WHEN ACCEPTABLE ANSWERS SHARE PARTS, FEEDBACK HELPS MORE

Anonymous authors
Paper under double-blind review

ABSTRACT

A correct observation about one part of an answer can help with many acceptable answers, or with very few. We study whether this shared structure changes the value of component-level feedback. Our controlled task asks language models to recover five terms from a forty-term catalogue. We vary overlap among acceptable answers while holding their number, the numerical data, report format, and verifier allowance fixed. Across 24 families and 11 models, the graded-recovery advantage of the full over a profile-substituted report increases with overlap: the interaction is $+0.419$ $[+0.323, +0.508]$ per unit of a chance-corrected overlap index. Correct-minus-no-report recovery changes from -0.012 at zero overlap to $+0.366$ at high overlap. A calibrated mechanical reader, defined by an explicit expected-overlap objective, has a graded interaction of $+0.440$ and a larger exact-recovery interaction than the models. Fixed-report and proposal-conditioned studies on external SAT formulas show related positive associations; the fixed-report slope is $+0.169$ $[+0.086, +0.251]$ after answer-count control. The construction, controls, and reader comparison show how a task property changes the usefulness of a specified report. They also separate recovering useful parts from assembling an acceptable whole. Code and data: anonymous repository.

1 INTRODUCTION

Many tasks allow more than one good answer. Well-performing prediction models can rely on different variables (Fisher et al., 2019). A Boolean formula can have many solutions and may have literals true in every satisfying assignment (Kilby et al., 2005). In either case, feedback about a component can narrow the choice without specifying a complete answer. Its usefulness depends in part on what the acceptable answers share.

Consider two collections of five-term models. In one, four terms recur in every acceptable model; only the fifth changes. In the other, different models use different terms. A report about frequently used components gives a clear starting point in the first collection. The same kind of report leaves more of the choice unresolved in the second. Figure 1 illustrates this difference without changing the number or size of the answers. This motivates our question: **does overlap among acceptable answers change how much a language model gains from correct component feedback?**

Feedback-driven systems use different information. ReAct interleaves reasoning with external observations (Yao et al., 2023); Self-Refine iterates on self-feedback (Madaan et al., 2023); CRITIC and Self-Debug use tools or execution results (Gou et al., 2024; Chen et al., 2024). Intrinsic correction differs from correction with external or oracle feedback (Huang et al., 2024). With-versus-without comparisons show what a system gains, but leave our task-side overlap question open. In equation search, LLM-SR proposes structures and evaluates them numerically (Shojaee et al., 2025a), while influence-guided feedback can guide term selection (Saveliev et al., 2026). In our setting, a useful component diagnosis still has to become compatible parts and a complete answer.

We build a controlled term-selection problem in which the numerical data and catalogue remain the same while the declared acceptable set changes. At each of three primary levels there are exactly 32 acceptable five-term supports from the same 40-term catalogue. The levels range from balanced component frequencies to a shared four-term core. A correct report contains a fit score and five component-profile/frequency slots. A matched control retains two original slots and replaces three profiles. Other arms supply a scalar score, an all-swapped report, or no fixed report.

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.

Different parts
Four acceptable answers

a	b	c	d	e
f	g	h	i	j
k	l	m	n	o
p	q	r	s	t

No component appears in every answer.

Shared parts
Four acceptable answers

a	**b**	**c**	**d**	e
a	**b**	**c**	**d**	f
a	**b**	**c**	**d**	g
a	**b**	**c**	**d**	h

Four components recur in every answer.

Figure 1: A schematic answer-set example, not an experimental result. Both collections contain four answers with five parts each. On the right, evidence about a recurring part applies to every acceptable answer. The experiment uses 32 accepted supports at each primary level and reports profiles that can correspond to several terms, rather than the unambiguous letters shown here.

The result is a strong change in feedback value. The full-minus-partial-sham graded-recovery interaction is $+0.419$ $[+0.323, +0.508]$, with positive point estimates on all 11 models and 23 of 24 families. The absolute scores matter too: correct feedback is close to no report at zero overlap, but exceeds it by 36.6 percentage points of graded recovery at high overlap. No-report recovery itself decreases as the accepted set changes.

The report construction deliberately makes recurring components easier to identify. What remains to be measured is how the models use that information, whether improvement reaches exact recovery, and what an explicit consumer can obtain from the report. We therefore compare the models with a calibrated expected-overlap reader. It optimizes a stated report model over fitting catalogue supports. Its graded interaction is similar in magnitude to the models', while its exact-recovery interaction is larger. Shared parts make this report useful to more than one kind of consumer, but recovering parts and completing an accepted answer are different outcomes.

The paper makes three contributions:

1. A controlled construction that varies overlap among acceptable answers while fixing answer count, support size, numerical data, report schema, and verifier allowance.
2. A measured relationship between that structure and language-model feedback value, interpreted through absolute arm scores, two recovery outcomes, and a named mechanical reader.
3. External fixed-report and own-proposal SAT checks that extend the question while distinguishing a controlled intervention from an observed association.

The resulting claim concerns these component reports and decision rules. It does not require a universal law of feedback or an automatic way to estimate answer-set overlap in open-ended scientific work.

2 Related work

Correction methods and the information they use. Reflexion and TextGrad use verbal feedback in broader optimization loops (Shinn et al., 2023; Yuksekgonul et al., 2024). A separate corrector can be trained to revise a generator using scalar or language feedback (Welleck et al., 2023). Self-Debug's variants differ in supplied and generated content, not merely wording (Chen et al., 2024). Our intervention instead changes the accepted-answer set while holding the consumer and report construction fixed.

Comparisons that make a correction result interpretable. Code self-repair and broader self-correction work emphasize baseline computation and feedback provenance (Olausson et al., 2024; Kamoi et al., 2024). Song et al. (2025) organize self-improvement around the relationship between generation and verification. Process and outcome supervision label intermediate steps and final answers, respectively (Lightman et al., 2024); here graded part recovery and exact acceptance remain separate outcomes. Full, profile-substituted, scalar, and no-report arms show what each comparison adds. Content-altering controls have precedents, including randomized demonstration labels (Min

et al., 2022); our contribution is their use in this controlled answer-set comparison, not the general idea of a content control.

Scientific components, stability, and explicit readers. Equation discovery includes physics-inspired symbolic regression (Udrescu & Tegmark, 2020) and search with learned reusable concepts (Grayeli et al., 2024). EoH and FunSearch instead evaluate candidate programs (Liu et al., 2024; Romera-Paredes et al., 2024). In classification, similarly performing models can make conflicting predictions (Marx et al., 2020). Our task fixes data and varies declared acceptable supports, rather than discovering a new law. LLM-SRBench evaluates recovered equations (Shojaee et al., 2025b), and ScienceAgentBench evaluates completed scientific programs (Chen et al., 2025); these final-artifact checks differ from graded part recovery. The overlap index uses an established chance-corrected stability form (Kuncheva, 2007; Nogueira et al., 2018), applied to acceptable supports rather than supports recovered under data resampling. The distinction between an information source and a decision rule is also classical (Blackwell, 1953; DeGroot, 1970). We instantiate it with a reader whose inputs, working assumptions, objective, and computation are explicit. This lets the experiment address model behavior without attributing every gain from an informative report to a model-specific reasoning ability.

3 A CONTROLLED COMPARISON OF SHARED ANSWER PARTS

3.1 ONE NUMERICAL PROBLEM, DIFFERENT ACCEPTABLE SETS

Each family fixes a catalogue of $M = 40$ complexity-matched terms and a numerical dataset. An answer uses $k = 5$ terms, with coefficients fitted by least squares. The terms are log-monomials over four active variables. Their exponent vectors span a four-dimensional space, so many five-term supports fit the same target exactly. This makes it possible to hold the numerical fitting problem fixed while changing which fitting supports receive acceptance credit.

Let $\mathcal{S}_f$ be the declared acceptable set for family f at a given level. Every member fits the data; fitting the data alone does not imply membership. The three primary constructions are:

- **Balanced:** each catalogue component appears in four of the 32 supports, with no universal component.
- **Two-term core:** two components recur in every support and the remaining three vary.
- **Four-term core:** four components recur in every support and the fifth varies.

All three contain 32 five-term supports. These are conditions within each catalogue/data family, not three unrelated task collections. Appendix A gives the generator and the additional descriptive levels.

We summarize shared structure with the chance-corrected overlap index L, also called localizability in the archived experiments. If q_j is component j's inclusion frequency across $\mathcal{S}_f$, define $p_j = q_j/k$, so $\sum_j p_j = 1$, and

$$L = \frac{\sum_{j=1}^{M} p_j^2 - 1/M}{1/k - 1/M}. \tag{1}$$

Balanced frequencies give $L = 0$; a single common support gives $L = 1$. The primary levels are approximately 0.00, 0.38, and 0.78. For two independently drawn acceptable supports, the construction identity in Appendix A gives expected shared-component counts of 0.625, about 2.3, and about 4.0, respectively. Analyses use the frozen, family-specific values rather than these rounded labels.

The data, catalogue, numerical fitting routine, support size, answer count, model settings, and evaluation allowance stay fixed across these levels. The accepted set and resulting score condition change. Thus the construction does not hold every meaning of task difficulty fixed: even an uninformed answer can have a different recovery score against a different accepted set. We show that change directly rather than treating it as an unknown defect.

Table 1: Fixed-report arms in the controlled construction. All arms use the same numerical observations and search procedure. Format and length matching apply to the full and sham reports, not to report versus no report.

Arm	Supplied fixed report
`full`	True fit score; five truthful profile/frequency slots.
`sw40`	Same score and frequency fields; two original slots retained, three profiles replaced.
`sw0`	Same score and frequency fields; all five profiles replaced.
`scalar`	True fit score, without component slots.
`none`	No fixed report; observations and proposal-fit archive remain.

3.2 The report and its controls

A `full` report gives the true fit score and five slots describing the most frequent components. Each slot contains a *profile*—variables, operators, and expression depth—and that component's inclusion frequency in the acceptable set. Several distinct terms can share a profile. The report therefore provides component evidence, not a complete accepted answer or the joint list of acceptable supports.

The report is fixed before the model starts. During search, all arms can receive numerical fit scores for their own proposals; the fixed component report is not updated from those proposals. Table 1 defines the arms. Here `sw40` retains two original slots and substitutes three profiles; `sw0` substitutes all five. The external studies define their controls separately.

The primary `full`$-$`sw40` contrast changes the reported profiles, not their frequency fields. Substitution need not create false values: at $L = 0$, every component occurs in $4/32$ accepted supports, so both substituted reports remain frequency-correct. At other levels, substitutions can also introduce misinformation. No report and `sw0` locate the contrast relative to omission and more substitutions. The matched format does not equalize decision value across overlap levels.

3.3 Model protocol, outcomes, and independent units

Models receive the term grammar, ten displayed numerical observations, the arm-specific report, and the archive of their proposals and fit scores. They make two proposal rounds of eight candidates and then one final commitment, under a 48-evaluation allowance. The configured proposal stage uses 16 charged slots, including unparseable proposals. Final recovery is scored after commitment. Model decoding limits and this schedule remain unchanged across arms. Appendix B gives the prompt contract, parsing rules, and resource accounting.

The primary graded outcome is the largest term-set F_1 with an acceptable support:

$$G(C, \mathcal{S}_f) = \max_{S \in \mathcal{S}_f} \frac{2|C \cap S|}{|C| + k}, \qquad E(C, \mathcal{S}_f) = \mathbf{1}\{C \in \mathcal{S}_f\}. \tag{2}$$

E is the exact outcome. Terms are compared by their frozen functional fingerprints; for two distinct five-term supports, G is simply the shared fraction. An unparsed final answer scores zero. We evaluate the committed answer, not the best candidate the model proposed. In Figure 1 (right), $\{a, b, c, d, x\}$ with $x \notin \{a, \ldots, h\}$ has $G = 0.8$ and $E = 0$, while $\{a, b, c, d, f\}$ has $G = E = 1$.

There are 24 families, two seeds, and 11 measurable model endpoints. Each arm–level mean contains 528 cells; the primary two-arm, three-level comparison contains 3,168. We average the paired full-minus-partial-sham differences across models and seeds inside each family and fit their slope on that family's L values. The headline interaction is the mean of these 24 family slopes. Intervals and sign-flip tests use families as the independent units. They do not count tokens, seeds, or 3,168 responses as independent tasks. The coefficient is a change in the feedback gap per unit of L, not an unconditional accuracy gain.

Table 2: Graded recovery at the three primary levels; 528 cells per entry. All levels have 32 accepted supports. The score is on a 0–1 scale. The last column is a descriptive difference of arm means, not an interaction coefficient or a separate significance claim.

L	None	Scalar	All swapped	Partial sham	Full	Full − none
0.00	0.391	0.395	0.411	0.435	0.379	−0.012
0.38	0.307	0.309	0.327	0.403	0.569	+0.262
0.78	0.265	0.262	0.195	0.360	0.631	+0.366

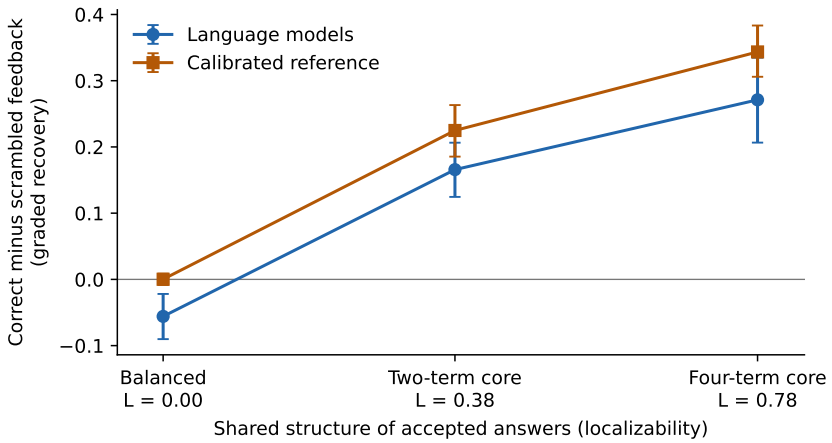

Figure 2: The full-minus-partial-sham graded-recovery gap grows with shared answer parts. Points summarize the complete 24-family, two-seed, 11-model grid at the three fixed-count levels. "Calibrated reference" is the expected-overlap reader defined in Section 5. Bars are pointwise 95% intervals from 20,000 family bootstrap resamples; lines are visual guides. They are not the interval for the interaction slope.

4 Correct component feedback helps more at high overlap

The primary graded interaction is $+0.419$ $[+0.323, +0.508]$, with family sign-flip $p = 5 \times 10^{-6}$. The point estimate is positive for every model and for 23 of 24 families. Figure 2 shows the change in the full-minus-partial-sham gap. Table 2 shows the absolute scores behind it.

The increase is visible against no report. At $L = 0$, full recovery is 0.379, versus 0.391 without a fixed report. At $L = 0.78$, full recovery is 0.631, versus 0.265 without one. The resulting full-minus-none means are -0.012, $+0.262$, and $+0.366$. Thus the main pattern is not visible only through subtraction of the substituted arm. At the same time, the declining no-report score shows that changing the acceptable set changes the recovery problem. The full report improves while the no-report baseline falls; both contribute to the widening gap.

Profile replacement matters. At high overlap the partial-sham report scores 0.360, above no report, whereas the all-swapped report scores 0.195, below it. Both shams keep the full report's percentages: at $L = 0.78$, the partial sham retains one or two original 100% slots, whereas the all-swapped report assigns the four 100% slots to decoy profiles. The scalar and no-report means are close at all three levels. These controls locate the useful signal in the component report rather than the mere presence of a fit-score field. At zero overlap all three profile reports are frequency-correct; the partial sham nevertheless has the highest observed mean.

What the construction lets us learn. Concentrated marginals make the deliberately constructed report more informative about recurring components. This explains why a positive trend is plausible, but it does not determine an LLM's recovery score. A consumer can misread a profile, choose a different term with the same profile, or assemble useful parts into a nonmember. The measured interaction and its consistency across models quantify that response. The exact outcome and reader

comparison test how far the response reaches. Appendix C reports the remaining level means, per-model estimates, and seed sensitivities.

5 WHAT AN EXPLICIT READER TELLS US

The reader's decision comes first. The calibrated expected-overlap reader receives the report, catalogue, and data-consistency constraint. It treats reported frequencies at face value, including slots whose profiles were substituted. Within a profile, it assigns component identities exchangeably; it spreads undisclosed inclusion mass over the remaining identities. Let π_j be the resulting inclusion probability. The reader chooses a data-consistent support maximizing

$$U(C) = \frac{1}{k} \sum_{j \in C} \pi_j. \tag{3}$$

This is expected overlap with an answer under the stated report model. It is not the benchmark's maximum-over-accepted-set score in Equation 2. The actual accepted set is used to evaluate the selected commitment, not to choose it.

The small catalogue allows exhaustive optimization over $\binom{40}{5} = 658{,}008$ supports, with data-inconsistent supports filtered out. An arm-independent failure mixture is calibrated using only scalar and no-report graded means, then held fixed for the diagnostic arms. The reader uses 48 charged verifier calls per run, but its offline enumeration and access to the explicit catalogue and structured report fields are additional resources. This is a transparent decision reference, not a wall-time-matched implementation comparison. Appendix D specifies the probability model, calibration, and tie handling.

Table 3: Feedback-by-overlap interactions for models and the calibrated expected-overlap reader. Both are evaluated on the committed-answer outcome in the first column. The last column is a paired difference of slopes with a 95% family-clustered interval.

Outcome	Models	Reader	Model − reader
Graded recovery	$+0.419$	$+0.440$	-0.021 $[-0.102, +0.058]$
Exact recovery	$+0.151$	$+0.486$	-0.334 $[-0.461, -0.222]$

The shared trend is not specific to a language-model consumer. The reader's graded interaction is $+0.440$, beside the models' $+0.419$. Their difference is -0.021 $[-0.102, +0.058]$ ($p = 0.62$). This comparison shows that an explicit rule can exploit the same task/report relationship. It does not establish equivalence: the interval allows small differences in either direction.

From reported parts to accepted answers. Table 4 separates the two kinds of absolute outcome. Model exact entries count accepted commitments out of 528 in each arm–level cell. Reader exact entries are calibrated expected rates, not observed model counts; they average the saved 24 family means, each based on 32 seeded reader runs.

Without a report, all fitting supports tie under the reader's objective. Its calibrated policy draws a fitting support uniformly with a fixed probability and returns an empty answer otherwise. The expected graded mean falls from 0.384 to 0.245 as overlap rises, compared with 0.391 to 0.265 for the models. The falling no-report score is therefore not unique to model behavior: the accepted set changes the baseline against which both consumers are scored.

At $L = 0$, all five report slots read 12.5%, each describing a component found in four of the 32 accepted supports. Full-report graded means remain close to each consumer's no-report baseline, and both full-report exact rates are zero. At $L = 0.38$, two slots read 100% and graded recovery rises. An accepted answer also needs three varying terms that occur together; marginal frequencies do not give that joint list. The models produce two accepted full-report commitments out of 528, while the reader's expected exact rate is 0.014.

At $L = 0.78$, four slots read 100%. A five-term commitment containing those four true components has $G \geq 0.8$, and any of 32 of the other 36 catalogue terms would complete an accepted support. Profiles can correspond to several terms, however; the report does not name term identities. The

Table 4: Absolute outcomes behind Table 3. The display column lists the full report's frequency fields in all 24 families: $2 \times 100\%$, for example, means two slots at 100%. Both shams keep these percentages and layout. The reader uses the same report fields, with its fixed calibration mixture included in the expectations. These are arm means and full-report exact outcomes, not slopes.

Models: observed commitments

	Graded mean			Exact count
Level and full-report display	None	Partial sham	Full	Full
$L = 0.00$: $5 \times 12.5\%$	0.391	0.435	0.379	$0/528$
$L = 0.38$: $2 \times 100\%$ 3×12.5–25.0%	0.307	0.403	0.569	$2/528$
$L = 0.78$: $4 \times 100\%$ $1 \times 3.1\%$	0.265	0.360	0.631	$62/528$

Reader: calibrated expectations

	Graded mean			Exact rate
Level	None	Partial sham	Full	Full
$L = 0.00$	0.384	0.378	0.378	0.000
$L = 0.38$	0.333	0.392	0.616	0.014
$L = 0.78$	0.245	0.378	0.722	0.375

models receive the grammar rather than a printed catalogue and reach a full-report graded mean of 0.631, with 62 of 528 commitments accepted. The catalogue-based reader reaches an expected graded mean of 0.722 and an expected exact rate of 0.375 (0.451 before the fixed calibration mixture). No model commitment in the other four arms is accepted at any of the three primary levels ($0/6{,}336$ pooled across four arms and three levels). These aggregates describe outcomes, not the causes of individual failures.

For exact recovery, the reader's interaction is $+0.486$, compared with $+0.151$ for the models. The difference is -0.334 $[-0.461, -0.222]$ ($p = 5 \times 10^{-6}$). The model improvement in parts therefore translates less strongly into accepted whole answers as overlap increases.

An explicit objective makes the comparison useful. Optimizing expected overlap removes search error for that working objective. It does not optimize exact recovery, the benchmark's maximum-over-set score, or the difference between arms. Those are evaluated outcomes, so the model–reader interactions remain empirical questions. The comparison also depends on the reader: across seven specified policies, the graded model-minus-reader point estimate ranges from $+0.255$ to -0.021. The within-model $+0.419$ result is unchanged by that choice. Appendix E gives the bracket, including the informationally privileged true-set policies, and the precision of the calibrated-reader comparison.

6 External SAT checks

The controlled construction makes the accepted set explicit. We next ask whether a related association appears in externally authored SAT formulas from SATLIB (Hoos & Stützle, 2000). An acceptable answer is now a satisfying assignment. Shared parts are variable values that recur across solutions. A literal true in every solution is a backbone literal (Kilby et al., 2005); our index also counts partial recurrence. For n Boolean variables with true-value marginals q_i, the same chance-corrected component index becomes $L = n^{-1} \sum_i (2q_i - 1)^2$. Here overlap is observed across formulas, not assigned by an experimenter. The two studies below also differ in when their reports are formed.

6.1 Fixed reports on 1,000 formulas

The `uf20-91` corpus contains 1,000 satisfiable formulas with 20 variables and 91 clauses. We enumerate all satisfying assignments to compute exact marginals and L. A frozen selection of

Table 5: External associations. Slopes are per unit of observed L and have study-specific outcomes and controls; they are not interchangeable effect sizes. All intervals shown here use formulas as the independent units. The common-corpus fits are exploratory and use the same 834 rows.

Setting and outcome	Population	Contrast and slope [95% interval]
Fixed report; graded	1,000 uf20 formulas; one model retained	Full − sham: $+0.213$ $[+0.145, +0.280]$; answer-count controlled: $+0.169$ $[+0.086, +0.251]$
Fixed report; graded	Same population and count control	Model − reader interaction: $+0.046$ $[-0.032, +0.124]$
Own-proposal report; exact	90 formulas, ten models, 834 usable first answers	Full − sham: $+0.385$ $[+0.200, +0.574]$
Own-proposal report; exact	Same common-corpus population	Full − none: $+0.214$ $[+0.040, +0.414]$

ten variables receives a marginal report. Full reports all ten truthfully; sham inverts four values, $q \mapsto 1 - q$, preserving confidence magnitudes and layout. The outcome is one minus normalized Hamming distance to the nearest satisfying assignment. Exact satisfaction is secondary.

A prospective ceiling rule excludes models whose full-arm graded mean exceeds 0.95 from the confirmatory population. GPT-5.5 and GPT-5.6 meet this rule; Opus-4.8 does not. Its full-minus-sham slope is $+0.213$ $[+0.145, +0.280]$ (sign-flip $p = 0.0002$). Because formulas with fewer solutions also tend to have higher overlap, we report answer-count control: the slope remains $+0.169$ $[+0.086, +0.251]$. The corresponding count-controlled *model-minus-reader* slope is a different contrast, $+0.046$ $[-0.032, +0.124]$. The exact-recovery slope is $+0.052$ $[-0.119, +0.223]$ ($p = 0.62$).

This external result supports the graded association for the retained model, not confirmation across all three models or an exact-recovery trend. Magnitude and layout matching do not make truthful and inverted marginals equally informative, and malformed answers remain part of the as-run outcome. Appendix F gives the population rule, all-model means, final token settings, and parser sensitivity.

6.2 Feedback about a model's own proposal

The common-corpus study gives ten models the same 90 formulas: 45 from `uf20-91` and 45 from the harder `uf50-218`. Each model makes one first attempt, shared by all four revision arms. The full report flags five variables most contradicted by the saved solution marginals among a selected pool of twelve. The partial sham keeps three slots and substitutes two decoys at matched confidence magnitudes. A schema control names the checked variables without values, and a no-report control asks for another attempt. These revisions use 16,000-token call limits, rather than the fixed-report study's 12,000; the two studies are not pooled. On the larger formulas, marginals, L, and the recorded solution count come from a bounded solution-finding procedure, not exhaustive enumeration.

Of 900 model–formula rows, 834 have usable shared first answers. The current analysis retains all of them, with no full-arm headroom screen. Controlling for corpus tier and recorded log answer count, the *exact-solving* full-minus-sham slope is $+0.385$ $[+0.200, +0.574]$. The full-minus-none component is $+0.214$ $[+0.040, +0.414]$, and none-minus-sham is $+0.170$ $[-0.012, +0.327]$. These are dependent exploratory components from one archive, not three replications. The raw final successes are $305/834$ with full feedback, 248 with none, 254 with schema, and 200 with sham. Those counts summarize mean success, not the L interaction.

The association thus extends to feedback about an actual first attempt, including comparison with another attempt without a report. Consumer choice still matters. In a separate five-model revision study, the model's mean feedback advantage over a simple marginal reader is $+0.056$ $[+0.011, +0.101]$, but the mean comparison reverses after one greedy variable flip is added to the reader. An unhinted complete solver solves the small formulas within 64 branching decisions, a different resource unit. Appendix H presents this reader ladder without treating it as solver superiority. Appendix G reports the common-corpus populations, formula-clustered sensitivities, and repair/damage counts.

7 DISCUSSION AND CONCLUSION

The task and the report belong in the same explanation. A report about recurring parts is particularly useful when acceptable answers share those parts. The controlled study turns that intuition into a measured comparison while preserving answer count and the numerical fitting problem. Its contribution is not a new definition of information or a claim that a positive slope follows for every consumer. It is evidence that the tested models respond strongly to this constructed relationship, with different consequences for graded and exact recovery.

This distinction is relevant to scientific agents. When several models explain a dataset, a component diagnosis can help select a useful subset without resolving the entire explanation. Evaluating only final fit can miss that selection problem; evaluating only component overlap can miss whether the selected parts form an acceptable whole. Reporting both, alongside a reader with a stated objective, makes the result easier to interpret. It is an evaluation practice, not evidence that an agent can cheaply enumerate all acceptable scientific explanations before deciding whether to buy feedback.

Scope follows the tested decision. The numerical construction is intentionally underdetermined, and its acceptance sets are authored. The external formulas supply a different origin for acceptable answers, but their overlap is observational, the fixed-report population retains one of three models, and the common-corpus reanalysis is exploratory. Neither design separates overlap from every possible property of a decision problem. A completed pair-credit probe gives a near-zero interaction, while an incomplete Boolean-circuit study gives positive directional evidence (Appendix I). Their restricted overlap range and partial coverage, respectively, make them scope checks rather than additional main confirmations.

The main finding is direct: within the controlled five-term task, correct component feedback helps more as acceptable answers share more parts. The model and explicit reader both show this pattern, while the reader converts it more strongly into exact accepted answers. External fixed-report and own-proposal studies support related associations. Together, the results identify a useful task/report relationship and a concrete way to study it: vary the acceptable set, show the absolute controls, and specify who reads the evidence and what that reader tries to achieve.

AI USE STATEMENT

Language models were experimental components. We also used generative AI to generate synthetic datasets; clean and reformat data; critique experiments; implement methods; analyze cases qualitatively; and simulate reviews. AI assisted literature search, software and artifact preparation, and editing and proofreading this paper. We did not use AI to develop conceptual frameworks, mathematical claims, proof ideas and proof text, propose and refine hypotheses, interpret results, or translate. Surveys, interviews, and transcription were not applicable. The authors reviewed all AI-assisted work, including code tests, saved-result replays and citation-source checks, and take responsibility for the final text, claims and artifacts.

REPRODUCIBILITY STATEMENT

Appendices A–H specify the tasks, models, controls and analyses. Appendix J maps the reported results to their code, data and offline replay procedures. Code and data are linked in the anonymous repository.

REFERENCES

David Blackwell. Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, 24(2):265–272, 1953. doi: 10.1214/aoms/1177729032.

Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. Teaching large language models to self-debug. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2304.05128.

Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, Vishal Dey, Mingyi Xue, Frazier N. Baker, Benjamin Burns, Daniel Adu-Ampratwum, Xuhui Huang, Xia Ning, Song Gao, Yu Su, and Huan Sun. ScienceAgentBench: Toward rigorous assessment of language agents for data-driven scientific discovery. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/f12b4df26344f3be803c06b555252efe-Abstract-Conference.html.

Morris H. DeGroot. *Optimal Statistical Decisions*. McGraw–Hill, 1970.

Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177):1–81, 2019. URL http://jmlr.org/papers/v20/18-760.html.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. CRITIC: Large language models can self-correct with tool-interactive critiquing. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2305.11738.

Arya Grayeli, Atharva Sehgal, Omar Costilla-Reyes, Miles Cranmer, and Swarat Chaudhuri. Symbolic regression with a learned concept library. In *Advances in Neural Information Processing Systems*, volume 37, pp. 44678–44709, 2024. doi: 10.52202/079017-1419. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/4ec3ddc465c6d650c9c419fb91f1c00a-Abstract-Conference.html.

Holger H. Hoos and Thomas Stützle. SATLIB: An online resource for research on SAT. In Ian P. Gent, Hans van Maaren, and Toby Walsh (eds.), *SAT 2000*, pp. 283–292. IOS Press, 2000. URL https://www.cs.ubc.ca/~hoos/SATLIB/.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *International Conference on Learning Representations (ICLR)*, 2024.

Ryo Kamoi, Yusen Zhang, Nan Zhang, Jiawei Han, and Rui Zhang. When can LLMs actually correct their own mistakes? A critical survey of self-correction of LLMs. *Transactions of the Association for Computational Linguistics (TACL)*, 12:1417–1440, 2024. doi: 10.1162/tacl_a_00713. URL https://aclanthology.org/2024.tacl-1.78/.

Philip Kilby, John Slaney, Sylvie Thiébaux, and Toby Walsh. Backbones and backdoors in satisfiability. In *AAAI-05*, pp. 1368–1373, 2005. URL https://cdn.aaai.org/AAAI/2005/AAAI05-217.pdf.

Ludmila I. Kuncheva. A stability index for feature selection. In *Proceedings of the 25th IASTED International Multi-Conference on Artificial Intelligence and Applications*, pp. 390–395, 2007.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/aca97732e30bcf1303bc22ac3924fd16-Abstract-Conference.html.

Fei Liu, Xialiang Tong, Mingxuan Yuan, Xi Lin, Fu Luo, Zhenkun Wang, Zhichao Lu, and Qingfu Zhang. Evolution of heuristics: Towards efficient automatic algorithm design using large language model. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32201–32223. PMLR, 2024. URL https://proceedings.mlr.press/v235/liu24bs.html.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-Refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46534–46594, 2023. doi: 10.52202/075280-2019. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91edff07232fb1b55a505a9e9f6c0ff3-Abstract-Conference.html.

Charles Marx, Flavio Calmon, and Berk Ustun. Predictive multiplicity in classification. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6765–6774. PMLR, 2020. URL https://proceedings.mlr.press/v119/marx20a.html.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022. URL https://arxiv.org/abs/2202.12837.

Tom M. Mitchell. Generalization as search. *Artificial Intelligence*, 18(2):203–226, 1982. doi: 10.1016/0004-3702(82)90040-6.

Sarah Nogueira, Konstantinos Sechidis, and Gavin Brown. On the stability of feature selection algorithms. *Journal of Machine Learning Research*, 18(174):1–54, 2018. URL https://jmlr.org/papers/v18/17-514.html.

Theo X. Olausson, Jeevana Priya Inala, Chenglong Wang, Jianfeng Gao, and Armando Solar-Lezama. Is self-repair a silver bullet for code generation? In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2306.09896.

Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M. Pawan Kumar, Emilien Dupont, Francisco J. R. Ruiz, Jordan S. Ellenberg, Pengming Wang, Omar Fawzi, Pushmeet Kohli, and Alhussein Fawzi. Mathematical discoveries from program search with large language models. *Nature*, 625(7995):468–475, 2024. doi: 10.1038/s41586-023-06924-6. URL https://doi.org/10.1038/s41586-023-06924-6.

Evgeny S. Saveliev, Samuel Holt, Nabeel Seedat, David L. Bentley, Jim Weatherall, and Mihaela van der Schaar. Influence-guided symbolic regression: Scientific discovery via LLM-driven equation search with granular feedback. *arXiv preprint arXiv:2605.29184*, 2026. URL https://arxiv.org/abs/2605.29184.

H. S. Seung, M. Opper, and H. Sompolinsky. Query by committee. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory (COLT)*, pp. 287–294, 1992. doi: 10.1145/130385.130417.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 8634–8652, 2023.

Parshin Shojaee, Kazem Meidani, Shashank Gupta, Amir Barati Farimani, and Chandan K Reddy. LLM-SR: Scientific equation discovery via programming with large language models. In *International Conference on Learning Representations*, 2025a. URL https://arxiv.org/abs/2404.18400.

Parshin Shojaee, Ngoc-Hieu Nguyen, Kazem Meidani, Amir Barati Farimani, Khoa D Doan, and Chandan K. Reddy. LLM-SRBench: A new benchmark for scientific equation discovery with large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 55325–55359. PMLR, 2025b. URL https://proceedings.mlr.press/v267/shojaee25a.html.

Yuda Song, Hanlin Zhang, Carson Eisenach, Sham Kakade, Dean Foster, and Udaya Ghai. Mind the gap: Examining the self-improvement capabilities of large language models. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2412.02674.

Silviu-Marian Udrescu and Max Tegmark. AI Feynman: A physics-inspired method for symbolic regression. *Science Advances*, 6(16):eaay2631, 2020. doi: 10.1126/sciadv.aay2631.

Sean Welleck, Ximing Lu, Peter West, Faeze Brahman, Tianxiao Shen, Daniel Khashabi, and Yejin Choi. Generating sequences by learning to self-correct. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=hH36JeQZDaO.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Mert Yuksekgonul, Federico Bianchi, Joseph Boen, Sheng Liu, Zhi Huang, Carlos Guestrin, and James Zou. TextGrad: Automatic "differentiation" via text. *arXiv preprint arXiv:2406.07496*, 2024.

A CONSTRUCTION OF THE CONTROLLED ANSWER SETS

This appendix gives the details needed to understand and reproduce the comparisons in the main paper. Appendices A and B describe the controlled task and model protocol; Appendix C gives additional results; Appendices D and E define and compare the readers. The external fixed-report, common-corpus, and separate five-model revision studies are in Appendices F, G, and H. The final two appendices cover the compact scope checks and reproduction coverage.

A.1 OBJECTS IN THE EXPERIMENT

A family is one generated catalogue and numerical dataset. A level selects a declared acceptable set within that family. An arm changes the report supplied to the consumer. A model–family–seed–level–arm cell produces one final commitment. These objects should not be interchanged: the primary estimate generalizes over the 24 families of this generator, not over $3{,}168$ unrelated scientific problems.

Table 6: Objects and quantities in the controlled study. The inclusion frequencies q_j sum to the support size; the normalized frequencies p_j sum to one.

Symbol	Meaning
$M = 40$	Number of fingerprint-distinct terms in a family catalogue.
$k = 5$	Required number of terms in an accepted support.
$\mathcal{S}_f$	Declared collection of acceptable supports for a family and level.
q_j	Fraction of supports in $\mathcal{S}_f$ that include component j.
$p_j = q_j/k$	Normalized component frequency used in Equation 1.
L	Chance-corrected concentration of these normalized frequencies.
C	Set of functional fingerprints in a parsed model commitment.
G, E	Graded and exact committed-answer recovery, Equation 2.
π_j	Component-inclusion probability under the reader's working report model.
$U(C)$	The reader's expected-overlap selection objective, not the benchmark maximum.

A.2 A CATALOGUE WITH MANY NUMERICALLY FITTING SUPPORTS

Each family selects four active variables from six positive-valued inputs. The catalogue contains 40 distinct log-monomial terms, represented by operator trees of depth at most four. The grammar uses powers, products, and quotients inside a logarithm. For example, $\log(x_1^2 x_2^{-1}) = 2\log x_1 - \log x_2$. A term therefore has an exponent vector over the four active variables. The catalogue vectors span a four-dimensional space.

The target coefficients over the active coordinates are $(2, -3, 5, 1)$. For input x, the target is the corresponding linear combination of the active log-inputs. Inputs are drawn uniformly from $[0.6, 2.4]$. The family has 160 development observations, 160 holdout observations, and 64 probe points. The base family seed is 20260814, with family identifiers entering the deterministic randomization. Every overlap level of a family reuses these arrays and the same catalogue.

Any five-term catalogue support whose exponent vectors have rank four spans the target. Least-squares fitting, with an intercept, can therefore recover the same target exactly. The generator checks rank and the resulting numerical fit. The probe points also distinguish term fingerprints, preventing two different catalogue identities from merely being duplicate numerical columns under the comparison rule.

The accepted set is a declared subset of these fitting supports. This separation is deliberate. Numerical fit supplies a shared feasibility problem, while the report supplies evidence about which fitting supports the acceptance rule rewards. A nonmember can fit perfectly and still receive no exact credit. The experiment consequently studies selection among acceptable structures under an authored acceptance rule, rather than identification of a uniquely data-determined physical law.

Terms share a bounded grammar and the frozen complexity controls. The construction does not assume that all expressions with similar numerical behavior have the same syntactic form. This distinction is important for the profile report and for scoring freely written model terms by their functional fingerprints.

A.3 THE THREE FIXED-COUNT LEVELS

Balanced. The generator creates 32 five-term supports in which each of the 40 components appears exactly four times. It checks that each support has distinct components and rank four. Consequently $q_j = 4/32 = 1/8$ and $p_j = 1/40$ for every component, giving $L = 0$. The selected top-five slots break a frequency tie; they are not a five-component answer certified by the accepted-set rule.

Two-term core. Two linearly independent components appear in every support. The generator varies the remaining three terms, preferring balanced use of the varying pool, rejecting duplicate supports and those with rank below four. The resulting 32 supports have a mean L of approximately 0.3822 across families. Its small family variation is retained in the slope calculation.

Four-term core. Four linearly independent components appear in every support. The fifth ranges over 32 distinct catalogue components. The four core frequencies are one, the 32 varying frequencies are $1/32$, and the remaining four frequencies are zero. Equation 1 gives $L = 0.778571\ldots$. The core can be identified from accurate component evidence, but the consumer must still interpret the reported profiles and produce an accepted support.

Table 7: Frozen construction levels. Only the first three enter the primary fixed-count interaction. Additional levels describe the construction over a wider range; they are not extra primary observations.

Level	$\lvert\mathcal{S}_f\rvert$	L	Construction
Balanced	32	0.000	Four appearances per catalogue component.
Two-term core	32	≈ 0.382	Two universal terms, three varying terms.
Four-term core	32	0.779	Four universal terms, 32 alternatives for the fifth.
Unique	1	1.000	One accepted fitting support.
Profile-grouped core	32	0.779	Four-term core; fifth-term alternatives concentrated in fewer profile classes.
Four disjoint supports	4	0.143	Four nonoverlapping five-term fitting supports.

The last two levels were called `hicount` and `lodiffuse` in the archive. The profile-grouped level shares the four-term-core level's component overlap and answer count, while changing how the varying terms group into visible profiles. The disjoint level has fewer accepted answers but low overlap. These levels help describe the joint roles of answer count and report granularity. They are not substituted for the three-level primary comparison.

A.4 WHY THIS INDEX MEASURES SHARED COMPONENTS

For independently and uniformly drawn $S, S' \in \mathcal{S}_f$,

$$\mathbb{E}|S \cap S'| = \sum_j q_j^2 = k^2 \sum_j p_j^2.$$

Thus L rescales average component agreement between acceptable supports, relative to balanced frequencies and complete agreement. This is a stability interpretation of an established chance-corrected form (Kuncheva, 2007; Nogueira et al., 2018), not a new information-theoretic quantity. Applying it to an accepted set is useful because that set describes which answers the verifier rewards. Connections to a version space or committee agreement are conceptual analogies, not additional experimental claims (Mitchell, 1982; Seung et al., 1992).

Neither this identity nor L alone ranks arbitrary feedback channels. A report may reveal component marginals, joint constraints, a complete answer, or unrelated information. Here the report construc-

tion is held fixed while the accepted-set construction changes. Its useful content changes with the component frequencies. That coupling is part of the experimental object, not an assumption that every equally long report has equal information.

A.5 Report generation and visible aliases

The full report sorts components by decreasing inclusion frequency, breaking ties by catalogue index, and selects five. A slot reveals:

- the variables used by the term;
- its expression depth;
- its unary- and binary-operator profile;
- the selected component's inclusion frequency across the accepted set.

It does not reveal the catalogue index, expression fingerprint, full expression tree, or a designated complete support. Distinct terms can have the same variables, depth, and operator profile. The frequency belongs to a selected component; it is not automatically the sum of frequencies of every term in its profile class.

The fit-score field is the true-world development R^2 of the accepted supports, which is one in this construction. The report renders frequencies as percentages to one decimal place. Profiles occupy fixed-width fields, and full and sham reports use the same headers, slot count, and frequency-field layout.

For the partial sham, a deterministic family/level/arm seed chooses three positions to replace. Decoys are different catalogue components, preferably with the same surface shape—depth and counts of variables and operators—but a profile outside the original top-slot profiles. The displaced frequency value remains in its slot. The generator avoids reusing a decoy. The all-swapped arm applies this operation to all five profiles. The score field and numerical task stay unchanged. The stored flags indicate retained or substituted slots, not a separate truth check. At the balanced level every component appears in four of 32 supports, so both controls preserve correct component frequencies. At other levels, a retained value may no longer match its substituted component.

The implementation's nominal size check is $5 \log_2 P + 15$, where P is the number of catalogue profiles and each frequency field receives a fixed three-bit allowance. Its spread across levels is zero, within the declared ± 0.5-bit tolerance. This is an encoding-size proxy. It is not measured mutual information, a compression theorem, or an equality of compatible-support counts. The two reported surface checks were a deterministic discriminator with AUC 0.5 at each level and a frontier-model forced-choice judge at chance. These check the tested cue classes; they do not make original and substituted reports indistinguishable to every consumer.

A.6 A worked scoring and report example

Consider a separate schematic example with only two accepted supports:

$$\mathcal{S} = \{\{a, b, c, d, e\}, \{a, b, c, d, f\}\}.$$

For clarity, suppose profiles uniquely name the letters. A full report can contain $a : 1$, $b : 1$, $c : 1$, $d : 1$, and $e : 1/2$. The omitted term f also has frequency $1/2$. A partial sham can retain $a : 1, b : 1$ and replace the other profiles with $g : 1, h : 1, i : 1/2$. It retains the five frequency fields and fit score. In this illustration, replacing the sham with full restores truthful entries and removes misleading ones.

A commitment $\{a, b, c, g, h\}$ shares three terms with its nearest acceptable support. It scores $G = 3/5$ and $E = 0$. The support $\{a, b, c, d, f\}$ scores one on both outcomes even though the report's fifth named slot is e. This shows why the report describes a set rather than prescribing one complete answer. The actual experiment has 32 primary supports and ambiguous profiles; this example is only an explanation of the contract.

B MODEL PROTOCOL, SCORING, AND INFERENCE

B.1 WHAT THE MODEL SEES AND PRODUCES

The prompt describes a target of the form $y = c_1T_1 + \cdots + c_5T_5$. It tells the model to propose term structures, without coefficients, because least squares fits the coefficients. It lists the six positive-valued inputs, the log-monomial grammar, and the depth limit. It also says that many supports fit equally well and that the report summarizes structures recurring across acceptable answers. The prompt prints the first ten development observations.

The LLM generates terms in this grammar; it is not given a printed list of all 40 catalogue expressions. Catalogue identities define the accepted supports and fingerprint-based evaluation. The reader, by contrast, uses the explicit catalogue for its enumerated reference calculation. This is one reason to describe the reader's information and representation rather than claim identical internal inputs.

Each proposal call requests eight new laws. A law is a line in the following format, with semicolons separating five terms:

```
LAW 1:  <term> ; <term> ; <term> ; <term> ; <term>
```

Two proposal rounds precede a final call requesting exactly one such line. The later calls include the same fixed report and the archive of distinct parseable proposals with their development fit scores. No exact accepted-set membership or graded recovery is returned during search.

All arms share this numerical proposal-fit channel. `None` omits the fixed report, not the observations, grammar, previous proposals, or their fit scores. `Scalar` adds the accepted-support fit-score field without the component slots. These definitions are necessary for interpreting the controls as run, rather than as a comparison with a consumer receiving no task information at all.

B.2 CALL LIMITS AND EVALUATION ACCOUNTING

The two proposal rounds each reserve eight evaluation slots. Missing or unparseable proposed laws consume their slots instead of earning free retries. Parseable proposals are fit by the frozen routine, and duplicate canonical proposals are not repeated in the visible archive. A final model call makes the scored commitment. The loop has an allowance of 48 verifier evaluations, but its configured two-round proposal stage uses 16 charged slots; the final commitment is then scored separately.

The controlled runner uses an $8{,}192$-token call cap. For the GPT-5.5 and GPT-5.6 reasoning endpoints it requests medium effort with a $4{,}096$ reasoning-token minimum; the other routes use their configured adapter behavior with a $2{,}048$ minimum. The study uses seeds zero and one, with deterministic round-specific request seeds. The experimental records retain requested and served identifiers, finish reasons, token accounting, and response hashes. Matching these settings within a model does not imply equal consumed tokens across arms, or equal computation between models and readers.

B.3 MODEL PANEL AND COMPLETE COVERAGE

The complete panel contains the 11 endpoints in Table 9. These are the labels used during the experiments, not claims about permanent public model availability or independent draws from an architectural population. The Opus-5 route returned content-filter refusals on this task and was excluded as unmeasurable; it is not treated as a twelfth zero-scoring consumer.

The complete archive contains $24 \times 2 \times 11 \times 6 \times 5 = 15{,}840$ cells: 24 families, two seeds, 11 models, six levels, and five arms. The primary subset uses three levels and two arms, yielding $3{,}168$ cells, or $1{,}584$ full/sham pairs. Each arm–level mean in the main table has 528 cells. The smaller eight-family records in the development archive are prefixes of this collection, not the population behind the headline numbers.

Table 8: Resource units used in the central comparisons. These are configured limits or accounting units, not measured equality of wall time or consumed tokens.

Consumer/study	Calls and allowance	Additional work or qualification
Controlled LLM	Two eight-proposal calls, one final call; 48-evaluation allowance; 8,192 output-token limit per call	16 charged proposal slots under this schedule; final scoring separate.
Expected-overlap reader	48 charged verifier calls per reader run	Offline enumeration of 658,008 supports and exact optimization of the working utility.
Fixed-report SAT	One assignment call per arm/formula; 12,000-token limit	Full and sham on 1,000 formulas per model; none on a frozen 250-formula subset.
Own-proposal SAT	One shared initial call, one revision per arm; 16,000-token limit per call	Four arms share the same parsed first answer; failed first answers do not produce ordinary four-arm revisions.
SAT reader ladder	Marginal update, then B greedy flips	A flip is not an LLM token, call, or verifier evaluation.
Complete SAT solver	Branching decisions with unit propagation	Separate solver budget; not matched to LLM computation.

Table 9: Model panel and descriptive fixed-count graded interactions. All 11 point estimates are positive. These are the stored per-model full-minus-partial-sham slopes, not eleven independent replications of the family-level statistical test.

Endpoint label	Graded interaction
GPT-5.6	+0.460
GPT-5.5	+0.455
GPT-5.4-mini	+0.409
Claude Opus-4.8	+0.370
Claude Opus-4.7	+0.285
Claude Opus-4.6	+0.374
Claude Sonnet-5	+0.445
Claude Sonnet-4.6	+0.481
Claude Sonnet-4.5	+0.409
Claude Haiku-4.5	+0.406
MAI-Code-1-Flash	+0.515

B.4 COMMITTED-ANSWER SCORING

The extractor reads the requested law lines and parses terms in the declared grammar. It does not execute a free-form generated program. Parsed terms are canonicalized and mapped to their numerical fingerprints on the frozen probe. Duplicate fingerprints collapse to one set element. As a result, the general F_1 expression in Equation 2, rather than an assumed five-element overlap for every output, is the scoring rule.

The scorer compares that set against each member of the declared acceptable set. Graded recovery is its maximum F_1; exact recovery requires equality with an accepted fingerprint set. An unparsed commitment receives zero on both. Development and holdout fit are recorded separately and do not substitute for recovery. The final commitment can differ from every earlier proposal; the scorer never awards best-of-pool credit instead.

An independent AST-based parser checks the complete primary raw-response population. It agrees on $3{,}059$ canonical commitments and scores and on the unparsed-answer zero rule for the remaining 109. There are no discrepancies in this census. It extends a previously fixed 132-cell sample and 16 unparsed diagnostic cases, with the parser fixed before those response inspections. The response hashes and matching model/seed records connect each scored commitment to its retained final response.

B.5 THE FIXED-COUNT INTERACTION

For family f and level ℓ, let

$$d_{f\ell} = \frac{1}{22} \sum_{m=1}^{11} \sum_{s=1}^{2} \left(G_{f\ell ms,\mathrm{full}} - G_{f\ell ms,\mathrm{sw40}} \right).$$

The family slope is the ordinary least-squares coefficient of $d_{f\ell}$ on the three frozen $L_{f\ell}$ values, with an intercept:

$$b_f = \frac{\sum_\ell (L_{f\ell} - \overline{L}_f)(d_{f\ell} - \overline{d}_f)}{\sum_\ell (L_{f\ell} - \overline{L}_f)^2}, \qquad \widehat{\beta} = \frac{1}{24} \sum_{f=1}^{24} b_f.$$

The exact interaction uses the same construction with E in place of G. For a reader comparison, the corresponding reader arm gap is subtracted inside each family and level before taking the slope. This preserves the pairing.

The reported two-sided sign-flip calculation uses $200{,}000$ random draws for the 24-family panel. The percentile interval uses $20{,}000$ family bootstrap resamples. Models, seeds, levels, and arms remain grouped with their family. The Monte Carlo sign-flip resolution explains the quoted 5×10^{-6} value. The analysis also records family signs and a one-sample t summary. Pointwise arm/level intervals in Figure 2 use the family bootstrap but are not intervals for the slope.

This estimand describes a response to the specified accepted-set construction. L is its scalar summary, not a separately randomized treatment purified of every other feature of the accepted set. The controls remove changes in data, catalogue, answer count, support size, report schema, and allowance; they do not make no-report recovery constant. The paired comparison is therefore reported together with the absolute arm means.

C ADDITIONAL CONTROLLED RESULTS

C.1 THE COMPLETE SIX-LEVEL MEANS

Table 10 reports the saved graded means for the complete panel. The three primary levels retain the fixed accepted-answer count. The other levels broaden the description: the unique-support level has the largest full-minus-none mean, while the four-disjoint-support level has a much smaller advantage despite its low answer count.

The full-minus-none differences for the additional levels are $+0.507$ for unique, $+0.352$ for the profile-grouped core, and $+0.089$ for four disjoint supports. The controlled primary interaction

Table 10: Complete-panel graded arm means. The first three rows are primary. Every arm–level entry averages 528 cells. Values are rounded from the saved complete-grid summary; the analysis uses unrounded scores and family-specific L.

Level	L	None	Scalar	All swapped	Partial sham	Full
Balanced	0.000	0.391	0.395	0.411	0.435	0.379
Two-term core	0.382	0.307	0.309	0.327	0.403	0.569
Four-term core	0.779	0.265	0.262	0.195	0.360	0.631
Unique	1.000	0.061	0.066	0.007	0.199	0.568
Profile-grouped core	0.779	0.265	0.262	0.200	0.378	0.617
Four disjoint supports	0.143	0.257	0.263	0.223	0.279	0.347

remains the three-level fixed-count result. We do not treat the six-level ordering as an independent replication or combine the count-changing rows into the primary test.

There are two visible changes across the primary levels. Full recovery rises from 0.379 to 0.631, while no-report recovery falls from 0.391 to 0.265. A balanced set spreads partial-credit opportunities over the catalogue; a concentrated set rewards a narrower collection of parts. Both the informative report and the scoring geometry therefore matter. The observed no-report trend is part of the defined task, not evidence that an otherwise fixed-difficulty harness drifted.

At $L = 0$, the partial-sham mean exceeds the full mean by 0.056 at displayed precision. This does not justify calling correctness intrinsically harmful, identifying a psychological mechanism, or treating the full/no-report difference as an equivalence result. It establishes an ordering of observed graded scores under that report construction. The positive primary slope describes how this ordering changes with the accepted set.

C.2 Variation across models, seeds, and families

The main interaction is positive on all 11 endpoint labels (Table 9), with descriptive estimates ranging from $+0.285$ to $+0.515$. It is positive on 23 of 24 family slopes. These sign summaries supplement, rather than replace, the family-clustered estimate.

The two seed-specific graded interactions are $+0.436$ and $+0.402$; the corresponding exact interactions are $+0.147$ and $+0.156$. The family-level seed correlation is 0.93. A saved crossed variance decomposition assigns approximately 34% of graded variation to families, 20% to family-by-model variation, and 0.4% to seeds (0.0% for exact). The important source of replication for inference is thus the family, not a large number of repeated seeds.

Family heterogeneity also appears in exact recovery. An exploratory comparison of even- and odd-numbered families gave 1.89% and 6.69%, respectively. Family identifiers enter as random-seed offsets; the checked rank, catalogue size, support size, and variable-use controls did not explain that split. We therefore retain all families rather than selecting an easier half. The family-clustered analysis preserves this variation.

C.3 Primary and secondary analysis status

The fixed-count levels and full/partial-sham contrast were specified before the run. The original reader comparison used a weak hand-built policy. The calibrated expected-overlap reader, the expanded reader bracket, and specification sensitivities are secondary analyses of the same completed model outcomes. Their role in the present paper is to interpret the stable within-model effect, not to turn a later comparator into the original preregistered policy.

A frozen inventory of earlier analyses contained 67 tests, five labelled confirmatory and 62 exploratory. The three positive tests in that confirmatory group survived both Holm and Benjamini–Hochberg correction; the other two were sham-discriminator null checks. These inventory counts do not describe every later external reanalysis. Under family-wise control of that earlier inventory, the $+0.115$ graded difference against the calibrated hand-built reader and the Boolean prefix did not survive. They remain exploratory here.

The current paper therefore gives the primary within-model comparison its own estimate and sampling unit, labels the named-reader differences as secondary, and reports external analyses with their own populations. It does not count alternative readers, nested prefixes, two seeds, or decomposed slopes as separate confirmations.

D THE CALIBRATED EXPECTED-OVERLAP READER

D.1 THE WORKING MODEL OF A REPORT

The report leaves two kinds of uncertainty: a profile may name several catalogue terms, and most components are not assigned to a reported slot. The reader makes both completions explicit. It treats each slot as truthful, with no substitution-position flags. A slot's identity is exchangeable among terms in its profile class, without replacement when several slots share that profile. Components not assigned to a reported slot share the remaining inclusion mass.

Let $r_1, \ldots, r_k$ be the frozen component frequencies in the report record. For a profile class g, let n_g be its number of catalogue terms, h_g its number of displayed slots, and R_g the sum of the displayed frequencies in those slots. The unreported inclusion frequency is

$$u = \frac{k - \sum_{s=1}^{k} r_s}{M - k}.$$

For component j in profile class g, the working inclusion probability is

$$\pi_j = \frac{R_g}{n_g} + \left(1 - \frac{h_g}{n_g}\right) u. \tag{4}$$

The resulting probabilities lie in $[0, 1]$ and sum to k on the frozen instances. In scalar and no-report arms, the vector is uniform, $\pi_j = k/M$.

This is a stated completion model, not recovery of the true joint distribution over acceptable supports. It assigns equal status to component identities the profile does not distinguish and treats the reported marginals literally. In a sham arm it therefore also trusts the substituted profiles and their retained frequencies. The true accepted set is not an input to this marginal-completion rule.

The saved reference calculation reads the structured slot frequencies underlying the report; the LLM rendering prints percentages to one decimal place. The reader thus uses the frozen numerical fields rather than reparsing the rounded text. Along with explicit catalogue access, this is part of the stated reference interface, not a claim of byte-identical input representations.

D.2 THE SELECTION OBJECTIVE AND THE EVALUATED OUTCOME

For a random size-k support $\widetilde{S}$ under the working model and a size-k commitment C,

$$\mathbb{E}\left[\frac{|C \cap \widetilde{S}|}{k}\right] = \frac{1}{k} \sum_{j \in C} \pi_j.$$

The reader selects a fitting support that maximizes this expected overlap. The equation uses only marginal inclusion probabilities, so it does not require reconstructing joint co-occurrence in $\mathcal{S}_f$.

The benchmark instead assigns $\max_{S \in \mathcal{S}_f} F_1(C, S)$ and exact membership. Expected overlap with one working-model draw and maximum overlap with the actual accepted set are different functions. The reader is scored on the latter outcomes after selection, just as the model is. It is optimal for its declared selection objective among feasible catalogue supports, not certified optimal for the benchmark's graded score, the exact score, or a feedback-by-overlap interaction.

This distinction is also why a reader can have a larger or smaller measured arm contrast without contradicting its per-arm optimization. Maximizing a utility in each arm does not maximize the difference of evaluated outcomes between two arms, much less the slope of that difference. The comparison is a concrete reference decision rule in the classical decision-theoretic sense (DeGroot, 1970); no uninstantiated value-of-information bound is used to explain the experimental coefficient.

D.3 Enumeration, ties, and charged calls

The reference enumerates all $\binom{40}{5} = 658{,}008$ catalogue supports. It retains the rank-four supports, the data-consistency condition of this construction, and computes their expected-overlap utility. It chooses a global maximizer. Ties within 10^{-12} are resolved by a seeded draw. If the marginal vector is uniform, all fitting supports tie, and the commitment is drawn from that feasible set.

For each family, level, and arm, 32 reader seeds average over ties. Each run charges a batch of 48 distinct feasible supports, including its selected commitment. The remaining batch members do not change the already-computed argmax. Across 24 families, six levels, five arms, and 32 seeds, this gives $23{,}040$ runs and $1{,}105{,}920$ charged calls. These runs are computational repetitions, not new independent task families.

The archived check verifies that every diagnostic commitment attains the exhaustive utility maximum within 10^{-12} and that every run has 48 charged calls. Those checks establish objective and accounting correctness. They do not charge the offline enumeration as 48 units of total computation. In particular, a consideration frontier whose first item is already ordered by this utility has its optimum at one considered candidate; constructing that ordering is not free.

D.4 Calibration on non-diagnostic arms

The uncalibrated reader always selects a feasible catalogue support. Models sometimes fail to do so. To align their average non-diagnostic recovery, the reader uses a global fitter-success probability ρ: with probability ρ it emits its selected support, and otherwise the empty support, which receives zero. Calibration changes this scalar mixture, not the report-completion rule.

Let t_a be the model's graded mean in non-diagnostic arm a and r_a the uncalibrated reader's corresponding mean, both across the six-level archive. The fitted coefficient is

$$\rho = \mathrm{clip}_{[0,1]} \frac{\sum_{a\in\{\mathrm{scalar,none}\}} r_a t_a}{\sum_{a\in\{\mathrm{scalar,none}\}} r_a^2}.$$

It is 0.832594 in the stored calculation. Neither full nor partial-sham outcomes enter this fit. The same coefficient then multiplies the reader's graded and exact expectations in every arm and level.

Table 11: Non-diagnostic calibration, using six-level graded means. Calibration matches these pooled arm means; it does not match every level or computational capability. The diagnostic arms are excluded from the fit.

Arm	Model target	Reader before	Reader after	Error
Scalar	0.259554	0.310069	0.258162	−0.001392
None	0.257751	0.311241	0.259138	+0.001387

The pooled arm root-mean-square error is 0.001389. Across individual level/arm means, it is 0.025289, with a maximum absolute error of 0.048859. The calibration therefore aligns selected pooled non-diagnostic outcomes rather than enforcing level-by-level equality. The calibrated reader's primary interactions are $+0.440$ graded and $+0.486$ exact.

E Reader sensitivity and statistical precision

E.1 What changes when the reader changes

The central model interaction does not depend on a mechanical reader. Subtracting a reader asks the more specific question of what remains relative to that policy. Table 12 gives the stored graded point estimates for seven policies. Their information and optimization procedures differ.

The hand-built policies differ in how many report slots they use, how they generate compatible candidates, and how they spend their search allowance. The calibrated hand-built policy matches scalar/no-report recovery with error below 0.002 and leaves a graded difference of $+0.115$

Table 12: Secondary reader bracket on the same completed controlled study. Model interaction is $+0.419$ throughout. True-set selection policies receive information not supplied to the LLM or the expected-overlap reader.

Reader	Reader interaction	Model − reader
Two-slot hand-built reader, budget 12	$+0.164$	$+0.255$
Intermediate hand-built reader	$+0.195$	$+0.224$
True-set candidate selector, calibrated	$+0.221$	$+0.198$
True-set candidate selector, ungated	$+0.264$	$+0.155$
Calibrated hand-built reader	$+0.304$	$+0.115$
Profile/frequency greedy, 48 calls	$+0.365$	$+0.054$
Calibrated expected-overlap reader	$+0.440$	-0.021

$[+0.046, +0.180]$. The stronger greedy reader leaves $+0.054$ $[-0.012, +0.118]$. These are secondary reader-relative intervals, not changes in the within-model result.

The true-set selectors use a profile-biased candidate generator with a 48-call allowance, retain numerical fitters, and select the candidate with the best graded score against the *actual* $\mathcal{S}_f$. They break ties using exact membership, profile agreement, and the seed. Supplying the true accepted set symmetrically in both report arms makes these policies informationally privileged; it does not make them fair same-report consumers or universal upper bounds on an interaction. Their role is a selection sensitivity on the generated candidate pool.

The calibrated true-set policy uses a fitter-success coefficient of 0.834 from scalar/no-report outcomes (arm RMSE 0.0012). Its model-minus-reader differences are $+0.198$ $[+0.117, +0.273]$ graded and -0.001 $[-0.080, +0.088]$ exact. These numbers do not identify a unique source of model reasoning. Different candidate generation, information, and selection rules can change the reader contrast.

E.2 Precision of the named-reader comparison

The calibrated expected-overlap comparison uses the same 24 family units as the primary study. Rebuilding its stored arrays returns a graded difference of -0.020987 and an exact difference of -0.334402, which give the rounded main-table values. The two-sided graded interval is $[-0.102, +0.058]$, with $p = 0.62$. The exact interval is $[-0.461, -0.222]$, with $p = 5 \times 10^{-6}$.

The precision analysis compares family sign-flip, percentile-bootstrap, and one-sample-t conventions, reporting the widest stated bound. The additional graded interaction has a one-sided 95% upper bound of $+0.051$. The corresponding 80%-power minimum detectable effect is 0.122 slope units. The first is a bound from the observed data; the second is a design-resolution calculation. They are not interchangeable.

Table 13: Precision summaries for model-minus-expected-overlap-reader interactions. They concern slopes, not mean recovery, regret, or the fraction of a gain caused by model reasoning.

Quantity	Stored value
Graded point estimate	-0.021
Graded one-sided 95% upper bound	$+0.051$
Graded 80%-power minimum detectable effect	0.122
Smallest reported symmetric equivalence margin	± 0.093
Exact point estimate	-0.334
Exact one-sided 95% upper bound	-0.227

The precision is enough to characterize a large reader-relative exact difference and limit the pooled additional graded slope. It does not settle small model-specific effects or establish equality. In particular, a null test and a coarse equivalence margin do not turn the reader into a ceiling on all possible consumers.

E.3 A SAVED SPECIFICATION SENSITIVITY

A separately fixed grid examined $1{,}008$ combinations: seven readers, two outcomes, three tie-breaking rules, two clustering units, three seed handlings, all models versus dropping the most influential model, and six levels versus the three fixed-count levels. This grid characterizes reader-relative analysis sensitivity. The main estimate continues to use families and the fixed-count levels.

Across the grid, the median reader-relative estimate is $+0.084$ (interquartile range $[+0.014, +0.155]$); 80.4% of specifications are positive and 53.4% are significantly positive under their own stated tests. Reader choice accounts for 54.3% of the spread in the saved variance decomposition. These percentages describe a chosen grid of analyses, not probabilities that a scientific claim is true or a causal allocation of model value.

Within the expected-overlap-reader portion, the median difference is -0.037 (interquartile range $[-0.145, +0.025]$); 33.3% are positive, none significantly positive, and 41.0% significantly negative. The practical conclusion is the same as in the main text: a reader-relative result needs a named reader, objective, and information set. The controlled full-minus-partial-sham result does not rely on selecting a favorable reader.

F EXTERNAL FIXED-REPORT SAT STUDY

F.1 BENCHMARK, ACCEPTABLE ANSWERS, AND OVERLAP

SATLIB `uf20-91` contains $1{,}000$ satisfiable random 3-SAT formulas with 20 variables and 91 clauses (Hoos & Stützle, 2000). The formulas were created independently of this study. Every satisfying assignment is acceptable; there is no authored subset of solutions playing the role of the term construction's declared accepted supports.

The study enumerates all 2^{20} assignments per formula using packed bitsets. It computes the solution count and each variable's true-value marginal over the complete solution set. No sampling approximation enters these fixed-report overlap values. The $1{,}000$ formulas have L between 0.175 and 1.000, with 465 distinct values and interquartile range 0.283. Solution counts range from one to 178, with median six.

To connect SAT to the component index, regard an assignment as a support containing one signed value per variable. There are $M = 2n$ signed components and $k = n$ selected ones. If q_i is the true-value marginal, the false-value marginal is $1 - q_i$. Substitution into Equation 1 gives

$$L = \frac{1}{n} \sum_{i=1}^{n} (2q_i - 1)^2.$$

The index is large when many variable values are shared across solutions. Its relation to solution count is measured rather than removed by construction.

F.2 REPORTS, PROMPTS, AND OUTCOMES

A frozen seed, 20260616, selects ten of the twenty variables and four of those ten for inversion. Variable selection does not depend on a model answer. Full reports the ten true marginals to two decimal places. Sham replaces four by $1-q_i$, retaining the other six. None gives the formula without a marginal report. The consumer makes one final assignment, not a two-round proposal-conditioned revision.

Inversion preserves the displayed confidence magnitudes $|2q_i - 1|$. The pre-run audit found identical character length and confidence multisets on all $1{,}000$ formulas. An enumerating consumer could nevertheless detect disagreement with the formula's solution set. Layout and magnitude matching are therefore presentation controls, not information symmetry.

The primary score is

$$G_{\mathrm{SAT}}(a, \mathcal{S}) = 1 - \frac{1}{n} \min_{s \in \mathcal{S}} d_H(a, s),$$

where d_H is Hamming distance. Exact satisfaction is a secondary binary outcome. Unparsed assignments score zero; they are not removed to condition the analysis on a successful output format.

The final call settings are medium reasoning effort, a 256-token reasoning minimum, 12,000 output tokens, temperature zero, and the frozen request seed.

F.3 Frozen reader and the population rule

The fixed-report reader is distinct from the term catalogue's calibrated expected-overlap reader. It sets each reported variable to $\mathbf{1}[q_i > 0.5]$ and fills the unreported variables uniformly. Its graded score is computed in exact expectation. Before model outcomes, its arm means were 0.5740 for none, 0.6123 for sham, and 0.7786 for full. Its full-minus-sham interaction was +0.1314. This establishes a useful reference trend for this SAT report, not a universal optimum.

The study fixes a ceiling rule: a model with full-arm mean graded recovery above 0.95 is excluded from the confirmatory moderator population. GPT-5.6 scores 0.9982 and GPT-5.5 scores 0.99965, so both are at ceiling under that rule. Opus-4.8 scores 0.93025 and is retained. The confirmatory result therefore has 1,000 formula units and one model, not three-model confirmation.

Table 14: Fixed-report SAT graded means and parsing rates. Full and sham each use 1,000 formulas per model; none uses a frozen random subset of 250. The none means are descriptive and are not paired full-suite means. All scores include parse failures as zero.

Model	Full	Sham	None	Parse full	Parse sham	Parse none
Opus-4.8	0.9303	0.6517	0.7578	1.000	0.960	0.928
GPT-5.5	0.9997	0.9895	0.9968	1.000	0.995	1.000
GPT-5.6	0.9982	0.9110	0.9626	1.000	1.000	1.000

The final archive has 6,750 cells: three models each have 1,000 full, 1,000 sham, and 250 none calls. Its overall parse rate is 0.9907, with zero recorded run errors. The remaining parse asymmetry is visible in Table 14; the as-run contrast includes the resulting score loss.

F.4 Association estimates and answer-count control

For the retained model, the OLS slope of the formula-level full-minus-sham graded gap on L is +0.213. The HC3 95% interval is $[+0.145, +0.280]$, the 100,000-draw bootstrap interval is $[+0.145, +0.279]$, and the 200,000-draw sign-flip p value is approximately 0.0002. The mean gaps by increasing overlap quartile are +0.227, +0.268, +0.301, and +0.316.

Overlap is correlated with the number of solutions: $r = -0.75$ with log count and $r = -0.52$ with raw count. Table 15 keeps the within-model and model-minus-reader adjustments separate.

Table 15: Fixed-report SAT slopes on L. Count controls do not change the formula sampling unit. "Model minus reader" subtracts the frozen SAT reader's corresponding arm contrast before fitting.

Outcome/contrast	Count covariate	Slope [95% interval]
Graded, full − sham	None	+0.213 [+0.145, +0.280]
Graded, full − sham	Solution count	+0.169 [+0.086, +0.251]
Graded, full − sham	Log solution count	+0.173 [+0.056, +0.290]
Graded, model − reader	None	+0.081 [+0.017, +0.146]
Graded, model − reader	Solution count	+0.046 [−0.032, +0.124]
Exact, full − sham	None	+0.052 [−0.119, +0.223]

The unadjusted model-minus-reader slope has sign-flip $p = 0.029$. After count control its interval crosses zero. The within-model graded association remains positive, while a positive additional model interaction is not established by the adjusted comparison. The exact slope has $p = 0.62$. These are different questions and different outcome scales.

The study does not observe $L = 0$. It cannot establish a zero-crossing at that value by extrapolating the fitted line. Nor does inversion equalize report value across formulas: inverting a marginal near one changes the evidence more than inverting one near one half. The experiment asks for an external association under this fixed report construction.

F.5 Collection amendments and retained checks

The final token settings replaced an earlier cap that starved reasoning responses. Before any hypothesis test, the ceiling and parse-rate checks showed much lower parsing in the unhinted and sham arms. The starved grid was excluded and collection restarted from zero under the $12{,}000$-token cap. The none arm was also reduced, for resource reasons, to a frozen random 250-formula subset per model; it was designated descriptive and does not enter the full/sham primary slope. These changes account for the final $6{,}750$ cells rather than the initially planned $9{,}000$.

An independent verifier re-enumerated 12 audited formulas and reparsed all $6{,}750$ responses. It found 18 parser disagreements (0.27%), all on Opus-4.8 and 17 in sham. In 13, the stricter primary parser rejected an answer recovered by the lenient parser. The stored sensitivity changes the headline slope from 0.2128 to 0.2109. The original primary result is retained, with this parser effect reported rather than silently rescored.

G Own-proposal feedback on the common corpus

G.1 Corpus and measurement on the two tiers

All ten models receive the same 90 formulas: 45 from `uf20-91` and 45 from `uf50-218`. The latter have 50 variables and 218 clauses (Hoos & Stützle, 2000). The panel comprises GPT-5.6, GPT-5.5, GPT-5.4-mini, Opus-4.8/4.7/4.6, Sonnet-5/4.6/4.5, and Haiku-4.5. The corpus spans recorded $L \in [0.289, 0.980]$. The corpus, models, arms, and primary contrast were fixed before collection.

On uf20, marginals and counts come from exact enumeration. On uf50, the saved preparation runs a bounded WalkSAT-style procedure with a $200{,}000$-flip budget. It records distinct satisfying assignments and computes their empirical marginals. Instances with fewer than two found solutions receive no estimate and are not eligible for this corpus selection. Thus the uf50 quantity is $\widehat{L}$ and its count is the number of found solutions, not a certified total number of satisfying assignments. Truthful reports on this tier faithfully report those saved marginals.

The selection procedure groups formulas by recorded solution count and selects contrasting overlap values within groups, then trims to the planned 45 formulas per tier with seed 20260825. In the resulting mixed corpus, the correlation between overlap and recorded log count is -0.362. Neither grouping nor subsequent regression makes the moderator randomized, and bounded solution finding need not sample all solutions uniformly. The exact final-solving outcome, however, is checked directly against the formula's clauses.

G.2 A shared start and four revisions

For each model–formula cell, the model first receives the formula and returns one complete assignment. That parsed assignment is shared unchanged across the four revision arms. This prevents a comparison between different starting answers from being mistaken for a feedback effect.

The diagnostic procedure considers a selected pool of twelve variables. For first assignment value a_i and saved true-value marginal q_i, its disagreement score is

$$d_i = \begin{cases} 1 - q_i, & a_i = 1, \\ q_i, & a_i = 0. \end{cases}$$

It flags the five largest disagreements within the pool. The four revisions are:

Full. Report the five flagged variables and their saved marginals, with the message that the current value looks wrong.

Partial sham. Retain three original slots. Replace two with variables less contradicted by the solution marginals, reporting values that oppose the model's current choices at the displaced slots' confidence magnitudes.

Schema. Name the same five variables as full and mark them "checked," without their marginal values or the "looks wrong" assertion.

None. Show the same previous attempt and ask the model to reconsider it without a diagnostic block.

In this study `sw40` means two of five slots replaced, whereas in the term construction it retains two of five original slots. We use "partial sham" in the narrative and state the substitution count rather than relying on that historical name. The schema arm preserves the identities selected by the diagnostic rule; it therefore does not remove every possible selection cue. The full/none comparison changes report presence and format along with the supplied values.

For a displaced confidence $c = |2q - 1|$, a decoy given value one by the first answer is reported with $(1 - c)/2$; a decoy given zero is reported with $(1 + c)/2$. Variable indices are padded to a fixed width and frequencies are printed to two decimal places. The saved 300-case report audit found character-length and magnitude-multiset agreement in every case. As in the fixed-report study, this is not a claim of indistinguishable information to a solver.

The initial and revision calls use medium reasoning effort, a 256-token reasoning minimum, 16,000 output tokens, temperature zero, and the study's request seed 20260823. The response format is a single assignment line containing every variable. The original call procedure allows retries after a delivery error or empty response; it does not choose a best answer by the held-out outcome. These settings differ from the fixed-report study and are not an additional manipulation inside its population.

G.3 Eligibility and the current estimand

Ten models on 90 formulas give 900 planned model–formula rows. One shared first call and four revisions give 4,500 planned logical call positions, not 4,500 delivered API responses: revisions are skipped after an unusable first answer, and retries or caching affect delivery. All 900 rows have archive records.

There are 834 parsed shared first answers, 64 first-answer run errors, and two non-error first-answer parse failures. The current main estimate uses all 834 eligible rows without a full-arm headroom screen. Second-answer run errors and unparseable outputs keep their original unsuccessful scores. It therefore describes revision after a usable first answer, not an unconditional all-900 success probability.

For row i, the analysis fits the exact-solving full-minus-sham difference to an intercept, recorded overlap, corpus tier, and the recorded log solution count. On uf20 the count is exact; on uf50 it is the found-solution count described above. Rows receive equal weight. The pooled coefficient is not the simple average of ten model-specific slopes.

Uncertainty resamples the 90 base formulas with all of their model–arm rows. The 4,000 bootstrap draws, seed 20260916, are shared across the current fits. This retains model dependence within a formula. The all-usable fit and its later decomposition were specified after outcomes were known, then frozen before their calculation. They remain exploratory pointwise analyses.

G.4 The all-usable estimate and sensitivities

Table 16 reports the all-usable estimate, model-intercept and delivery sensitivities, and the earlier screened population for comparison. No model answer or original outcome is changed by these fits.

The earlier screen removed model–tier strata with observed full-arm solve rates outside $[0.02, 0.95]$. It retained 440 rows and excluded nine of twenty strata. Because this screen uses full-arm outcomes, it was not a preregistered population definition. The present main text instead uses the unscreened eligible population.

The screened coefficient's original bootstrap interval was $[+0.249, +0.835]$. Table 16 uses the later shared bootstrap stream and reproduces the same $+0.541$ point estimate. The paired interval for the population difference conditions on the observed screen; it is not a causal estimate of screening bias or unconditional post-selection inference.

The earlier screen left only GPT-5.4-mini with retained observations in both tiers. Its uf20-only estimate was $+0.536$ $[+0.201, +0.875]$ on 313 rows; uf50 alone gave $+0.473$ $[-0.132, +1.144]$

Table 16: Common-corpus slopes of the exact-solving gap on recorded overlap. All rows use 90 formula clusters. Intervals are exploratory, pointwise percentile intervals from shared formula resamples, not simultaneous confidence statements.

Population or fit	Rows	Slope [95% interval]
All usable shared first answers	834	$+0.385$ $[+0.200, +0.574]$
All usable, plus model intercepts	834	$+0.395$ $[+0.208, +0.587]$
All usable, excluding paired run errors	798	$+0.338$ $[+0.153, +0.534]$
Earlier full-arm-screened population	440	$+0.541$ $[+0.254, +0.826]$
Earlier screen, excluding paired run errors	411	$+0.523$ $[+0.224, +0.817]$
Paired all-usable minus screened slope	—	-0.156 $[-0.306, -0.014]$

on 127. Leave-one-model or leave-one-stratum estimates ranged from $+0.459$ to $+0.701$. These sensitivities describe that screened population, not the all-usable population or new experiments.

For the current all-usable rows, the unadjusted within-model slopes are positive for seven models and negative for three. The positive pooled coefficient therefore does not mean every endpoint has a positive association. The saved fits and resamples are full-rank, with normalized design condition numbers from 14.63 to 19.13. No condition-number rule selects observations or changes the reported estimator.

G.5 Separating the two report contrasts

On the same 834 rows and design matrix, the full-minus-none slope is $+0.214$ $[+0.040, +0.414]$; none-minus-sham is $+0.170$ $[-0.012, +0.327]$. Using the same formula draws preserves the identity

$$\beta_{\text{full}-\text{sham}} = \beta_{\text{full}-\text{none}} + \beta_{\text{none}-\text{sham}}$$

at unrounded precision in every draw. It does not justify adding the marginal interval endpoints: the bootstrap correlation between the two components is -0.472.

On a common 790-row population without a run error in any of these three arms, the component slopes are $+0.203$ $[+0.018, +0.411]$ and $+0.138$ $[-0.056, +0.299]$, with bootstrap correlation -0.488. This excludes a different population; it is not an imputation or a correction for missing outcomes. The full-minus-none component supports an association beyond comparison with corrupted feedback, while still changing report presence and format.

G.6 Final successes, repair, and damage

Final exact successes on the 834 eligible rows are 305 with full, 248 with none, 254 with schema, and 200 with sham. The full count exceeds none by 57, and sham is below none by 48. These are descriptive mean-level summaries, not the decomposition of the overlap slope.

Table 17 separates changes from the shared first answer among non-error responses. Repair means a wrong first answer becomes correct; damage means a correct first answer becomes wrong. Recorded delivery errors are a separate column, not called semantic damage. Non-error parse failures retain the original unsuccessful score.

Table 17: Common-corpus transitions among non-error revisions. The four transition columns sum to "Rows"; Rows plus Errors equals 834 for each arm. These counts are descriptive, not additional tested contrasts.

Arm	Rows	Repair	Damage	Stay correct	Stay wrong	Errors
Full	820	110	13	195	502	14
Sham	804	43	43	157	561	30
Schema	806	50	1	204	551	28
None	807	41	1	207	558	27

Sham has 51 as-run damage events when failures of delivery are included; eight are run errors after a correct first answer, leaving the 43 non-error events shown. There is one non-error parser failure in sham and none in the other arms. Across the upper two prefixed overlap bins, sham has 22 run errors and full has seven. The paired run-error exclusion in Table 16 shows how the estimate changes when that eligibility condition changes.

The archive records formula-level overlap, but the rendered diagnosis depends on the model's first assignment. The cell records do not retain the actual disclosed slot lists, and the response cache does not retain request messages. Consequently the reported reanalysis does not adjust for actual slot confidence or reconstruct the exact historical prompt from a formula-wide proxy. The result is an observed formula-level association under the recorded revision protocol.

H Reader dependence in the separate revision study

H.1 A different population from the common corpus

A separate, prospectively specified study uses the same shared-first-answer, five-slot revision design but assigns substrates by model family. A disjoint 200-cell pilot selected exact solving as the primary outcome: clause fraction nearly saturates the GPT reasoning models on easier formulas, while exact solving floors Claude models on uf50. The resulting design assigns GPT-5.5 and GPT-5.6 to uf50, and three Opus endpoints to uf20. It plans 132 and 146 distinct formulas in the respective substrate sets, giving 702 model–formula cells; 698 are complete. It is not the ten-model, 90-formula common-corpus population.

The primary mean full-minus-partial-sham exact-solving gap is $+0.264$. Its slope on recorded overlap is $+0.620$ $[+0.413, +0.828]$; the saved bootstrap interval is $[+0.419, +0.821]$. The gaps by increasing overlap quartile are $+0.097$, $+0.260$, $+0.312$, and $+0.384$. The point estimate is positive on four of five models and individually significant on three.

The original quoted intervals and sign flips operate on pooled model–formula rows, and the standalone reader standard error also uses those rows. They are not new formula-clustered estimates. We retain their original scope here rather than pooling them with the current formula-clustered common-corpus intervals.

Against the shared first answer, full gains $+0.288$ in mean exact success and partial sham gains $+0.024$. Their overlap slopes are $+0.608$ $[+0.397, +0.819]$ and -0.012 $[-0.228, +0.203]$. Adding first-round performance or recorded log count as covariates gives full/sham slopes $+0.621$ and $+0.622$, respectively. These are sensitivities within the same study.

H.2 A simple marginal reader

The standalone reader receives the model's parsed first assignment and the same five reported marginals. It sets each flagged variable to $\mathbf{1}[q > 0.5]$ and leaves the remaining variables unchanged. This is a simple component-wise update, not a complete SAT solver or a maximizer of exact-solving probability under the formula.

Its mean full-minus-sham gap is $+0.208$, compared with the model's $+0.264$. The difference in these mean gaps is $+0.056$ $[+0.011, +0.101]$. The reader's own overlap slope is $+0.163$ $[-0.007, +0.334]$; the model-minus-reader slope is $+0.457$ $[+0.206, +0.708]$. This comparison shows what the revision loop adds beyond that update rule, not what it adds beyond all non-model computation.

H.3 Adding greedy repair changes the mean comparison

The reader ladder first applies the marginal update, then uses B steps of greedy local search, flipping the variable that most increases the number of satisfied clauses. Its zero-step implementation differs from the standalone reader above because it reconstructs the commitment within the ladder's code path. We retain both stored results and do not treat their 0.010 mean-gap difference as a new effect.

At one greedy flip, the mean comparison reverses: the reader's full/sham gap is larger than the model's. At larger budgets, search can solve more instances in both report arms, reducing the

Table 18: Saved reader-strength ladder in the separate five-model study. Mean-gap differences and overlap-slope differences are distinct quantities. The quoted intervals retain the original pooled-row analysis.

Greedy flips B	Reader mean gap	Model − reader mean gap	Model − reader slope
0	$+0.218$	$+0.046$ $[+0.002, +0.089]$	$+0.476$ $[+0.231, +0.721]$
1	$+0.350$	-0.086 $[-0.133, -0.039]$	$+0.595$ $[+0.332, +0.858]$
2	$+0.251$	$+0.013$ $[-0.038, +0.064]$	$+0.854$ $[+0.574, +1.134]$
5–100	$+0.107$	$+0.156$ $[+0.111, +0.202]$	$+0.587$ $[+0.334, +0.841]$

reader's message gap again. Improved absolute solver performance need not produce a monotone difference between its arms. The model-minus-reader overlap interaction remains positive within this tested ladder, while the mean ordering is reader-dependent.

H.4 A COMPLETE SOLVER ANSWERS A DIFFERENT RESOURCE QUESTION

On the uf20 stratum, a DPLL solver with unit propagation and no report solves 0.371 of the instances within eight branching decisions and 1.000 within 64. The corresponding loop success is 0.459. Branching decisions, greedy flips, and model calls are different resource units. This result therefore does not supply a matched speed comparison, but it makes clear that the model result is not superiority over competent complete SAT solving.

The lesson relevant to the main paper is positive and specific: a report's effect should be evaluated for a stated consumer. A simple marginal update, a local repair rule, and a complete solver can use the same task information very differently. Their mean feedback gaps and overlap interactions answer different questions.

I COMPACT SCOPE CHECKS

I.1 A PARTIALLY COMPLETED BOOLEAN-CIRCUIT CONSTRUCTION

A second construction asks models to synthesize five-gate Boolean circuits matching a target truth table. The declared accepted circuits all compute that table, and the report describes recurring gate components using the same five-arm comparison. This changes the objects and verifier relative to the term selection study.

A compute cap stopped collection after 20% of the planned grid. The retained balanced prefix contains eight complete families, 11 models, six levels, five arms, and one seed: $2{,}640$ cells. Partial families beyond that prefix are excluded. Its overlap values span 0.057 to 0.829. This is an observed incomplete study, not a completed replication.

Table 19: Directional results in the Boolean prefix. Intervals use its eight family units. The two-sided sign-flip floor is 0.0078. These exploratory prefix estimates do not describe an unobserved completion of the grid.

Contrast and outcome	Interaction [95% interval]
Model full − partial sham, graded	$+0.653$ $[+0.555, +0.762]$
Model full − partial sham, exact	$+1.009$ $[+0.845, +1.187]$
Model − specified reader, graded	$+0.035$ $[-0.008, +0.090]$
Model − specified reader, exact	-0.321 $[-0.432, -0.194]$

The full/sham point estimates are positive on all eleven models. Parsing and strict output conformance also vary with overlap. Restricting to three models that conform on every cell gives graded and exact interactions $+0.542$ and $+0.895$, with model-minus-reader differences -0.076 and -0.435. Equal-count levels give a graded model interaction $+0.521$ and a reader-relative difference -0.031. These sensitivities support the observed direction while preserving the partial-coverage qualification.

The stored check recomputed all $2{,}640$ scores from raw responses, with zero parse, score, or frozen-world hash disagreements. It also rechecked all 48 overlap values and 320 accepted circuits. These checks verify the reported prefix; they do not increase its eight-family sample size.

I.2 A COMPLETED PAIR-CREDIT PROBE

A separate probe changes the credit and report to involve component pairs. Its complete grid contains 24 families, six levels, five arms, three models, and two seeds: $4{,}320$ cells. No commitment is malformed, and raw-response rescoring has no recorded mismatches.

The graded correctness-by-overlap interaction is $+0.005$ $[-0.125, +0.137]$ ($p = 0.94$). The exact interaction is $+0.075$ ($p = 0.063$), with only eleven exact-recovery events in the entire grid and nineteen of the twenty-four families contributing none. The declared conjunction of outcomes is not supported.

Three measured properties limit what this null says about non-additive credit. First, the report names five pairs, whereas a five-part answer has ten pairs, and those pairs are not directly emitted as an answer. The report/action relation has changed along with the utility. Second, realized pair overlap L_2 spans only 0.031–0.211 on the common normalization. The stored common-normalization interaction is $+0.017$ with an approximate interval $[-0.40, +0.44]$, which does not exclude the main additive estimate.

Third, the consumer's problem becomes much closer to additive after seeing the actual message. Before conditioning, an additive top-k rule is optimal on 12.5% of constructed instances, with a median utility shortfall of 47.7%. After conditioning on the report, it is optimal on 42.4%, with median shortfall 3.2%. The nominally harder utility therefore does not ensure a harder reported decision.

This result is relevant scope evidence. It neither confirms the main interaction under pair credit nor demonstrates a general boundary at non-decomposability. The more general evaluation point is to inspect the decision after conditioning on the information the consumer actually receives.

J REPRODUCTION SCOPE AND CLAIM TRACEABILITY

J.1 WHAT AN OFFLINE REPRODUCTION CHECKS

The score-and-statistics package verifies content hashes before reading inputs and rejects missing or changed files. Its core replay uses canonical commitments, frozen accepted sets and scorers, saved reader arrays, and the common-corpus outcome table. It can check scores and statistics without making model calls or running generated programs.

For the primary construction, replay rescans the committed terms under the frozen score contract, reconstructs the fixed-count subset, and groups all model/seed observations inside each family before computing the contrast. The raw-response parser check has complete coverage of the $3{,}168$ primary cells, not every raw prompt and response in all five arms and six levels. The larger $15{,}840$-cell archive provides canonical commitments and scores.

For the expected-overlap reader, the replay uses frozen policy arrays and the recorded calibration/inference conventions. It reproduces the interaction difference and its precision summaries. Full offline enumeration is documented as the original reference calculation; using its saved arrays does not claim to rerun that computation.

For the common corpus, replay reproduces the point fits, checks the saved joint-resample intervals, and checks the component identity within every shared draw. It preserves the formula cluster, initial-answer eligibility, original outcome scores, and run-error flags. It does not reconstruct missing request messages or actual-slot confidence fields.

J.2 PERSISTENT SCIENTIFIC SOURCES

Table 20 names the scientific records associated with the central claims. These are artifact filenames. The construction and SAT studies have separate protocol, data, and analysis records; a common label such as "full" does not make their reports identical.

Table 20: Source map for the retained main claims. Filenames identify saved records. The text specifies the relevant populations so that a prefix or screened analysis cannot silently replace the reported estimate.

Claim	Saved record family	Population/section
Controlled interaction and absolute means	`analysis_full_verify.json`; `SEED_MULTIPLICITY.json`; frozen family manifest	24 families; Section 4
Expected-overlap reader	`BAYES_OPTIMAL_READER_RESULT.json`; `EQUIVALENCE_RESULT.json`	Same families; Section 5
Fixed-report SAT	`TRACK_M_RESULT.json`; `TRACK_M_COUNTCONTROL_RESULT.json`; SATLIB formula records	1,000 formulas; Appendix F
Common-corpus revisions	Retained outcome table, fixed-fit summaries, and stored joint formula resamples	834 rows, 90 formulas; Appendix G
Separate revision/reader study	`TRACK_N_RESULT.json`; standalone-reader and greedy-ladder summaries	698 complete rows; Appendix H

The term generator, report renderer, model schedule, and reader objective are described by the frozen modules `fiber_env.py`, `fiber_report.py`, `fiber_pilot.py`, and `bayes_optimal_reader.py`. They are named to make the algorithmic contract traceable, not to promise that an arbitrary current environment can regenerate the original service responses. The SAT records likewise distinguish exact uf20 enumeration from the bounded uf50 solution collection.

J.3 ACCESS, ANONYMITY, AND LIMITS OF REPLAY

The anonymous repository provides code and packaged replay records. Main-paper replay covers saved commitments and scores, frozen reader arrays, fixed-report SAT responses, and common-corpus outcomes and resamples, with the coverage limits above. These checks do not regenerate service responses or rerun the original full reader enumeration. The separate large numerical archives for other completed scientific studies are not part of the lightweight source package and are not required for this paper's statistical replay. SATLIB clause files are obtained separately; the default statistical replay does not require them. Endpoint labels describe the experimental configuration, and requested/served records support response verification. Offline score reproduction does not require continued access to those endpoints. New generations would be a different reproducibility exercise, with their own stochastic, service-version, and delivery variation.

These replay checks cover the reported claims. Separate scientific and programming studies are not counted as additional confirmations of the controlled overlap result.