/20 (60.0%)
Fixed orders	1,778/1,920 (92.6%)	541/620 (87.3%)	12/20 (60.0%)

Edge signs agree across all three trials on 1,359/1,600 edges (84.9%) for repeats and 1,358/1,600 (84.9%) for order conditions. Among 26 question–message instances showing a sign reversal in at least one repeat trial, 18 show it in at least two; this conditional 18/26 statistic is not a population prevalence estimate. Order disagreement exceeds repeat disagreement by 0.52 percentage points descriptively. This supports canonical order as a fixed working condition, not order invariance, equivalence, or attribution of all remaining variation to decoding alone. No independent repeat audit of the expanded 1,154-question training population is available.

E.7 Immediate and Lookahead Oracle Diagnostics

A reveal means sharing one additional message. For a remaining message budget b, the lookahead oracle examines combinations of further messages and finds the best recorded final outcome reachable by revealing at most b additional messages:

$$V_0^\star(S) = U_g(S; q), \tag{2}$$

$$V_b^\star(S) = \max\left\{U_g(S; q), \max_{m \notin S} V_{b-1}^\star(S \cup \{m\})\right\}. \tag{3}$$

If no message remains, only stopping is feasible. Immediate greedy selection uses positive current Δ rather than future reachability. A gap between the two detects combinations that immediate selection can miss. Both oracles use measured counterfactual outcomes unavailable to the deployable selector.

A secondary cost-sensitive version subtracts $\lambda c(m)$ from each reveal, where $c(m)$ is a stated communication cost and λ is fixed without test outcomes. This penalty does not imply another receiver call for each reveal: deployment makes one terminal receiver call. Primary reporting uses explicit budgets and actual realized cost. Tie handling prefers stopping when no accuracy improvement is available; tied positive reveals and their reporting order remain explicit.

For static diagnostics, leave-one-out value measures $U(\mathcal{M}) - U(\mathcal{M} \setminus \{m\})$. The BANDMAS-style two-context score (Dong & Lin, 2026) is

$$\tfrac{1}{2}\left[U(\{m\}) - U(\varnothing) + U(\mathcal{M}) - U(\mathcal{M} \setminus \{m\})\right]. \tag{4}$$

Shapley and Banzhaf scores summarize a message by averaging the change caused by adding it to different subsets of other messages. They average these add-one effects over contexts with their respective weights. Each compresses a message's context-specific effects into one number. Computing them on a test lattice creates an oracle diagnostic, not a test-time feature for a learned selector.

Table 28 quantifies the immediate target's limitations on training data. At one reveal, immediate and lookahead coincide. Lookahead improves with additional budget, reaching the unrestricted best-subset value by budget four. Sixty-three of 1,154 questions (5.5%) are planning-rescue cases:

Table 28: **Immediate versus lookahead headroom, with an expected-random reference.** GPT-OSS source-training mathematics, $n = 1,154$ questions, starting at the empty set. Entries are empirical terminal accuracies on the same recorded lattices, with the tie rule above. Only the two oracle columns inspect utilities when selecting. Expected random averages all $\binom{5}{b}$ size-b subsets exactly. Budgets limit additional reveals; they do not equal realized cost.

Budget	Immediate oracle	Lookahead oracle	Expected random
0	54.2%	54.2%	54.2%
1	77.4%	77.4%	59.7%
2	77.4%	80.8%	61.0%
3	77.4%	82.6%	62.0%
4	77.4%	82.8%	62.0%
5	77.4%	82.8%	63.5%

lookahead can succeed but immediate greedy cannot. This is a limitation of the chosen target, not evidence for an untested reinforcement-learning gain.

At every nonzero budget, immediate greedy uses 0.232 messages per question on average, whereas the expected-random reference uses exactly b messages. Under binary utility, any positive first reveal reaches utility one; with STOP preferred on subsequent zero-gain ties, the immediate oracle cannot benefit from a larger cap. Its roughly 17-point advantage over random at budget one is label-aware selection headroom, not a learned improvement or a matched-realized-cost comparison. Oracle optimization on a single recorded table can exploit replay noise; its accuracy is not a repeated-call guarantee.

E.8 Receiver Transfer and Adaptation

The receiver comparisons hold question and candidate-message texts fixed and use the same GPT-OSS evaluator. Fixed-selector transfer evaluates an unchanged selection under each receiver's outcomes. None of the evaluated selectors receives receiver identity as an input, and no receiver-adapted selector is included. These comparisons therefore measure the performance of fixed choices across receiver conditions, not receiver-specific adaptation. Exact paired cohorts are given in Section E.10.

E.9 Extended Measurement Analyses

E.9.1 Populations and recorded utility

The expanded mathematics collection contains all 58 held-out questions and 137 additional questions in lower difficulty tiers. Under Gemma, receiver-only accuracy is 0.847 on the additional 137 versus 0.759 on the held-out 58; reveal-all is 0.898 versus 0.707. Their pooled 195-question result therefore changes the difficulty mixture. The four additional domains have matched complete lattices under both receivers, with intersections 200, 198, 188 and 180 for materials QA, engineering entrance, college science and biology. These intersections, rather than the smaller of the two marginal counts, define paired comparisons.

Each recorded utility is a binary outcome of one stateless replay at a canonical message order. It is not the expected probability of success. Successful requests without a parseable answer receive zero; operationally unresolved cells remain missing. Complete lattice analyses exclude questions missing any of their 32 outcomes.

E.9.2 Structural diagnostics and headroom

On the 1,154 source-training questions, reveal-all accuracy is 63.5% and the recorded best-subset accuracy is 82.8%. Receiver-only answering succeeds on 625 questions (54.2%), one message is sufficient on another 268 (23.2%), two or more are needed on 63 (5.5%), and no subset succeeds on 198 (17.2%). The corresponding held-out proportions are 67.2%, 10.3%, 10.3% and 12.1% over 58 questions. These are recorded existence statements, not predictions available to a deployed selector.

Reference-dependent headroom. A question is *unreachable* if every subset fails; *endpoint-solved* if either receiver-only answering or reveal-all succeeds; and *intermediate-only* if both endpoints fail but another subset succeeds. The last category measures headroom above an outcome-aware chooser of the better endpoint, not headroom above either fixed baseline.

Table 29: Reference-dependent headroom on complete outcome tables. Endpoint-solved means receiver-only or reveal-all succeeds.

receiver	domain	n	unreachable	endpoint-solved	intermediate-only
Gemma	JEEBench	199	3	192	4
Gemma	LAB-Bench	182	16	147	**19**
Gemma	SciBench	199	31	157	11
Gemma	MaScQA	200	2	193	5
four shifted domains, pooled unique questions		780	52	689	**39** (5.0%)
Omni-MATH-2, source receiver		58	7	44	**7**

Across the four expanded transfer domains, the two endpoints disagree on 102 of the 689 endpoint-solved questions. Adding the 39 intermediate-only questions gives at least $141/780$ questions on which a selection can change correctness. Even both-successful endpoints do not guarantee that every intervening subset succeeds.

E.9.3 Compact-representation values and resampling

Table 30 reports the full-cohort counterparts of Figure 4. The question gain compares a question-specific ordering with the global ordering. The remaining budget gain compares budget-specific selection with that question-specific ordering. Their sum is the total gain over a global ordering.

The reported bootstrap keeps the fitted global and per-question orders fixed and resamples questions. Re-selecting the global order inside each resample is a separate sensitivity analysis. On SciBench it gives a bootstrap mean question gain of 0.0324 with interval $[0.0201, 0.0462]$, compared with the fixed-order estimate 0.0362 and interval $[0.0221, 0.0523]$. The bootstrap mean is not substituted for the original sample estimate. Per-budget values appear in Appendix E.11.1. Nonnegative gaps follow from nested decision classes; their signs alone are not independent evidence.

E.9.4 Attribution summaries

Scores use singleton contrasts $U(\{m\}) - U(\varnothing)$, leave-one-out contrasts $U(\mathcal{M}) - U(\mathcal{M} \setminus \{m\})$, or Shapley/Banzhaf context averages. Each induces a message order with canonical-index tie-

Table 30: **Hindsight decision representations on five expanded Gemma cohorts.** Mean accuracy over exactly-k budgets 1–5. Total gain is budget-specific selection minus global ordering; question gain is question-specific minus global; budget gain is budget-specific selection minus question-specific ordering. Total gain equals the two component gains up to rounding. Paired question bootstrap, 2,000 resamples, seed 20260923, fitted orders held fixed. All classes use recorded outcomes to select and are not deployable.

Domain	n	Global order	Question-specific	Total gain	Question gain	Budget gain
Omni-MATH-2	195	0.8369	0.8718	$+0.0369$ $[+.0215, +.0544]$	$+0.0349$ $[+.0205, +.0513]$	$+0.0021$ $[.0000, .0051]$
MaScQA	200	0.9320	0.9500	$+0.0190$ $[+.0090, +.0320]$	$+0.0180$ $[+.0090, +.0300]$	$+0.0010$ $[.0000, .0030]$
JEEBench	199	0.9176	0.9437	$+0.0281$ $[+.0141, +.0432]$	$+0.0261$ $[+.0121, +.0402]$	$+0.0020$ $[.0000, .0050]$
SciBench	199	0.7508	0.7869	$+0.0382$ $[+.0241, +.0543]$	$+0.0362$ $[+.0221, +.0523]$	$+0.0020$ $[.0000, .0050]$
LAB-Bench	182	0.6813	0.8011	$+0.1242$ $[+.0923, +.1560]$	$\mathbf{+0.1198}$ $[+.0890, +.1516]$	$+0.0044$ $[.0000, .0110]$

Brackets give 95% intervals. A budget-gain interval reaching zero is not an equivalence claim.

breaking. The table averages accuracy across exactly-k budgets one through five on the expanded Gemma cohorts. Intervals are paired question bootstraps; up/down/tied refer to each question's budget-averaged difference.

Table 31: **Singleton and leave-one-out attribution orders.** Accuracy is averaged over exactly-k budgets one through five. Differences are singleton minus leave-one-out; intervals are paired question bootstraps.

domain	n	singleton	LOO	diff [95% CI]	up/down/tied
Omni-MATH-2	195	0.8533	0.8369	$+0.0164$ $[+0.0051, +0.0287]$	14/3/178
MaScQA	200	0.9380	0.9300	$+0.0080$ $[+0.0000, +0.0180]$	8/3/189
JEEBench	199	0.9286	0.9136	$+0.0151$ $[+0.0050, +0.0261]$	13/3/183
SciBench	199	0.7598	0.7558	$+0.0040$ $[-0.0080, +0.0161]$	12/10/177
LAB-Bench	182	0.7308	0.7198	$+0.0110$ $[-0.0110, +0.0319]$	26/19/137

The singleton–leave-one-out mean-difference sign-flip tests give Holm-adjusted $p = 0.0355$ on mathematics and 0.0376 on engineering entrance within the five-domain family. Exact sign tests of how often one method wins instead give 0.0636 and 0.0851. These test different aspects of a non-binary budget-averaged difference. The comparisons rest on 17 and 16 discordant questions, respectively. Pooling with the separate Shapley–Banzhaf comparisons into ten tests removes both sign-flip rejections.

Fully tied five-message scores occur on 80.5% of questions for singleton, 88.7% for leave-one-out, and 74.3% for Shapley and Banzhaf. The latter two have identical tie *status* on every question, not necessarily identical score vectors or orders; their subset agreement is 83–98.5% and correctness agreement 92–99% across domains. Every paired accuracy-difference interval for Shapley versus Banzhaf contains zero.

A retrospective stratification of singleton versus leave-one-out yields 52 up/0 down where leave-one-out is fully tied (865 questions), versus 21 up/38 down where it is not (110 questions). The latter sign-test $p = 0.036$ is exploratory. A constant-score control also demonstrates substantial dependence on the canonical tie rule. These findings do not establish either attribution method's general superiority. In particular, ties on constant-outcome tables reflect the absence of a correctness-based choice, rather than necessarily a lossy estimator.

E.10 Matched Cross-domain and Receiver Comparisons

Biology sample composition. The eligible biology pool is concentrated in sequence reasoning. A 200-question draw contains 151 sequence-reasoning, 40 protocol-reasoning and nine cloning questions. The corresponding pilot had 15, three and one of nineteen. Consequently the biology findings should not be read as representative of all LAB-Bench task types. The expanded Gemma analysis uses 182 complete questions; source-receiver completeness and matched intersections are given in Appendix E.9.

Matched outcomes under two receivers. All comparisons below use identical question and candidate-message texts, restricted to complete lattices under both receivers. The fixed selector does not observe receiver identity, so it chooses the same subset on each side. The first table compares correctness of that fixed choice; the second compares directional survival of each receiver's smallest successful choice, restricting to questions solvable under both. Exact McNemar tests are Holm-corrected separately within each four-domain family. These questions differ from whether either direction has high absolute survival.

Table 32: Fixed-selector accuracy under two receivers on matched questions and candidate pools.

domain	n	OSS acc	Gemma acc	diff	b (Gemma>OSS)	c (OSS>Gemma)	p_{Holm}
MaScQA	200	0.9100	0.9300	$+0.0200$	6	2	0.689
JEEBench	198	0.8990	0.8939	-0.0051	5	6	1.000
LAB-Bench	180	0.6111	0.6500	$+0.0389$	16	9	0.689
SciBench	188	0.8191	0.7553	-0.0638	3	15	**0.030 (survives)**

Only college science rejects the equal-accuracy null after correction. Paired 95% question-bootstrap intervals (20,000 resamples, seed 20260923) are $[-0.005, 0.050]$ for materials QA, $[-0.035, 0.025]$ for engineering entrance, $[-0.0167, 0.0944]$ for biology and $[-0.106, -0.021]$ for college science.

Table 33: Survival of each receiver's smallest successful choice on questions solvable under both receivers.

domain	n (solvable both)	OSS→Gemma survives	Gemma→OSS survives	b	c	p_{Holm}
MaScQA	192	185/192 (96.4%)	181/192 (94.3%)	10	6	0.909
JEEBench	191	178/191 (93.2%)	177/191 (92.7%)	7	6	1.000
LAB-Bench	160	121/160 (75.6%)	112/160 (70.0%)	23	14	0.563
SciBench	160	142/160 (88.8%)	152/160 (95.0%)	3	13	0.085

No directional contrast survives correction. The following alternative denominators answer a different descriptive question: among questions solvable under the source of a frozen choice, how often does that choice survive under the other receiver? Because each direction has its own denominator, the two rates are not a matched directional contrast.

Table 34: Directional choice survival using each source receiver's solvable-question denominator.

domain	OSS→Gemma / solvable-under-OSS	rate	Gemma→OSS / solvable-under-Gemma	rate
MaScQA	185/193	95.9%	181/198	91.4%
JEEBench	178/191	93.2%	177/195	90.8%
LAB-Bench	121/163	74.2%	112/164	68.3%
SciBench	142/165	86.1%	152/161	94.4%

Overlap of successful subsets. Per-question Jaccard overlap, averaged where the union is non-empty, is 0.918 $[0.884, 0.948]$ for materials QA, 0.899 $[0.863, 0.932]$ for engineering entrance, 0.848 $[0.803, 0.890]$ for college science and 0.633 $[0.576, 0.687]$ for biology. This measures overlap of sets of successes, not selection accuracy. The receiver collections differ in run conditions and date as well as model identity, and GPT-OSS also grades both conditions; within-receiver repeat evidence is reported separately.

Missingness. Materials QA has no matched-cohort exclusions. Engineering entrance loses one question per side. Biology has eighteen Gemma-side timeouts; their completed GPT-OSS cells average 0.443 accuracy versus 0.600 on the matched population. College science has eleven GPT-OSS-side evaluator-parse failures; their completed Gemma cells average 0.270 versus 0.755 on matched questions. These are associations in the observed sample, not estimates of the missing sample or proof of a bias direction.

E.10.1 Sensitivity to the excluded college-science question

One Gemma-side college-science question has $31/32$ outcomes. The unresolved evaluator verdict is for `scibench:diff:page344-14`, subset $\{m_1, m_2, m_3\}$ (canonical subset index seven). All 31 observed outcomes fail. The primary population therefore contains 199 of 200 intended questions. We consider both possible values of the missing cell:

The exclusion increases both reported endpoint accuracies by 0.0037 relative to the intended 200-question population, whichever value the unresolved cell takes. Its contribution to intermediate-only headroom and informative ranking pairs remains undetermined. This sensitivity does not impute the cell or change the primary analysis.

E.11 Smallest successful subsets

Section E.9.2 reports the training-population figures. The held-out split is broadly similar but not identical: 67.2% of its questions are answered correctly with no messages, 10.3% need one, 10.3% need two or more, and 12.1% are answered incorrectly under every subset. The one-message share is roughly half the training value and the two-or-more share roughly double, on 58 questions where one

Table 35: Sensitivity of college-science summaries to either possible grade for the unresolved subset.

quantity	if $U(\{m_1, m_2, m_3\}) = 0$	if $U(\{m_1, m_2, m_3\}) = 1$
bucket for this question	unreachable	**intermediate-only**
contributes informative pairs	no	**yes** (31 pairs, $< 0.5\%$ of the base)
unreachable count	$31/199 \to 32/200$	$31/199 \to 31/200$
intermediate-only count	$11/199 \to 11/200$	$11/199 \to \mathbf{12/200}$
informative-question count	62	**63**
receiver-only answering accuracy	$0.7337 \to 0.7300$	$0.7337 \to 0.7300$
reveal-all accuracy	$0.7387 \to 0.7350$	$0.7387 \to 0.7350$
adds an oracle-reachable question	no	**yes**

question moves a proportion by 1.7 points, so the two splits should be read as compatible in shape rather than as matching estimates.

E.11.1 Per-budget detail for the compact-representation gaps

Section E.9.3 reports question gain (question-specific order minus global order) and budget gain (budget-specific selection minus question-specific ordering) as an aggregate over exact sizes one through five. Table 36 gives the per-size values using the same paired bootstrap and seed.

Table 36: Per-size hindsight gains on the five primary Gemma cohorts. Question gain compares question-specific and global orderings; budget gain compares budget-specific selection with a question-specific ordering.

domain	k=1	k=2	k=3	k=4	k=5
question gain: question-specific order − global order					
Omni-MATH-2	+0.046	+0.036	+0.062	+0.031	+0.000
MaScQA	+0.040	+0.025	+0.010	+0.015	+0.000
JEEBench	+0.045	+0.030	+0.030	+0.025	+0.000
SciBench	+0.060	+0.035	+0.040	+0.045	+0.000
LAB-Bench	+0.170	+0.176	+0.165	+0.088	+0.000
budget gain: budget-specific selection − question-specific ordering					
Omni-MATH-2	+0.000	+0.005	+0.000	+0.005	+0.000
MaScQA	+0.000	+0.000	+0.000	+0.005	+0.000
JEEBench	+0.010	+0.000	+0.000	+0.000	+0.000
SciBench	+0.005	+0.005	+0.000	+0.000	+0.000
LAB-Bench	+0.011	+0.011	+0.000	+0.000	+0.000

Every domain reaches exactly 0.000 on both gaps at budget five, where all three rows converge to revealing every message by construction; this is a boundary identity, not an estimated null result, and is reported for completeness rather than as evidence. Biology's question gain is largest at budgets one through three and falls at budget four, so the aggregate figure in the main text is not driven by a single budget. No domain or budget shows the question-specific order beating the unrestricted oracle, which would indicate an error in either row; this non-negative-regret property was asserted and checked in the underlying computation rather than assumed.

E.12 Interactions between messages

Interactions between messages are real in this data and run in both directions. Over the 1,154 training questions we compare each message pair against the better of its two members alone. On 166 questions (14.4%) at least one pair strictly outperforms both of its members, a complementary effect in the recorded pair outcomes. As Section 5 shows, an *outcome-aware* policy restricted to one order per question recovers nearly all achievable accuracy despite them. That is a statement about a restricted class of oracle policies on this data, and it does not show that interactions are unimportant to a selector that has to learn the ordering.

F Training Details and Additional Selector Results

This appendix separates how a selector is trained from what its choices accomplish. A subset scorer assigns a score to an entire candidate combination. A pairwise selector learns by comparing two combinations. A sequential selector instead chooses one message at a time and can stop; the resulting sequence of choices is called a rollout. A checkpoint is a saved version of the trained model. The untrained reference uses the corresponding model before selector training. A rank is a position in the score ordering, with rank one best. An endpoint rank gap compares the average ranks at two subset sizes, rather than the numerical scores themselves.

F.1 Selector Training and Evaluation

F.1.1 Shared setup and information access

The selectors use Qwen3-4B-Instruct-2507 (Qwen Team, 2025), revision `cdbee75f17c01a7cc42f958dc650907174af0554`, with LoRA (Hu et al., 2022): rank 16, scaling 32, dropout 0.05, and bf16 computation. Only adapter weights are trained. The original subset scorer and sequential selector use seed 20260908. An additional subset-scorer run uses seed 20260925; its results and unequal training budget are reported in Appendix A.1. Other configurations have one run each.

Inputs contain the question, candidate texts and selection state. Reference answers, recorded utilities, proposal correctness and dataset identifiers are excluded. Candidate handles identify actions. The selector can read messages it does not forward. The receiver answers only after selection ends, in canonical message order; it does not provide intermediate feedback to the selector.

Model-specific development splits. The subset scorer uses 1,154 source-training questions and the separate 19-question development set for monitoring binary cross-entropy. That development set also selects the decision-time cost penalty. The correctness-only and cost-aware pairwise selectors instead use 1,009 questions for training and a 145-question development carve-out of the 1,154-question pool. Their checkpoint metric is development pairwise accuracy. The sequential gate is selected by balanced binary negative log-likelihood on the source-development states; its ranker uses the run-designated checkpoint described below. None uses the 58 held-out questions, target-domain outcomes or Gemma outcomes for fitting or selection.

Table 37: **Selector configurations.** Each row states the information available to the selector, its training signal and its decision rule.

Configuration	Information available	Training signal and decision rule
Trained subset scorer	Question + all five messages	Subset correctness; choose the highest-scoring feasible subset.
Untrained subset scorer	Same	No supervision (initialization); same decision rule.
Original pairwise-contrastive	Same	Ordered subset pairs; choose the highest-scoring feasible subset.
Correctness-only pairwise	Same	Correctness-labeled pairs; same decision rule.
Cost-aware pairwise	Same	Pairs labeled by correctness and subset size; same decision rule.
Message-content-blind control	Question + message *count* only	Correctness-labeled pairs; same decision rule.
Sequential, trained ranker	Question + candidates + selected set	Separately trained ranker and gate; reveal until the gate stops.
Sequential, untrained ranker	Same	Trained gate with ranker at initialization; same rollout.

F.1.2 SUBSET-SCORING SELECTOR

For a candidate subset S, the score is the logit difference between the answer tokens "Yes" and "No". Binary cross-entropy supervises that score against $U_g(S; q)$. All 36,928 source-training cells have unit weight: 22,578 positive and 14,350 negative labels. Successful subsets receive the same target regardless of message count; there is no cost term. The prompt exposes subset size, so size is available as a predictor.

Training uses AdamW with constant learning rate 10^{-4}, micro-batch one, gradient accumulation eight and development monitoring every 250 steps. Both primary-scorer runs use the same 19-question development set and select the lowest development BCE reached. The original run completes its search through update 5,250 and selects update 2,750 (BCE 0.42764). The added run stops at a 5.5-hour time limit, after 11 development evaluations, and selects its last evaluated update, also 2,750 (BCE 0.50097). Equal selected steps therefore do not imply equal search budgets. The added run ends by the time limit, not early stopping.

The untrained reference is the original run's initialization. At inference, each model scores all 32 subsets and chooses the highest-scoring feasible one, breaking ties toward fewer messages, then canonical subset index. Binary cross-entropy measures probability prediction, whereas deployment depends on the selected subset. Figure 10 shows the added run's development history; held-out outcomes for both runs appear in Appendix A.1.

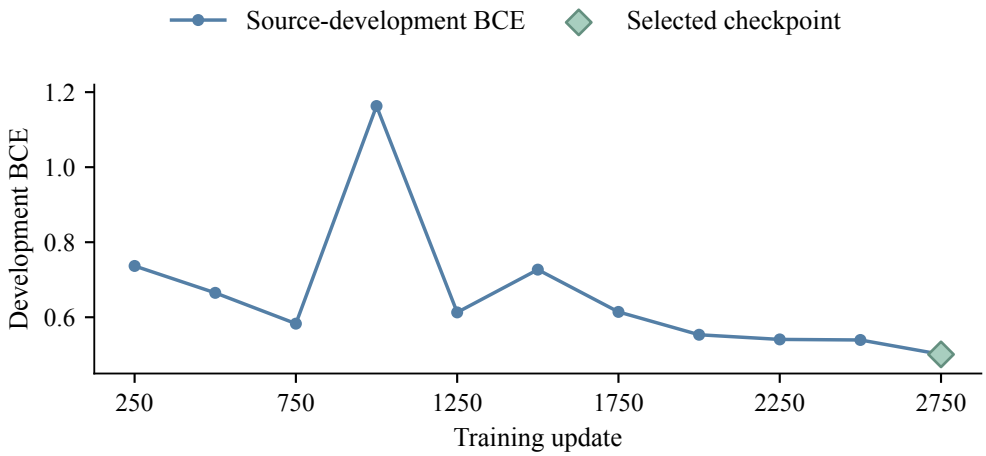

Figure 10: **Development history of the added subset-scorer run.** Source-development binary cross-entropy (BCE) at all 11 recorded evaluations for seed 20260925. The mint diamond marks the frozen checkpoint at update 2,750, the minimum reached before the 5.5-hour time limit. It is a marker on the blue loss curve, not a separate metric. Values follow the delivered training log, rounded to four decimals except the selected endpoint. Development improvement is distinct from held-out selection performance.

F.1.3 PAIRWISE SELECTORS AND CONTENT CONTROL

The original pairwise-contrastive selector uses ordered success-versus-failure subset pairs. Its 77,258 pairs contain no success-versus-success cost comparisons. It is a separate configuration from the correctness-only selector of the cost-supervision study.

In this comparison, "cost-aware" refers to a preference for fewer messages reaching the receiver, rather than a measured saving in total system resources. Both selectors use 133,372 pair presentations. The correctness-only selector presents its 66,686 distinct correctness pairs twice; the cost-aware selector presents them once plus 66,686 pairs preferring the smaller of two successful subsets. The selectors therefore differ in correctness-example exposure as well as cost supervision. Development has 10,572 success-versus-failure pairs and no cost pairs. Early stopping uses pairwise cross-entropy with patience four; checkpoint selection maximizes pairwise accuracy, with ties to the earlier checkpoint.

The correctness-only selector runs to step 1,125 and selects 375; the cost-aware selector runs to 625 and selects 125. Both have a maximum budget of 1,500 steps. These are not matched-compute checkpoints. At the selected updates, the archived development pairwise accuracies are 0.575 for correctness-only and 0.479 for cost-aware training, compared with a 0.5 chance reference on strict pairs. The latter checkpoint therefore does not demonstrate successful source-development discrimination; its communication behavior should not be taken as evidence of improved ranking.

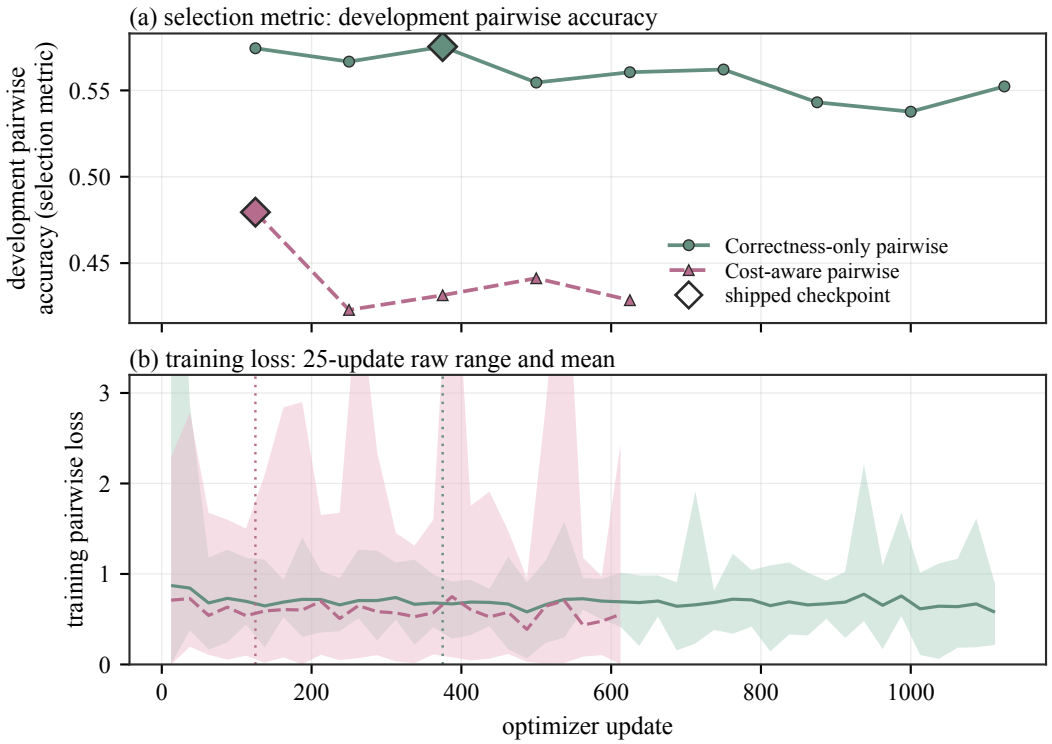

Figure 11: **Pairwise-selector training histories.** (a) Development pairwise accuracy, the checkpoint-selection metric; diamonds mark update 375 for correctness-only training and update 125 for cost-aware training. (b) Training pairwise loss, displayed as the raw minimum–maximum range and mean within each 25-update block. The two loss heights are not directly comparable because the models see different pair sets and token exposure. Color, marker and line style jointly identify the two configurations.

The message-content-blind control retains the question and subset-size information while blanking message content. It uses the correctness-only selector's architecture and training pairs. Its results test changes induced by removing message content, not removal of the question itself. Adapter identities are in Appendix F.2.9.

F.1.4 Sequential selection and stopping

The sequential selector chooses one remaining message or STOP at each step. A separate trained gate decides whether to continue. Two configurations share that gate and use either the trained ranker (step 3,200) or its untrained initialization. The receiver is called once after selection terminates.

The immediate supervision compares each reveal with stopping:

$$A_{\text{imm}}(m \mid S, q) = \Delta_g(m \mid S, q), \qquad A_{\text{imm}}(\text{STOP} \mid S, q) = 0. \tag{5}$$

The gate has a single continue target when some reveal has positive immediate advantage, and a stop target otherwise. The ranker compares only actions with strictly different advantages; tied actions do not form training pairs. The gate's continue target is therefore myopic. A combination can succeed even if no single addition does. On the held-out empty states it targets stop on $52/58$ questions, including six with a successful larger subset.

The gate uses class-weighted binary cross-entropy, with continue and stop weights 5.020208 and 0.555307. The ranker uses pairwise logistic loss: within each state, it averages $\log(1 + \exp[-(s_i - s_j)])$ over message pairs with $A_{\text{imm}}(i) > A_{\text{imm}}(j)$, where s_i is the score for message i at that state. States without a strict pair contribute zero ranking loss; these comprise $30{,}599$ of $35{,}774$ training states (85.5%). The ranker scores remaining-message tokens in one forward pass at each state; the gate separately scores continue and stop. Neither uses a joint listwise action loss.

The component records specify AdamW at learning rate 10^{-4}, micro-batch one, gradient accumulation eight and sequence length $8{,}192$. The recorded sequential LoRA configuration targets the query, key, value and output attention projections and the gate, up and down feed-forward projections. The gate stops after $6{,}750$ updates and selects step $3{,}250$ by development balanced binary negative log-likelihood. The ranker stops after $5{,}750$ updates. Its evaluated step-$3{,}200$ adapter is run-designated, not selected by the intended terminal-accuracy rule: the recorded source-development rollouts all stopped immediately and could not distinguish checkpoints. No held-out result is used to choose it. This limits interpretation of the sequential comparison as a test of a well-selected policy.

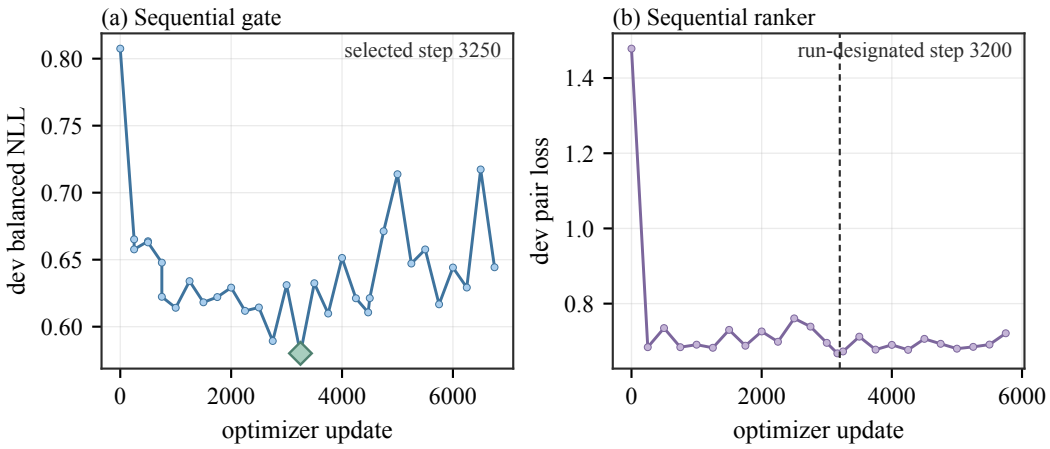

Figure 12: **Sequential gate and ranker development histories.** (a) Gate class-balanced negative log-likelihood (NLL). The mint-green diamond marks the selected gate checkpoint at training update $3{,}250$; blue circles show the logged development evaluations. (b) Ranker strict-pair logistic loss. The gate and ranker use different objectives and separate vertical axes. Ranker update $3{,}200$ was run-designated rather than selected by an executed validation rule and is not itself a logged evaluation point, so it appears as a dashed reference without a marker.

The evaluated rollouts use the remaining budget in the prompt consistently with the external budget cap. Expanded-cohort outcomes are reported in Appendix F.4.

F.1.5 Inference rules and costs

Inference ties prefer fewer messages, then the lower subset index. This differs from the utility-reference ranking, which averages uniformly over ties, and from supervision, which retains admissible tied actions. On expanded mathematics and materials QA, changing exact score ties from largest to smallest subset reduces mean messages by only about 0.10; it does not account for the several-message excess over the smallest recorded successful subset.

The subset scorer submits 32 inputs per question. Each includes the question and all candidate texts, so input length does not decrease when the selected subset is smaller. The following counts use the selector's tokenizer and full templated prompts. Tokens per question are medians of question-level sums; tokens per input are medians over individual inputs, so one is not necessarily 32 times the other.

Inputs can be batched, so their count is not a model-forward-call count. These figures exclude output tokens, generation of the candidate pool, runtime and energy. The sequential selector uses at most six decision inputs per question. Receiver-visible message count is the primary communication measure; it does not imply total-compute savings.

Table 38: Selector input-token accounting on primary cohorts. Medians exclude candidate generation and output tokens.

domain	questions	median tokens/question	median tokens/input
MaScQA	200	26,464	821
SciBench	199	31,720	990
JEEBench	199	32,536	1,019
Omni-MATH-2	195	33,928	1,027
LAB-Bench	182	46,352	1,462
pooled	975	32,248	1,012

Evaluation contract. All test choices are made without recorded outcomes. They are scored by looking up the selected subset in the matching question, candidate-pool, receiver and prompt record. That lookup is an evaluation of the recorded outcome, not an independent serving replay. Question-paired comparisons use the same feasible sets and report actual communication alongside accuracy. Repeated receiver outcomes and selector-seed variation are separate sources of uncertainty.

F.2 Extended Selector Analyses

F.2.1 Reference ranks and size preference

For mean rank $\bar{r}(k)$ among size-k subsets, with rank one best, define

$$D_{\text{total}} = \bar{r}(0) - \bar{r}(5), \quad D_{\text{empty}} = \bar{r}(0) - \bar{r}(1), \quad D_{\text{nonempty}} = \bar{r}(1) - \bar{r}(5). \tag{6}$$

The components add exactly. They are endpoint differences, not slopes, and do not identify separate modules within the selector. Positive values indicate a preference for larger subsets. The comparison reference ranks all recorded successes above failures and averages uniformly over ties. With $N = 32$ and k successful subsets, expected ranks are $(k+1)/2$ for a success and $(N+k+1)/2$ for a failure. Therefore, per question,

$$\mathbb{E}[D_{\text{total}}] = 16\,[U(\mathcal{M}) - U(\varnothing)]. \tag{7}$$

The reference total gap and endpoint accuracy lift are the same quantity in different units. Alternative utility-consistent tie conventions can have different size profiles.

Table 39 compares the score ranks assigned to the endpoint subset sizes.

Table 39: **Utility-consistent reference rank gaps.** Uniform random tie-breaking within successful and unsuccessful subsets; expected ranks computed exactly. Parentheses are question-level standard errors. The total gap equals 16 times the full-minus-empty accuracy difference, so these are two units of one quantity. R denotes receiver shift; R+D denotes receiver and domain shift. The two test-mathematics cohorts overlap.

				reference endpoint rank gaps (ranks)			label lift
Domain	split	shift	N	D_{total}	D_{empty}	$D_{\text{non-empty}}$	5 minus 0
no shift on either axis — Omni-MATH-2 under the source receiver							
Omni-MATH-2	held out	—	58	+0.83 (0.72)	−0.44 (0.58)	+1.27 (0.51)	+0.052
Omni-MATH-2	development	—	19	+0.84 (0.82)	+0.67 (1.05)	+0.17 (0.84)	+0.053
receiver shifted, domain held							
Omni-MATH-2	expanded	R	195	+0.33 (0.42)	−0.16 (0.33)	+0.49 (0.21)	+0.021
receiver and domain both shifted							
MaScQA	expanded	R+D	200	−0.08 (0.33)	+0.13 (0.26)	−0.21 (0.17)	−0.005
LAB-Bench	expanded	R+D	182	−0.62 (0.58)	**−1.30** (0.47)	+0.69 (0.36)	−0.038
JEEBench	expanded	R+D	199	+0.08 (0.37)	−0.11 (0.31)	+0.19 (0.19)	+0.005
SciBench	expanded	R+D	199	+0.08 (0.37)	−0.13 (0.34)	+0.21 (0.25)	+0.005

Matched trained–untrained comparisons. A selector's scores depend on the question and candidate messages, not receiver outcomes. A receiver change can alter the reference, but cannot alter the

fixed selector's scores on identical inputs. Expanded versus held-out mathematics also changes the question mixture. Table 50 reports raw gaps and differences each against its reference on the same questions.

The trained-minus-untrained non-empty gap is 10.35 ranks $[7.95, 12.63]$ on expanded mathematics, 11.62 $[8.82, 14.27]$ on its disjoint 137 questions, and 11.82 $[10.31, 13.30]$ on materials QA. These are paired question-bootstrap 95% intervals. The fitted materials-QA curve decreases from mean rank 29.8 at size zero to 5.1 at size five. Its size preference exceeds its initialization's, but the base selector itself also favours larger non-empty subsets there.

The trained empty-versus-singleton difference from the reference remains unresolved on expanded mathematics: 0.63 $[-0.74, 1.95]$ on 195 questions. The untrained non-empty difference on expanded mathematics is also unresolved, 1.11 $[-0.81, 3.03]$. None implies indifference.

Full size profiles. Endpoint differences need not imply a monotone preference. The following mean ranks at sizes zero through five show the reference profiles and the available corresponding selector profiles. Rank one is best. Source mathematics uses GPT-OSS; expanded cohorts use Gemma. These curves are descriptive; paired endpoint comparisons above carry the uncertainty estimates.

Table 40: Mean utility-reference rank by subset size, with rank one best.

reference, mean rank of 32	$\|S\|$=0	1	2	3	4	5
Mathematics, held out (N=58)	16.53	16.98	16.42	16.45	16.42	15.71
Mathematics, expanded (N=195)	16.57	16.73	16.53	16.48	16.30	16.24
MaScQA (N=200)	16.48	16.35	16.50	16.55	16.54	16.56
LAB-Bench (N=182)	15.62	16.92	16.60	16.36	16.38	16.24

Table 41: Mean selector rank by subset size for the reported training configurations and populations.

policy	split	$\|S\|$=0	1	2	3	4	5
Trained	MaScQA, n=200	29.8	26.1	19.7	12.9	7.3	5.1
Untrained	MaScQA, n=200	31.2	18.2	18.3	16.0	10.8	8.9

F.2.2 Successful alternatives and excess communication

A size preference can occur among subsets that all succeed without misordering successes and failures. The following comparison separates an endpoint size gap recomputed within successful subsets from the success–failure boundary gap. Positive boundary gap means failures rank below successes.

Table 42 separates size preference among successful subsets from the ranking of successes against failures.

Table 42: **The size preference operates among subsets that all succeed.** The within-success comparison is the endpoint size gap recomputed over each question's succeeding subsets only. The success–failure comparison is the mean rank of failing subsets minus succeeding ones — whether the selector separates winners from losers at all. Brackets are 95% question bootstraps; n is the number of questions contributing. The math rows are the 195-question expanded set, which contains all 58 held-out questions.

Policy	set	Within successes	Success–failure
Trained	Omni-MATH-2, n=195	$+\mathbf{14.22}$ $[+11.52, +16.86]$ (n=175)	$+2.13$ $[-0.25, +4.47]$ (n=46)
Trained	MaScQA, n=200	$+\mathbf{24.66}$ $[+22.86, +26.30]$ (n=195)	$+1.16$ $[-2.38, +4.69]$ (n=31)
Untrained	MaScQA, n=200	$+21.50$ $[+20.04, +22.89]$ (n=195)	-0.73 $[-3.84, +2.50]$ (n=31)

Table 43 relates successful alternatives to the additional messages selected.

Table 44 gives the full-cohort paired accuracy and communication comparisons.

Table 43: **Untrained subset-scorer decisions on solvable questions.** Gemma expanded cohorts. Wrong: selected subset fails; costly: succeeds with more messages than the smallest recorded success; minimal: succeeds at that smallest size. Excess is mean selected size minus the smallest-success size on the same solvable questions, with 95% question-bootstrap intervals.

Policy	set	solv.	WRONG	COSTLY	MIN.	msgs	orac.	excess
Untrained	Omni-MATH-2, 195	177	20 (11.3%)	153 (86.4%)	4 (2.3%)	2.59	0.11	$+2.47$ $[+2.30, +2.65]$
Untrained	MaScQA, 200	198	9 (4.5%)	186 (93.9%)	3 (1.5%)	3.43	0.07	$+3.37$ $[+3.19, +3.55]$

Table 44: **Trained subset-scorer decisions on five expanded Gemma cohorts.** Categories and excess communication use solvable questions; accuracy differences use all questions. Successful-subset count is averaged over all complete questions. Intervals are paired question-bootstrap 95% intervals. The mathematics cohort contains the 58 held-out questions. All excess-message intervals exclude zero; none of these accuracy difference intervals does.

Domain	succ.	solv. / n	WRONG	COSTLY	MIN.	excess messages
LAB-Bench	20.9	166 / 182	**47 (28.3%)**	108 (65.1%)	11 (6.6%)	$+2.19$ $[+1.93, +2.44]$
SciBench	23.5	168 / 199	24 (14.3%)	143 (85.1%)	1 (0.6%)	$+3.37$ $[+3.17, +3.56]$
Mathematics*	26.4	177 / 195	17 (9.6%)	145 (81.9%)	15 (8.5%)	$+2.86$ $[+2.63, +3.10]$
JEEBench	29.1	196 / 199	18 (9.2%)	176 (89.8%)	2 (1.0%)	$+3.70$ $[+3.53, +3.86]$
MaScQA	29.6	198 / 200	12 (6.1%)	183 (92.4%)	3 (1.5%)	$+4.03$ $[+3.84, +4.20]$

paired accuracy differences on the same questions — every interval contains zero

Domain	vs revealing everything	vs receiver-only answering
LAB-Bench	-0.0165 $[-0.0769, +0.0440]$	$\mathbf{-0.0549}$ $[-0.1264, +0.0165]$
SciBench	-0.0151 $[-0.0402, +0.0101]$	-0.0101 $[-0.0603, +0.0402]$
Mathematics*	-0.0205 $[-0.0615, +0.0205]$	$+0.0000$ $[-0.0410, +0.0410]$
JEEBench	-0.0201 $[-0.0452, +0.0000]$	-0.0151 $[-0.0603, +0.0302]$
MaScQA	$+0.0100$ $[-0.0150, +0.0350]$	$+0.0050$ $[-0.0350, +0.0400]$

* the expanded mathematics set: same domain as training, shifted receiver. The other four shift both axes.

The categories are mutually exclusive over questions with at least one recorded success. Unsolvable questions are excluded from these category rates, but remain in accuracy comparisons. Excess communication uses the same solvable population for the selected and smallest-success sizes. It is relative to a hindsight reference on one replay, not a measured minimum communication requirement under repeated serving.

F.2.3 DIRECT RANKING QUALITY

Concordance asks whether scores rank successful subsets above unsuccessful ones. It uses only mixed-outcome questions and weights questions equally, rather than allowing domains with many informative pairs to dominate. Top-choice regret additionally measures the recorded outcome missed by the chosen subset.

Paired trained-minus-untrained concordance is 0.1192 $[0.0344, 0.2052]$ on expanded mathematics, 0.0997 $[0.0508, 0.1511]$ on biology, 0.0657 $[-0.0045, 0.1379]$ on college science, 0.0580 $[-0.0671, 0.1841]$ on materials QA and 0.0536 $[-0.0328, 0.1400]$ on engineering entrance. All point estimates are positive; two pointwise intervals exclude zero. Absolute trained concordance remains near chance on these expanded cohorts. Materials QA has only 31 mixed-outcome questions despite its 200-question cohort, illustrating the resolution lost to constant-outcome tables.

On the 137 mathematics questions outside the held-out set, trained concordance is 0.6132 $[0.5127, 0.7152]$. This estimate differs from the full-cohort value, illustrating population dependence. Score dispersion alone cannot establish ordering quality; widely separated scores can still have the wrong order.

Table 45: **Direct ordering quality.** Concordance averages, per mixed-outcome question, the fraction of success–failure subset pairs ordered correctly, counting score ties as one half. Chance is 0.5. Regret is recorded best utility minus selected utility, averaged over all questions. Intervals are question-bootstrap 95% intervals. All rows use the Gemma receiver.

Policy	set	n	mixed q	pairs	concordance	top-1 regret
Trained	Omni-MATH-2 expanded	195	46	6,078	$0.5696\ [0.4947, 0.6442]$	0.087
Untrained	Omni-MATH-2 expanded	195	46	6,078	$0.4505\ [0.3816, 0.5220]$	0.103
Trained	MaScQA	200	31	3,494	$0.5350\ [0.4250, 0.6468]$	0.060
Untrained	MaScQA	200	31	3,494	$0.4770\ [0.3792, 0.5777]$	0.045
Trained	LAB-Bench	182	100	16,229	$0.5166\ [0.4657, 0.5678]$	0.258
Untrained	LAB-Bench	182	100	16,229	$0.4169\ [0.3735, 0.4610]$	0.275
Trained	SciBench	199	62	6,880	$0.5624\ [0.4934, 0.6307]$	0.121
Untrained	SciBench	199	62	6,880	$0.4968\ [0.4318, 0.5619]$	0.116
Trained	JEEBench	199	40	4,505	$0.5553\ [0.4647, 0.6454]$	0.090
Untrained	JEEBench	199	40	4,505	$0.5017\ [0.4087, 0.5937]$	0.101

F.2.4 Held-out pairwise selector outcomes

Appendix A.6 reports the three identified pairwise checkpoints on the same 58 held-out questions under source and transfer receivers. Their accuracy differences from receiver-only answering remain unresolved. The expanded-cohort comparisons below cover selection behavior across budgets.

F.2.5 Decision-time communication penalties

A descriptive sweep chooses $\arg\max_S \sigma(s(S)) - \lambda|S|$ from the subset scorer's existing scores. It uses no additional training or receiver calls. The table uses at-most-five choices on expanded Gemma cohorts. The expanded mathematics set contains the held-out question IDs.

Table 46: Decision-time message-penalty sweep. Accuracy and mean messages use an at-most-five budget.

	Omni-MATH-2 expanded, 195		MaScQA, 200	
λ	accuracy	messages	accuracy	messages
0 (the rule used above)	0.8205	2.87	0.9300	4.08
0.02	0.8205	1.35	0.9350	1.84
0.05	0.8256	0.79	0.9300	1.03
0.1	0.8256	0.47	0.9300	0.80
0.2	0.8256	0.19	0.9250	0.26

The frozen development rule selects $\lambda = 0.002$ on 19 questions, rather than one of the larger displayed penalties. It saves only 0.14–0.32 messages across the reported held-out settings, with unchanged recorded accuracy. Other points describe a sweep over test outcomes and are not validation-selected operating points. The $\lambda = 0.05$ point reduces messages from 2.87 to 0.79 on expanded mathematics and from 4.08 to 1.03 on materials QA, with accuracy differences 0.005 and 0.000. These small differences do not establish equivalent accuracy.

In several domains, increasing the penalty approaches the receiver-only baseline, which already matches or exceeds the unpenalized selector at the point estimate. Therefore a large communication reduction alone does not establish improved content selection. Selector inference overhead is reported in Appendix F.1 and is excluded from receiver-visible message count.

F.2.6 Cost-aware supervision

The correctness-only pairwise selector ranks successes over failures; the cost-aware selector also ranks smaller successes over larger successes. Their training, development criterion and unequal

exposure are specified in Appendix F.1. The following results compare the resulting checkpoints; they do not isolate an objective change at matched training compute.

Table 47 reports the evaluated cost-aware and correctness-only configurations.

Table 47: **Correctness-only and cost-aware pairwise selectors.** Accuracy at budgets one, three and five; paired differences cost-aware minus correctness-only. Message differences are at budget five. Mathematics held out uses GPT-OSS; expanded cohorts use Gemma. The pooled 1,033 rows include 58 repeated question identities across receivers. The selectors differ in training exposure and selected step (Appendix F.1).

Budget 1 Condition	Correctness-only	Cost-aware	Δ accuracy
Omni-MATH-2, held out (n=58)	0.690	0.621	−0.0690*
Omni-MATH-2, expanded (n=195)	0.810	0.805	−0.0051
MaScQA (n=200)	0.935	0.920	−0.0150
LAB-Bench (n=182)	0.643	0.626	−0.0165
JEEBench (n=199)	0.899	0.920	+0.0201
SciBench (n=199)	0.744	0.749	+0.0050
pooled (n=1,033)	0.8025	0.7967	−0.0058

Budget 3 Condition	Correctness-only	Cost-aware	Δ accuracy
Omni-MATH-2, held out (n=58)	0.690	0.672	−0.0172
Omni-MATH-2, expanded (n=195)	0.831	0.831	+0.0000
MaScQA (n=200)	0.910	0.915	+0.0050
LAB-Bench (n=182)	0.676	0.643	−0.0330
JEEBench (n=199)	0.910	0.915	+0.0050
SciBench (n=199)	0.749	0.734	−0.0151
pooled (n=1,033)	0.8103	0.8025	−0.0077

Budget 5 Condition	Correctness-only	Cost-aware	Δ accuracy	Δ messages
Omni-MATH-2, held out (n=58)	0.724	0.672	−0.0517	−3.24
Omni-MATH-2, expanded (n=195)	0.841	0.831	−0.0103	−3.16
MaScQA (n=200)	0.920	0.915	−0.0050	−3.08
LAB-Bench (n=182)	0.714	0.643	−0.0714	−2.69
JEEBench (n=199)	0.899	0.915	+0.0151	−2.88
SciBench (n=199)	0.739	0.734	−0.0050	−2.99
pooled (n=1,033)	0.8190	0.8025	−0.0165	−2.98

* Held-out mathematics at budget one has bootstrap interval $[-0.138, -0.017]$ and exact raw $p = 0.125$ on few discordant questions. The smallest raw p among thirty cells is biology at budget five (0.0294); none survives Holm or Benjamini–Hochberg correction. Pooled budget-five accuracy difference is -0.0165 $[-0.0339, +0.0010]$; message difference is -2.982 $[-3.031, -2.931]$.

The six conditions contain 975 distinct question identities in 1,033 rows, because 58 mathematics questions appear under both receivers. Dropping their duplicated source-receiver rows yields a pooled budget-five accuracy difference of -0.014 $[-0.033, 0.004]$, consistent with the unresolved full-pool contrast. Per-condition results avoid treating the duplicated rows as independent replications.

At budget five, the cost-aware selector rescues 36 rows the correctness-only selector fails and harms 53 it succeeds on. Among the selectors' 187 and 204 failures, respectively, 110 and 127 have a successful alternative within budget. Failures recoverable only by sending more rise from 1 to 43; those recoverable only by sending less fall from 86 to 19. Categories allowing both directions overlap and are not summed. Cost supervision is associated with a different failure profile, not elimination of selection errors.

The biology budget-five contrast (9 up, 22 down, 151 tied) also appears in the 20-comparison panel in Appendix F.4.1. Its raw exact $p = 0.0294$ is the same; Holm adjustment gives 0.883 in the thirty-cell family and 0.589 in the twenty-cell family. Neither rejects at 0.05, and these are not independent findings.

Decision-time near-ties. A separate descriptive rule uses the correctness-only selector's raw scores: among subsets within ε of the maximum, choose the smallest. The grid $\{0, 0.25, 0.5, 1, 2\}$ is fixed before inspecting outcomes. At budget five, reported (accuracy, messages) pairs are $(0.8190, 4.13)$ at 0, $(0.8025, 1.52)$ at 0.25, $(0.7890, 0.98)$ at 0.5, and $(0.8141, 0.02)$ at 1. The cost-aware selector gives $(0.8025, 1.15)$. Equal rounded accuracy at two points does not establish equivalent expected accuracy. The comparison provides no resolved advantage of cost supervision over a decision-time rule. Across these same rows, receiver-only accuracy is 0.8132, reveal-all 0.8141, and recorded best-subset accuracy 0.9255.

Size versus content discrimination. On cross-size pairs, the cost-aware selector prefers the smaller subset in 89.4% of cases, versus 8.0% for the correctness-only selector. These are score-order properties, not proof of reliance on a particular input field. Holding size fixed, cost-aware minus correctness-only accuracy is -0.010 $[-0.102, 0.081]$ at size one, 0.015 $[-0.064, 0.093]$ at size two and 0.029 $[-0.058, 0.117]$ at size three, on 197, 204 and 171 informative questions. Neither selector resolves an advantage over the corresponding per-question random-choice success rate at these sizes.

F.2.7 Message-content-blind control

The control retains question and subset-size information while removing message content. The table compares the message-blind control with the correctness-only selector at budget five. τ_b measures agreement between their subset orderings; top-choice agreement concerns selected subsets, not correctness. Exact McNemar tests are pointwise.

Table 48: Message-blind versus correctness-only selection at budget five. Agreement of choices and agreement of correctness are distinct.

set	n	τ_b (order)	Top choice	Full acc.	Blind acc.	Δ msgs	p_{McNemar}
Omni-MATH-2 (expanded)	195	$+0.725$	41.0%	0.8410	0.8359	$+0.31$	1.000
MaScQA	200	$+0.636$	27.5%	0.9200	0.9200	$+0.60$	1.000
JEEBench	199	$+0.643$	20.1%	0.8995	0.9146	$+0.77$	0.375
SciBench	199	$+0.674$	31.2%	0.7387	0.7387	$+0.53$	1.000
LAB-Bench	182	$+0.622$	14.8%	0.7143	0.6703	$+0.70$	0.185
Omni-MATH-2, held out	58	$+0.714$	50.0%	0.7241	0.7069	$+0.05$	1.000
Omni-MATH-2, development	19	$+0.743$	47.4%	0.5789	0.5789	-0.05	1.000

None of the seven accuracy contrasts rejects at 0.05. Message differences on the five expanded domains are positive with paired question-bootstrap intervals excluding zero: 0.313 $[0.185, 0.441]$ on mathematics, 0.600 $[0.485, 0.715]$ on materials QA, 0.769 $[0.638, 0.899]$ on engineering, 0.528 $[0.402, 0.653]$ on college science, and 0.703 $[0.544, 0.857]$ on biology ($10{,}000$ resamples, seed 20260923). Materials QA has identical correctness on every question despite different choices on $145/200$. This documents concealed behavioural differences, not equivalence of policies or evidence that content has no value.

A post-hoc choice-diversity check finds five distinct selected subsets for the blind selector on materials QA versus sixteen for the content-aware selector; full-set selection is 79.5% versus 30.0%. Across expanded domains, the counts are 5–6 versus 16–20 distinct choices. Same-size comparisons have exact score ties on 59.1–66.0% for the blind selector versus 7.4–10.8% for the content-aware selector. Removing content reduces choice diversity and increases communication; unchanged accuracy does not establish that content is unused.

F.2.8 Outcome-aware restrictions on source mathematics

On the source-held-out set, the best globally fixed order loses 8.6, 10.3, 13.8 and 10.3 accuracy points to the exactly-k optimum at budgets one through four. On the $1{,}154$ source-training questions those losses are 15.9, 14.9, 12.9 and 11.0 points. Allowing a separate order per question greatly reduces the loss, with remaining source-training gaps of 0.6–1.6 points. These complement the expanded Gemma analysis and remain hindsight comparisons. An order obtained from complete outcomes can encode interactions; it is not evidence that learning may ignore them.

F.2.9 EVALUATED CONFIGURATION IDENTITIES

Table 49: Evaluated checkpoints and corresponding initialization references.

Configuration	Evaluated checkpoint
Subset scorer	step 2,750
Untrained subset scorer	matched initialization
Correctness-only	step 375
Cost-aware	step 125
Message-content-blind control	step 250
Original pairwise-contrastive	step 1,250
Sequential, trained ranker	ranker step 3,200; trained gate
Sequential, untrained ranker	ranker initialization; same trained gate

The six scorers choose by argmax over feasible subsets. The two sequential configurations choose through a rollout. Shared initialization weights do not imply the same policy when the inference rules differ.

F.3 SUPPORTING RANK COMPARISONS

Table 50 separates preference for larger subsets from ordering quality. It compares endpoint ranks against a matched reference, complementing the success–failure concordance in Table 45.

Table 50: **Matched comparison: each selector minus its own reference, same questions, same component.** The reference is the utility-consistent ranking under uniform random tie-breaking (Appendix F.2.1), computed exactly on the same questions and differenced question by question before bootstrapping (20,000 resamples over questions). Seed 20260923. Intervals containing zero do not resolve the sign of the difference.

Component	Selector	Reference	Difference (95% CI)
Trained · Omni-MATH-2 · receiver shift · $n = 195$			
Total	$+12.42$	$+0.33$	$+12.09$ $[+9.33, +14.72]$
Empty	$+0.46$	-0.16	$+0.63$ $[-0.74, +1.95]$
Non-empty	$+11.95$	$+0.49$	$+11.46$ $[+9.25, +13.60]$
Untrained · Omni-MATH-2 · receiver shift · $n = 195$			
Total	$+17.01$	$+0.33$	$+16.68$ $[+15.10, +18.19]$
Empty	$+15.40$	-0.16	$+15.56$ $[+14.26, +16.91]$
Non-empty	$+1.61$	$+0.49$	$+1.11$ $[-0.81, +3.03]$
Trained · MaScQA · receiver and domain shifts · $n = 200$			
Total	$+24.75$	-0.08	$+24.84$ $[+22.98, +26.54]$
Empty	$+3.71$	$+0.13$	$+3.58$ $[+2.78, +4.40]$
Non-empty	$+21.04$	-0.21	$\mathbf{+21.25}$ $[+19.45, +22.90]$
Untrained · MaScQA · receiver and domain shifts · $n = 200$			
Total	$+22.21$	-0.08	$+22.30$ $[+20.84, +23.69]$
Empty	$+12.99$	$+0.13$	$+12.86$ $[+11.83, +13.93]$
Non-empty	$+9.22$	-0.21	$+9.43$ $[+7.74, +11.06]$

F.4 FULL BUDGET RESULTS

The following tables report the expanded Gemma cohorts at each budget. The matched source-receiver comparison appears in Appendix A.6.

F.4.1 EXPANDED RESULTS AT EVERY BUDGET

The five expanded cohorts contain 195, 200, 199, 199 and 182 questions for mathematics, materials QA, engineering entrance, college science and biology. All use Gemma outcomes. The action space

and standing inference tie rule are identical within each comparison. Baselines that always use b messages are identified separately from policies that may use fewer.

Tables 51, 52, 53, and 54 give the same evaluation at budgets one through four.

Table 51: **Expanded Gemma results, budget 1.** Each cell is accuracy (%) / mean messages on the five cohorts. Learned selectors may use fewer than the budget. Reveal-all is an external reference; only the hindsight oracle uses outcomes to choose.

Selector	Omni-MATH-2	MaScQA	JEEBench	SciBench	LAB-Bench
Trained subset scorer	82.1 / 0.84	93.0 / 0.98	87.4 / 0.98	74.4 / 0.97	65.9 / 0.96
Untrained subset scorer	81.0 / 1.00	94.5 / 1.00	88.9 / 1.00	74.4 / 1.00	59.9 / 1.00
Correctness-only	81.0 / 1.00	93.5 / 1.00	89.9 / 1.00	74.4 / 1.00	64.3 / 1.00
Cost-aware	80.5 / 1.00	92.0 / 1.00	92.0 / 1.00	74.9 / 1.00	62.6 / 1.00
Message-blind	78.5 / 1.00	92.5 / 1.00	89.9 / 1.00	72.9 / 1.00	64.8 / 1.00
Pairwise-contrastive	82.0 / 1.00	93.0 / 0.94	90.0 / 0.97	73.9 / 0.95	64.3 / 0.99
Sequential, trained ranker	80.5 / 0.56	95.0 / 0.28	88.9 / 0.43	72.4 / 0.73	64.8 / 0.82
Sequential, untrained ranker	78.5 / 0.56	93.5 / 0.28	89.5 / 0.43	73.9 / 0.73	67.0 / 0.82
Receiver only	82.1 / 0.00	92.5 / 0.00	91.0 / 0.00	73.4 / 0.00	70.9 / 0.00
Reveal-all*	84.1 / 5.00	92.0 / 5.00	91.5 / 5.00	73.9 / 5.00	67.0 / 5.00
Random (exactly b)	81.0 / 1.00	93.3 / 1.00	90.2 / 1.00	72.6 / 1.00	62.7 / 1.00
Fixed order	79.5 / 1.00	93.0 / 1.00	87.9 / 1.00	71.9 / 1.00	61.5 / 1.00
At-most-1 hindsight oracle†	89.7 / —	99.0 / —	97.5 / —	82.4 / —	87.9 / —

Table 52: **Expanded Gemma results, budget 2.** Each cell is accuracy (%) / mean messages on the five cohorts. Learned selectors may use fewer than the budget. Reveal-all is an external reference; only the hindsight oracle uses outcomes to choose.

Selector	Omni-MATH-2	MaScQA	JEEBench	SciBench	LAB-Bench
Trained subset scorer	82.1 / 1.64	93.5 / 1.92	89.9 / 1.94	72.4 / 1.87	64.8 / 1.58
Untrained subset scorer	81.0 / 1.60	94.5 / 1.71	88.9 / 1.67	72.9 / 1.70	63.2 / 1.51
Correctness-only	81.0 / 1.98	92.0 / 1.96	91.0 / 1.99	74.4 / 1.99	63.7 / 1.96
Cost-aware	83.1 / 1.24	91.5 / 1.03	91.5 / 1.07	73.4 / 1.18	64.3 / 1.16
Message-blind	82.1 / 2.00	91.5 / 1.94	90.5 / 1.97	72.9 / 1.97	65.4 / 2.00
Pairwise-contrastive	81.5 / 2.00	93.0 / 2.00	91.5 / 2.00	73.4 / 2.00	63.2 / 2.00
Sequential, trained ranker	81.0 / 1.02	95.0 / 0.47	90.0 / 0.62	73.9 / 1.16	64.3 / 1.49
Sequential, untrained ranker	83.6 / 0.99	94.0 / 0.46	90.5 / 0.62	73.9 / 1.18	62.1 / 1.46
Receiver only	82.1 / 0.00	92.5 / 0.00	91.0 / 0.00	73.4 / 0.00	70.9 / 0.00
Reveal-all*	84.1 / 5.00	92.0 / 5.00	91.5 / 5.00	73.9 / 5.00	67.0 / 5.00
Random (exactly b)	82.3 / 2.00	92.4 / 2.00	90.6 / 2.00	72.9 / 2.00	64.8 / 2.00
Fixed order	81.0 / 2.00	92.0 / 2.00	90.5 / 2.00	71.4 / 2.00	67.0 / 2.00
At-most-2 hindsight oracle†	90.3 / —	99.0 / —	97.5 / —	83.9 / —	91.2 / —

F.4.2 Paired accuracy differences at budget five

Differences are question-paired bootstrap estimates (2,000 resamples, seed 20260923). For binary outcomes the exact two-sided sign test on discordant pairs is equivalent to exact McNemar. Holm correction covers the twenty displayed comparisons: four method pairs across five domains. The cost-aware and correctness-only selectors have unequal training exposure, so their contrast is between checkpoints rather than an isolated objective effect.

No comparison survives Holm correction. Biology's cost-aware contrast has the smallest raw p (0.0294) but adjusted $p = 0.589$. The SciBench pairwise-contrastive comparison has five wins and no losses, giving exact $p = 2/2^5 = 0.0625$ despite a percentile-bootstrap interval above zero. Sparse discordance can make that interval a poor substitute for the exact test. Materials QA's content-blind comparison has zero discordant outcomes; its zero-width empirical bootstrap interval reflects this sample and is not evidence of population equivalence.

Table 53: **Expanded Gemma results, budget 3.** Each cell is accuracy (%) / mean messages on the five cohorts. Learned selectors may use fewer than the budget. Reveal-all is an external reference; only the hindsight oracle uses outcomes to choose.

Selector	Omni-MATH-2	MaScQA	JEEBench	SciBench	LAB-Bench
Trained subset scorer	82.6 / 2.32	93.0 / 2.78	89.9 / 2.85	72.9 / 2.66	65.4 / 2.01
Untrained subset scorer	80.5 / 2.23	94.5 / 2.32	87.9 / 2.50	72.9 / 2.34	63.7 / 2.00
Correctness-only	83.1 / 3.00	91.0 / 2.94	91.0 / 2.91	74.9 / 2.96	67.6 / 2.85
Cost-aware	83.1 / 1.25	91.5 / 1.03	91.5 / 1.07	73.4 / 1.19	64.3 / 1.16
Message-blind	81.5 / 3.00	91.5 / 3.00	91.0 / 3.00	74.4 / 3.00	62.6 / 3.00
Pairwise-contrastive	80.5 / 2.90	92.5 / 2.67	90.0 / 2.54	73.9 / 2.87	63.7 / 2.26
Sequential, trained ranker	81.5 / 1.38	93.0 / 0.62	91.0 / 0.78	73.4 / 1.50	68.7 / 2.07
Sequential, untrained ranker	80.5 / 1.33	93.5 / 0.61	91.0 / 0.76	74.9 / 1.43	68.1 / 1.94
Receiver only	82.1 / 0.00	92.5 / 0.00	91.0 / 0.00	73.4 / 0.00	70.9 / 0.00
Reveal-all*	84.1 / 5.00	92.0 / 5.00	91.5 / 5.00	73.9 / 5.00	67.0 / 5.00
Random (exactly b)	82.6 / 3.00	92.0 / 3.00	90.8 / 3.00	73.9 / 3.00	66.3 / 3.00
Fixed order	83.1 / 3.00	91.5 / 3.00	89.5 / 3.00	75.9 / 3.00	63.7 / 3.00
At-most-3 hindsight oracle†	90.8 / —	99.0 / —	98.0 / —	84.4 / —	91.2 / —

Table 54: **Expanded Gemma results, budget 4.** Each cell is accuracy (%) / mean messages on the five cohorts. Learned selectors may use fewer than the budget. Reveal-all is an external reference; only the hindsight oracle uses outcomes to choose.

Selector	Omni-MATH-2	MaScQA	JEEBench	SciBench	LAB-Bench
Trained subset scorer	82.1 / 2.72	93.0 / 3.54	89.4 / 3.56	72.9 / 3.15	65.9 / 2.29
Untrained subset scorer	80.5 / 2.49	94.5 / 3.06	88.4 / 2.85	72.9 / 2.64	63.7 / 2.28
Correctness-only	84.1 / 3.75	91.5 / 3.63	91.0 / 3.55	74.4 / 3.63	70.3 / 3.51
Cost-aware	83.1 / 1.25	91.5 / 1.03	91.5 / 1.07	73.4 / 1.19	64.3 / 1.16
Message-blind	82.1 / 3.79	91.5 / 3.57	91.0 / 3.74	74.4 / 3.64	65.4 / 3.68
Pairwise-contrastive	82.0 / 3.21	92.0 / 3.50	91.5 / 3.32	74.9 / 3.21	65.9 / 3.31
Sequential, trained ranker	83.1 / 1.62	93.5 / 0.72	89.5 / 0.85	74.9 / 1.70	65.9 / 2.45
Sequential, untrained ranker	82.6 / 1.70	93.0 / 0.72	89.5 / 0.91	74.9 / 1.60	64.8 / 2.42
Receiver only	82.1 / 0.00	92.5 / 0.00	91.0 / 0.00	73.4 / 0.00	70.9 / 0.00
Reveal-all*	84.1 / 5.00	92.0 / 5.00	91.5 / 5.00	73.9 / 5.00	67.0 / 5.00
Random (exactly b)	83.7 / 4.00	92.1 / 4.00	91.3 / 4.00	75.0 / 4.00	66.1 / 4.00
Fixed order	84.1 / 4.00	93.0 / 4.00	91.0 / 4.00	75.4 / 4.00	63.2 / 4.00
At-most-4 hindsight oracle†	90.8 / —	99.0 / —	98.5 / —	84.4 / —	91.2 / —

G Measurement Protocol and Definitions

G.1 Protocol and Statistical Conventions

A recorded condition fixes the question, five previously generated messages, instructions to the receiver, evaluator, and order of message presentation. Changing the candidate-message generator would change the pool and is distinct from the receiver comparisons reported here. Each domain uses its own evaluation population; representative source testing and hard-stress questions are not pooled.

Both receiver conditions use a concise prompt asking the receiver to answer using the available information with a $32{,}768$-token receiver output limit and a $4{,}096$-token evaluator output limit, at temperature zero. These settings are limits, not actual token usage. Calibration questions are excluded from training and final testing. Completed responses without an extractable final answer count as zero; unresolved receiver or evaluator requests remain missing. Complete-subset analyses use only questions with all subsets recorded. Missingness can depend on generation length or answer format and need not be random (Appendix E.10).

Accuracy and message-count comparisons compare methods on the same questions. A question-paired bootstrap resamples whole questions, keeping each question's outcomes under both methods together. Structural rates use question rather than edge resampling, since a question's 32 states and

Table 55: **Budget-five paired selector comparisons.** Differences are row-method minus reference-method accuracy. Up/down/tied counts questions on which that difference is positive, negative or zero.

Domain	Diff.	95% CI	Up/down/tied	Raw p	Holm p	Verdict
Cost-aware minus correctness-only						
Omni-MATH-2	-0.0103	$[-0.0462, +0.0308]$	$6/8/181$	0.7905	1.0000	not rejected
MaScQA	-0.0050	$[-0.0300, +0.0200]$	$3/4/193$	1.0000	1.0000	not rejected
JEEBench	$+0.0151$	$[-0.0201, +0.0503]$	$8/5/186$	0.5811	1.0000	not rejected
SciBench	-0.0050	$[-0.0452, +0.0352]$	$8/9/182$	1.0000	1.0000	not rejected
LAB-Bench	$\mathbf{-0.0714}$	$[-0.1319, -0.0110]$	$9/22/151$	0.0294	0.5890	not rejected
Message-content-blind minus correctness-only						
Omni-MATH-2	-0.0051	$[-0.0256, +0.0103]$	$1/2/192$	1.0000	1.0000	not rejected
MaScQA	$\mathbf{+0.0000}$	$[+0.0000, +0.0000]$	$0/0/200$	1.0000	1.0000	no discordance
JEEBench	$+0.0151$	$[-0.0050, +0.0352]$	$4/1/194$	0.3750	1.0000	not rejected
SciBench	$+0.0000$	$[-0.0201, +0.0201]$	$2/2/195$	1.0000	1.0000	not rejected
LAB-Bench	-0.0440	$[-0.0989, +0.0110]$	$10/18/154$	0.1849	1.0000	not rejected
Pairwise-contrastive minus subset-scoring						
Omni-MATH-2	$+0.0000$	$[-0.0308, +0.0308]$	$5/5/185$	1.0000	1.0000	not rejected
MaScQA	-0.0100	$[-0.0350, +0.0150]$	$2/4/194$	0.6875	1.0000	not rejected
JEEBench	$+0.0201$	$[-0.0050, +0.0503]$	$6/2/191$	0.2891	1.0000	not rejected
SciBench	$+0.0251$	$[+0.0050, +0.0452]$	$5/0/194$	0.0625	1.0000	not rejected
LAB-Bench	$+0.0165$	$[-0.0440, +0.0714]$	$16/13/153$	0.7111	1.0000	not rejected
Sequential (trained ranker) minus subset-scoring						
Omni-MATH-2	$+0.0154$	$[-0.0256, +0.0564]$	$10/7/178$	0.6291	1.0000	not rejected
MaScQA	-0.0050	$[-0.0300, +0.0200]$	$3/4/193$	1.0000	1.0000	not rejected
JEEBench	$+0.0151$	$[-0.0151, +0.0503]$	$7/4/188$	0.5488	1.0000	not rejected
SciBench	$+0.0050$	$[-0.0352, +0.0452]$	$8/7/184$	1.0000	1.0000	not rejected
LAB-Bench	$+0.0330$	$[-0.0275, +0.0989]$	$20/14/148$	0.3915	1.0000	not rejected

80 add-one edges are correlated. Confidence intervals are pointwise unless a correction is stated. Binary paired tests use exact discordant-pair tests; budget-averaged non-binary contrasts may use sign-flip tests. Each family and denominator is stated with its results. A small or non-significant contrast does not establish equivalence, and pointwise bootstrap intervals do not replace corrected tests. Training-seed variation is not estimated by question bootstraps.

G.2 Measurement conventions and boundaries

Order and terminal outcomes. Although S is written as a set, the receiver consumes an ordered prompt. Messages in every subset follow the same fixed relative order. A separate order audit changes only that presentation order; identical-repeat checks estimate the variation that occurs without reordering. Selector-facing candidate order is a different variable and never silently changes the receiver evaluation conditions.

A successful request with no parseable final answer is a completed no-answer outcome with $U_g(S; q) = 0$, including both early stopping and output-ceiling exhaustion. It remains in accuracy and cost denominators. Operational failures, such as authentication or connection errors, have undefined utility until resolved and are not converted to zeros (Appendix E.2).

Scope of message effects. Effects are conditional on the fixed question, candidate texts, receiver, prompt and ordering. Averaged or context-specific effects describe recorded outcomes; neither alone establishes that a selector can predict them.