LAYER-EDIT COVERAGE BEYOND SAMPLING: FROM CANDIDATES TO SELECTED ANSWERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Can layer-execution edits supply correct answers that ordinary sampling misses, and can a selector recover them? We study skipping, repetition, and substitutions that combine the two in frozen language models. Our grid contains approximately 14.2 million correctness evaluations over 1,261 paths, five mathematical difficulty strata, and four prompt conditions on Qwen2.5-3B-Instruct, with a Qwen3-4B comparison. Across three model–difficulty groups of 565 questions each, the exhaustive 1,260-edit pool covers 15–56 questions missed by all 57 unedited-model samples. Yet question-held-out path ranking and direct 57-path voting correctly answer only 0–3 of these additional questions and trail 57-sample self-consistency by 23–57 correct answers per group. Shortlists exclude many available answers, and voting rarely recovers those retained. The edit grid also has reproducible structure: average substitution performance tracks constituent effects, while 45–51% of discovered substitution events with failed constituents succeed again on the rewrite prompt. Tested pre-generation rankers show no improvement over a structural prior, despite informative post-generation correctness signals. These findings separate additional candidate coverage from the ability to select a correct answer. They use original-checker scores and previously inspected questions and outputs; unequal coverage pools and unmeasured end-to-end costs preclude an efficiency claim.

1 INTRODUCTION

Changing which layers a language model executes can change its answers. The practical question is whether these edits supply correct answers that ordinary sampling misses, and whether a system can select them. We address both parts with an exhaustive grid of layer edits and question-held-out selection experiments. The grid contains additional correct candidates, but the tested selectors rarely return those answers and underperform self-consistency. This gap makes candidate generation and answer selection separate empirical questions for layer-edit inference.

We measure *coverage*, the presence of any correct candidate, and *output accuracy*, the correctness of the answer actually returned. An oracle establishes coverage by inspecting correctness labels. A selector must choose without the target question's labels and must account for damage to questions the unedited model already solves. We compare its output with self-consistency (SC), which votes over ordinary samples (Wang et al., 2023). We distinguish candidate counts from end-to-end computation.

Our grid pairs layer skipping and repetition with substitutions that skip one layer and repeat another, preserving the number of layer executions per token. This construction lets us examine constituent effects alongside coverage and selection. The principal findings are:

- **Additional coverage.** The exhaustive edit pool covers 15–56 questions per group that all 57 unedited-model samples miss (§3.1).
- **Limited selection.** Direct structural voting correctly answers only 0–3 of these additionally covered questions per group (§3.2).
- **Replicable composition exceptions.** Of the substitution events discovered to succeed while both constituent edits fail every discovery prompt, 45–51% succeed again on rewrite (§4.2).

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.

Layer removal and adaptive computation can support efficient inference (Fan et al., 2020; Men et al., 2024; Gromov et al., 2024; Elbayad et al., 2020; Schuster et al., 2022; Raposo et al., 2024). Layer reuse also changes reasoning behavior (Dehghani et al., 2019; Saunshi et al., 2025; Li et al., 2025; Heakl et al., 2026): recent studies report gains from training-free block recurrence (Chen et al., 2026) and learned selective iteration (Fu et al., 2026). Our study complements these approaches by measuring coverage and selected answers for a finite edit pool at fixed checkpoints; §7 distinguishes their operators and decision rules.

The coverage and output comparisons use three model–difficulty groups of 565 questions each. Composition and prediction analyses then characterize the candidate pool and tested signals (§4.2–5). We use the original checker's scores and treat the analyses that reuse inspected questions as exploratory. §2 specifies the measurement and selection protocols; §6 collects their limitations. Verification scripts and result-to-ledger mappings support reproduction (Appendix A).

2 Setup and Measurement Protocol

Layer edits and evaluation data. For the default L-layer schedule $\pi_0 = (1, \ldots, L)$, KEEP executes the unedited model, SKIP(i) removes layer i, and REPEAT(j) executes layer j twice. A substitution SUB$(i \leftarrow j)$ skips i and repeats j adjacent to its original occurrence, preserving the order of other layers. The four operators execute L, $L-1$, $L+1$, and L layers per token, respectively. Qwen2.5-3B-Instruct (Qwen Team, 2024) has $L = 36$. Excluding two deletion boundary positions and 34 diagonal substitutions leaves 34 SKIP, 36 REP, and 1,190 SUB edits, plus KEEP: 1,261 paths. The SUB-only pool contains the 1,190 substitutions; oracle references may also retain KEEP.

Five disjoint DART-Math (Tong et al., 2024) difficulty strata contain $N = 565$ questions each, taking the first retained questions up to the common minimum across available grids. Difficulty is defined by a reference model's failure rate. Every path is evaluated under 0-shot, 1-shot, 2-shot, and rewrite prompts, giving approximately 14.2M binary outcomes. Rewrite changes one fixed instruction wrapper while keeping the question body. The model comparison uses Qwen3-4B (Qwen Team, 2025) on difficulty 5. Stored correctness matrices support coverage and single-path evaluation; we grade sampled and voted answers with the original equivalence checker. These scores share a checker convention but are not a fresh, uniform regrading of all candidates. Scoring and provenance limitations are in §6.

Discovery, evaluation, and coverage. Discovery validity D requires correctness on all three 0/1/2-shot prompts. Evaluation correctness C scores rewrite; four-condition validity is $R = D \wedge C$. Average correctness averages the four observations. An evaluation-only oracle counts a question as covered when KEEP or any candidate edit is C-correct. A discovery-constrained oracle also requires the recovering edit to be D-valid. Both preserve correct KEEP answers and use evaluation labels to establish existence. Prompt transfer evaluates a discovery-selected choice on C; question-held-out fitting separates training and evaluation questions. These are distinct splits. A decision that reads the target question's discovery labels is a *conditional-transfer diagnostic*. Our primary selectors use supervised training but no target correctness labels at decision time.

Training and ranking paths. The primary selectors use five question-grouped folds, keeping each question's three discovery-prompt feature views together. Qwen2.5 difficulty-3 fits pool training strata 2+3; difficulty-5 fits pool 4+5. Qwen3 fits use difficulty 5. Within each training fold, we standardize layer-18 user-body mean prefill features, reduce them to 256 principal components, and fit 71 independent L2 logistic probes: one for each SKIP, REP, and KEEP action. Each action's D-validity label is repeated across its three feature views. We apply the fitted pipeline to held-out rewrite features.

SKIP and REP receive their predicted log-odds, SUB receives the sum of its constituent scores, and KEEP receives its own score. This factorization uses constituent predictions to rank substitutions; its relation to measured composition is examined in §4.2. Held-out correctness never enters fitting, eligibility, or ranking. The multilayer ranker in §5 is a separate prediction diagnostic.

Returning an answer. Table 1 summarizes the primary comparisons. Single-path selection executes the highest-ranked path, including KEEP when it wins. Direct structural voting uses exactly the top 57 paths, with one ballot per path and KEEP included only if shortlisted. It strips outer whitespace, excludes empty strings, and returns the exact-string plurality winner without semantic clustering. Ties follow the first supporting rank; an all-empty shortlist returns an empty answer, scored incorrect. The fixed-path reference maximizes D-validity on training-fold questions. SC samples KEEP at temperature 0.7, top-$p = 1$, with a 2,048-token output cap and returns the most frequent extracted answer. The width 57 is inherited from earlier cover experiments, not selected as an optimum for these fits.

Table 1: Primary selectors use training labels or generated answers, never target correctness.

Method	Decision	Generated answers
Single-path selector	Execute the highest-ranked path, with KEEP eligible.	1
Direct structural vote	Return the plurality answer from the top 57 paths.	57
Training-selected fixed path	Execute the best path on training-fold discovery labels.	1
KEEP self-consistency (SC)	Return the plurality answer from unedited-model samples.	57

Notes. Counts exclude feature acquisition. Equal counts do not imply equal end-to-end cost.

Output accuracy and uncertainty. A recovery corrects a KEEP error; a destruction replaces a correct KEEP answer with an incorrect one. We score the returned answer:

$$\text{Net} = \text{recovered} - \text{destroyed}, \qquad \Delta\text{Acc} = \text{Net}/565. \tag{1}$$

Post-hoc refund instead restores correct KEEP answers after consulting correctness. For example, 60 recoveries and 57 destructions yield net 3, not 60. Refund-normalized diagnostics remain separate (Appendix A). For the primary output comparisons, 95% intervals use 10,000 question-bootstrap resamples with seed 20260915, conditional on fitted decisions and stored outputs. Same-question differences are paired; disjoint strata are resampled independently. Other diagnostics state their intervals separately. Coverage, output accuracy, and prediction are reported in §3.1, §3.2, and §5, respectively. Full fitting details, historical protocols, and resampling conventions are retained in companion Appendix B.1 and Appendices B and E.3.

3 From Candidate Coverage to Output Accuracy

3.1 Layer edits cover questions missed by ordinary sampling

We compare the exhaustive edit pool with 57 stored SC responses for each question in three model–difficulty groups: 565 questions per group and 96,615 sampled responses in total. Additional edit coverage counts questions missed by every SC sample but answered correctly by at least one edit. The full pool adds 15, 42, and 56 such questions (Figure 1a). The SUB-only pool gives the same increments, and all these questions are also KEEP errors. This is coverage beyond SC's candidate pool, rather than its voted answer; the 1,260-edit and 57-sample pools have unequal sizes.

The distinction between coverage and selection also matters within SC. At Qwen2.5 difficulty 5, its 197 vote errors comprise 137 questions with a correct SC sample, 42 with no correct SC sample but a correct edit, and 18 covered by neither pool. Edits therefore add candidates for some failures, while many others already have a correct ordinary sample. Treating SC checker exceptions as potentially correct lowers the three additional-coverage counts to 13, 36, and 45. These are sensitivity bounds, not confidence limits. Excluding known empty but matrix-correct structural cells leaves the primary counts unchanged. Complete counts and intervals are in Appendix G.2.

3.2 Direct voting realizes little of the additional coverage

We next evaluate the selectors defined in §2. The single-path selector retains KEEP as an eligible action; the direct voter returns its winning string. Table 2 reports both recoveries and destruction relative to KEEP.

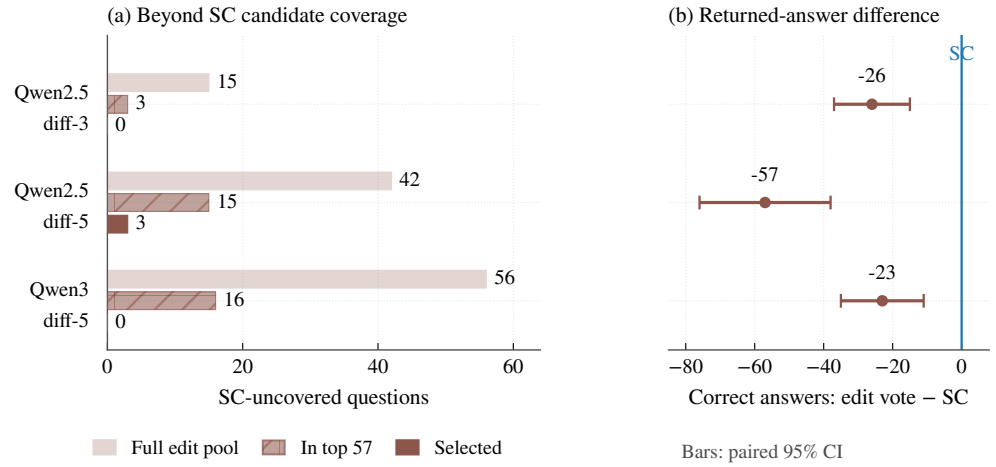

Figure 1: Additional edit coverage is rarely selected, and direct voting trails SC. (a) SC-uncovered questions covered by the full edit pool, retaining a nonempty matrix-correct answer in the top 57, and answered correctly by the voter. (b) Direct-vote minus SC correct counts, with paired 95% question-bootstrap intervals; $N = 565$ per group. Panel (a) uses unequal candidate pools and different scoring records; neither panel matches end-to-end costs.

Table 2: Direct structural voting trails self-consistency in all three groups.

Model / stratum	Method	rec/dst	Net	95% interval
Qwen2.5 / diff-3	Single-path selector	27/103	−76	−97 to −55
	Direct 57-path vote	46/14	+32	17 to 47
	SC@57	61/3	+58	43 to 73
Qwen2.5 / diff-5	Single-path selector	32/181	−149	−175 to −123
	Direct 57-path vote	60/57	+3	−18 to 24
	SC@57	78/18	+60	42 to 79
Qwen3 / diff-5	Single-path selector	18/74	−56	−74 to −38
	Direct 57-path vote	26/31	−5	−20 to 10
	SC@57	29/11	+18	6 to 30

Notes. $N = 565$; KEEP correct counts are 461/308/352. Net and intervals are in questions. Single paths use matrix scores; votes use the original answer checker. Intervals condition on fitted decisions and outputs.

Single-path selection reduces accuracy in every group; training-selected fixed paths instead net $0, +1, +2$. Direct voting improves on KEEP at Qwen2.5 difficulty 3, while its other two intervals span zero. It trails SC@57 by 26, 57, and 23 correct answers; the paired intervals exclude zero (Figure 1b).

The additionally covered questions show where answers are lost. The top-57 lists retain a nonempty, matrix-correct answer for only 3, 15, and 16 of the 15, 42, and 56 questions. The final voters correctly answer only 0, 3, and 0. Thus limited recovery involves both shortlisting and final selection. Availability uses stored C-matrix labels, while final answers are scored by the original checker; these counts compare distinct records and are not a uniformly regraded or causal loss decomposition.

Candidate availability and voting accuracy can disagree. Across all Qwen3 difficulty-5 questions, structural shortlist availability slightly exceeds SC coverage (456 versus 451), yet the structural voter returns fewer correct answers (347 versus 370). A larger set of covered questions therefore need not yield a more accurate voter. Checks of KEEP insertion, return mapping, and solver convergence do not remove the observed deficit (companion Appendix B.3). Historical forced-KEEP voting and target-label-assisted classification use different protocols; their width and supervision controls are retained in Appendices B and G.3. The next section examines the structure of the edit pool independently of these selection outcomes.

4 Structure and Transfer of Layer-Edit Success

4.1 Discovery validity narrows evaluation-prompt coverage

The preceding comparison asks what edits add beyond sampling. We now examine how edit success transfers across prompts and relates to its constituent operations. At Qwen2.5 difficulty 5, KEEP answers 308 of 565 questions correctly. Searching the full path pool with evaluation labels recovers another 238, for oracle accuracy $546/565$. Requiring the recovering path to succeed on all three discovery prompts yields 125 recoveries and accuracy $433/565$ (Table 3). Most of this constrained coverage remains in the SUB-only pool: 121 recoveries, compared with 237 under C-only scoring. Thus the layer-count-preserving edits themselves contain substantial coverage.

Table 3: Requiring discovery validity reduces oracle recovery.

Stratum	KEEP	Full: eval-only	Full: disc.+eval	SUB: disc.+eval	Reduction
diff-1	534	30	25	25	5
diff-3	461	102	71	70	31
diff-5	308	238	125	121	113

Notes. $N = 565$. Recovery columns count additional correct questions beyond KEEP. Disc.+eval requires D and C; full and SUB pools contain 1,260 and 1,190 edits. Both oracles retain correct KEEP answers. Reduction is the loss in full-pool recoveries.

The corresponding full-pool counts at difficulty 3 are 102 and 71. On Qwen3-4B difficulty 5, they are 155 and 85 against a KEEP baseline of $352/565$. The 113-question reduction at Qwen2.5 difficulty 5 measures the effect of requiring discovery correctness, including the resulting restriction of admissible candidates. Since both references still choose with evaluation labels, this contrast quantifies coverage rather than performance of a decision frozen before evaluation.

4.2 Constituent regularity and successful exceptions

An additive prediction formed from measured SKIP and REP success rates correlates 0.91–0.95 with the question-averaged substitution grid (Figure 6). At the question level, constituent success is informative but insufficient. Let TT denote that both constituents are D-valid, succeeding on all three discovery prompts. Table 4(a) reports the probability that SUB is R-valid, succeeding on all four prompts, conditional on TT and across the full SUB pool. At difficulty 5, TT raises this probability from 10.7% to 57.7%, leaving substantial failures. Discovery-valid constituents therefore provide a useful but incomplete predictor of four-prompt substitution validity. The TT-selected question–path population changes with stratum, so this contrast does not identify a causal effect of difficulty (Appendix D.2).

Table 4: Constituent success predicts substitution validity, but failed constituents can yield successful substitutions.

(a) Successful constituents: conditional versus base success

Stratum	All substitutions	Both constituents D-valid	Ratio
diff-1	56.5%	81.2%	1.44×
diff-3	29.6%	67.9%	2.29×
diff-5	10.7%	57.7%	5.39×

(b) Failed constituents: discovery and rewrite replication

Stratum	Replicated/discovered	Replication	Matched FF	Ratio
diff-1	50/99	50.5%	5.1%	9.93×
diff-3	134/292	45.9%	4.1%	11.16×
diff-5	197/438	45.0%	2.9%	15.55×

Notes. (a) R-validity among 1,190 SUB paths. (b) Question–path events with both constituents failing every discovery prompt; discovered substitutions succeed on all three. Matched FF rates include discoveries; ratios use unrounded rates.

Success after both constituents fail. Some substitutions succeed on all discovery prompts even when their SKIP and REP constituents fail on every one of those prompts (strict FF→✓). Across the three strata, 45–51% of discovered events also succeed on rewrite (Table 4(b)). At difficulty 5, 197 of 438 discoveries replicate, compared with a 2.9% rewrite success rate among all strict-FF opportunities. The resulting $15.55\times$ enrichment compares discovered events with that matched population, which includes the discoveries. Exact denominators and the distinct discovery-conversion rate are given in Appendix D.1. These are replicable behavioral exceptions. Multiple events can share a question; the retained intervals resample events rather than question clusters. Figure 5 and Appendix D.1 give the distributions and resampling details.

These findings establish behavioral structure in the measured edit grid. The associated residual analysis measures a mixture of question-specific margins, interaction, and noise; it does not isolate a cause of selection difficulty. Its full estimator, response-grid sensitivity, and behavioral-consistency criterion are retained in Appendix D. We next examine prediction signals for baseline correctness and path selection.

5 Prediction Signals for Detection and Selection

Answer agreement predicts baseline correctness. Table 5A compares four post-generation signals for 0-shot KEEP correctness. Agreement of five structural probes with KEEP has the highest measured AUROC. Each arm uses its own available records, so the values are not uniformly matched-population comparisons. They measure whether a signal can identify a correct baseline answer, rather than choose a replacement.

Table 5: Correctness detection and prefill path ranking measure different prediction tasks.

A. Post-generation detection: 0-shot KEEP correctness

Signal	diff-3	diff-5
Structural-probe agreement ($K = 5$)	0.972	0.929
Prompt-condition agreement	0.911	0.886
Temperature agreement ($n = 5$)	0.433	0.603
Answer log-probability	0.471	0.527

B. Prefill ranking: R-validity at difficulty 5

Predictor / training population	AUROC	Increment	Paired 95% CI
Cross-fitted structural prior	0.6765	—	—
Ranker, all questions	0.6653	−0.0112	$[-0.0352, 0.0135]$
Ranker, baseline errors only	0.6279	−0.0486	$[-0.0800, -0.0183]$

Notes. B reports mean within-question AUROC and paired increments over the same prior. Full signal definitions and detection intervals: Appendix G.4.

An inspected detector workpoint improves on KEEP but exceeds the specified damage cap at difficulty 5. A separate cross-condition gate loses accuracy. These outcomes illustrate why detecting errors is insufficient: the system must also select a useful replacement and avoid destroying correct answers. Probe and threshold inspection on the same questions makes these workpoints exploratory. Their exact protocols, uncertainty, and costs are in Appendices G.4 and B.2. Five-sample agreement here is distinct from the 57-sample answer vote in §3.

Prefill rankers show no improvement over a structural prior. Table 5B evaluates a separate question-conditioned ranker, which scores factorized path embeddings against an attention-pooled prefill representation from all 36 layers. Unlike the layer-18 marginal probes used for output selection, it is evaluated on within-question R-validity ranking. Both training regimes use the same 125 held-out difficulty-5 evaluation questions: KEEP is incorrect on rewrite, and the 1,261-path pool contains both R-valid and R-invalid paths, including KEEP. "All questions" and "baseline errors only" refer to training populations. The structural prior assigns each path its validity rate over all training-fold questions, using no features of the evaluation question.

The all-question ranker's increment over that prior has an interval spanning zero; the error-only increment is negative. Neither tested regime demonstrates an improvement. Alternative pooling, multilayer features, nonlinear heads, and an external embedding are documented in Appendix C.1.

Ranking quality must be evaluated through an output decision. AUROC averages pairwise comparisons. Output gain instead depends on the chosen answer, KEEP decisions, and destruction on baseline-correct questions. An oracle-assisted latent-score simulation illustrates this distinction under refund and no-refund accounting (Appendix C.4); its thresholds belong to the specified score model and candidate pool. The observed output results in §3 provide the direct test of whether a selection rule realizes useful coverage.

6 Limitations

Exploratory evidence and uncertainty. The primary selectors exclude held-out correctness from fitting and decisions, but reuse previously inspected questions and stored answers. They are exploratory analyses, not prospective confirmation. Earlier conditional-transfer controls additionally use target discovery labels. Question-bootstrap intervals condition on fitted decisions, saved outputs, and checker outcomes; they exclude repeated fitting, generation, and model-selection uncertainty. Intervals crossing zero do not establish equivalence or absence of useful gain. The question sets are retained prefixes of available grids, not random population samples.

Scoring and provenance. The original checker can err without raising an exception: it accepts 2^{4950} against reference 2^{5050} in one audited Qwen2.5 question. We retain its scoring convention and disclose exception and missing-answer sensitivity separately (companion Appendix B.4). The structural matrices have not all been independently regraded. Qwen2.5 voting replay also depends on historical question/path alignment: agreement with stored references and feature ordering does not independently verify generation-time provenance. These findings concern recorded checker outcomes rather than independently verified mathematical correctness.

Compute and method coverage. The 1,260-edit oracle is not budget-matched to 57 SC samples. Even the 57-candidate voting comparison lacks complete end-to-end costs, including generation and feature extraction. Equal candidate or per-token layer counts do not establish equal cost. Our selectors do not reproduce stronger learned routers or shared-pool answer verifiers; their failures do not establish that other selection rules cannot exploit edit coverage.

Generalization and mechanism. The primary evidence concerns mathematics, two related Qwen checkpoints, and one rewrite wrapper. MMLU supplies descriptive evidence rather than a deployment replication (Appendix E.4). Independent tasks, model families, and new-question confirmation remain open. Four fixed templates do not identify generation variance. FF intervals resample events rather than question clusters. Constituent failures do not identify a computational mechanism, and the residual estimator mixes question-specific margins, interaction, and noise. Neither analysis identifies the cause of the observed selection gap.

7 Related Work

Efficient inference and routing. Layer removal and adaptive computation change how much of a model executes (Fan et al., 2020; Men et al., 2024; Gromov et al., 2024; Elbayad et al., 2020; Schuster et al., 2022; Raposo et al., 2024). Routing also changes which computation an input receives: CoLa (Li et al., 2025) searches paths with label-guided MCTS; Dr.LLM (Heakl et al., 2026) and PoLar (Li et al., 2026) learn decisions from offline search traces, using intermediate hidden states and an external question encoder, respectively. Task-level routing (Batorski et al., 2026) changes the unit of adaptation. These approaches motivate our question about selecting useful edits. Our finite grid instead provides paired constituent outcomes and separates candidate coverage from the outputs of specified selectors; it does not reproduce the broader action spaces and learned routers of Dr.LLM or PoLar.

Looping and adaptive depth. Reusing computation ranges from ordinary layer repetition to altered recurrence operators and adapted models. Training-free looping (Chen et al., 2026) uses damped block updates, while LoopUS (Park et al., 2026) adapts models with gated recurrence and depth supervision. TaH (Fu et al., 2026) learns token-level iteration; RecurTrace (Wang et al., 2026) adds loop-time memory and learned halting. Both use oracle supervision during training, not target labels at inference. LoopFormer (Jeddi et al., 2026) trains for variable compute budgets, and T-LoopFormer (Yu et al., 2026) routes token-level depth from hidden states. Our ordinary skip/repeat substitutions isolate a smaller intervention at frozen checkpoints. Their results characterize this edit family rather than the performance of adaptive looping as a whole.

Candidate coverage and answer selection. Test-time scaling evaluation already distinguishes candidate coverage from system accuracy (Hariri et al., 2026). PoLar's single-output accuracy and pass@5 similarly describe different outcomes. CAPS (Lin et al., 2026) studies adaptive pairwise verification on shared candidate pools. We apply the coverage–selection distinction to layer edits and track both shortlist availability and returned answers. This comparison identifies what the tested ranking and plurality rules realize; shared-pool verification remains a separate, untested way to select from these candidates.

Self-consistency. SC aggregates ordinary sampled answers (Wang et al., 2023) and provides a direct multi-sample reference for structural voting. We separately measure questions missed by every SC sample and errors made by its vote despite a correct sample. This distinguishes the role of an additional candidate source from the role of a better answer selector. Our coverage comparison uses unequal pool sizes, while the output comparison fixes candidate count; neither establishes equal end-to-end cost. Appendix F gives detailed protocol comparisons and Tables 17–18 distinguish evaluated methods from open comparisons.

8 Conclusion

In the evaluated Qwen math grids, layer edits cover questions missed by ordinary sampling, but the tested selectors rarely recover those answers and remain below self-consistency. Both shortlisting and final voting leave additional coverage unrealized. The grid also contains substitutions that succeed repeatedly despite failed constituents, while the tested prefill rankers show no improvement over a structural prior.

These findings motivate evaluating a candidate source together with the answer selection rule that uses it. A useful extension must translate coverage into returned-answer accuracy, accounting for damage to correct baseline answers. Shared candidate pools, independent questions, verified grading, and measured generation and selection costs would make that test more conclusive.

AI USE STATEMENT

Large language models were used as writing and review aids, for arithmetic and cross-section consistency checks, and for drafting analysis and verification code. The authors remain responsible for experimental design, estimators, registration history, verification, and all reported claims. AI-assisted revisions require author review. Exploratory results, failed criteria and unresolved reproduction limitations are disclosed in the text and appendices.

REPRODUCIBILITY STATEMENT

Data and generation settings are documented in Appendix E; feature construction and routing protocols are in Appendices C.1 and B. The retained analysis archive includes correctness matrices, prefill features, source snapshots, frozen decisions, and answer caches. For the new offline experiments, saved fold assignments, preprocessing/probe parameters, and candidate/answer ledgers support conditional replay of fitted predictions and vote winners (companion Appendix B.5). Artifact availability and the limits of rescoring, refitting, and regeneration are described in Appendix E.2 and companion Appendix B.5; checker and alignment checks are in companion Appendix B.4. The detailed companion supplies verification entrypoints, the result-to-ledger mapping, and a minimal ledger-reduction example (companion Appendix B.6). The example requires the retained archive; its expected counts and accounting identities are given there. The companion also provides extended analyses (Appendix C), the English registration and deviation record (Appendix I), and correction history (Appendix K). A public release of the retained code, tensors, and analysis artifacts with file hashes is planned.

REFERENCES

Paweł Batorski, Abtin Pourhadi, Akylgali Aitaza, Przemysław Spurek, and Paul Swoboda. MACRO: Markov chain routing of transformer layers. *arXiv preprint arXiv:2608.05872*, 2026. URL https://arxiv.org/abs/2608.05872.

Hugh Blayney, Álvaro Arroyo, Johan Obando-Ceron, Pablo Samuel Castro, Aaron Courville, Michael M. Bronstein, and Xiaowen Dong. A mechanistic analysis of looped reasoning language models. *arXiv preprint arXiv:2604.11791*, 2026. URL https://arxiv.org/abs/2604.11791.

Lizhang Chen, Jonathan Li, Chen Liang, Ni Lao, and Qiang Liu. Training-free looped transformers. *arXiv preprint arXiv:2605.23872*, 2026. URL https://arxiv.org/pdf/2605.23872.

Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Lukasz Kaiser. Universal transformers. In *International Conference on Learning Representations (ICLR)*, 2019. URL https://arxiv.org/abs/1807.03819.

Maha Elbayad, Jiatao Gu, Edouard Grave, and Michael Auli. Depth-adaptive transformer. In *International Conference on Learning Representations (ICLR)*, 2020. URL https://arxiv.org/abs/1910.10073.

Angela Fan, Edouard Grave, and Armand Joulin. Reducing transformer depth on demand with structured dropout. In *International Conference on Learning Representations (ICLR)*, 2020. URL https://arxiv.org/abs/1909.11556.

Tianyu Fu, Yichen You, Zekai Chen, Guohao Dai, Huazhong Yang, and Yu Wang. Think-at-Hard: Dynamic looped transformers for improved reasoning. In *International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2511.08577v4.

Andrey Gromov, Kushal Tirumala, Hassan Shapourian, Paolo Glorioso, and Daniel A. Roberts. The unreasonable ineffectiveness of the deeper layers. *arXiv preprint arXiv:2403.17887*, 2024. URL https://arxiv.org/abs/2403.17887.

Mohsen Hariri, Weicong Chen, Nahal Shahini, Vikash Singh, Kai Ye, Amirhossein Samandar, Debargha Ganguly, Sreehari Sankar, Yanyan Zhang, Shouren Wang, Jerry Peng, Biyao Zhang,

Michael Hinczewski, and Vipin Chaudhary. Test-time scaling in reasoning LLMs: Inference regimes, evaluation, and reproducibility. *arXiv preprint arXiv:2608.04001*, 2026. URL https://arxiv.org/abs/2608.04001.

Ahmed Heakl, Martin Gubri, Salman Khan, Sangdoo Yun, and Seong Joon Oh. Dr.LLM: Dynamic layer routing in LLMs. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://arxiv.org/abs/2510.12773.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://arxiv.org/abs/2009.03300.

Ahmadreza Jeddi, Marco Ciccone, and Babak Taati. LoopFormer: Elastic-depth looped transformers for latent reasoning via shortcut modulation. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://arxiv.org/abs/2602.11451.

Ferdinand Kapl, Emmanouil Angelis, Kaitlin Maile, Johannes von Oswald, and Stefan Bauer. From growing to looping: A unified view of iterative computation in LLMs. *arXiv preprint arXiv:2602.16490*, 2026. URL https://arxiv.org/abs/2602.16490.

Ziyue Li, Yang Li, and Tianyi Zhou. Skip a layer or loop it? test-time depth adaptation of pretrained LLMs. *arXiv preprint arXiv:2507.07996*, 2025. URL https://arxiv.org/abs/2507.07996.

Ziyue Li, Yang Li, and Tianyi Zhou. Skip a layer or loop it? learning program-of-layers in LLMs. In *International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2606.06574.

Fangzhou Lin, Shuo Xing, Peiran Li, Siyuan Yang, Qianwen Ge, Kazunori Yamada, Ziming Zhang, Haichong Zhang, and Zhengzhong Tu. CAPS: Cascaded adaptive pairwise selection for efficient parallel reasoning. *arXiv preprint arXiv:2605.15513*, 2026. URL https://arxiv.org/abs/2605.15513.

Xin Men, Mingyu Xu, Qingyu Zhang, Bingning Wang, Hongyu Lin, Yaojie Lu, Xianpei Han, and Weipeng Chen. ShortGPT: Layers in large language models are more redundant than you expect. *arXiv preprint arXiv:2403.03853*, 2024. URL https://arxiv.org/abs/2403.03853.

Taekhyun Park, Yongjae Lee, Dohee Kim, and Hyerim Bae. LoopUS: Recasting pretrained LLMs into looped latent refinement models. *arXiv preprint arXiv:2605.11011*, 2026. URL https://arxiv.org/abs/2605.11011.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

David Raposo, Sam Ritter, Blake Richards, Timothy Lillicrap, Peter Conway Humphreys, and Adam Santoro. Mixture-of-depths: Dynamically allocating compute in transformer-based language models. *arXiv preprint arXiv:2404.02258*, 2024. URL https://arxiv.org/abs/2404.02258.

Nikunj Saunshi, Nishanth Dikkala, Zhiyuan Li, Sanjiv Kumar, and Sashank J. Reddi. Reasoning with latent thoughts: On the power of looped transformers. *arXiv preprint arXiv:2502.17416*, 2025. URL https://arxiv.org/abs/2502.17416.

Tal Schuster, Adam Fisch, Jai Gupta, Mostafa Dehghani, Dara Bahri, Vinh Tran, Yi Tay, and Donald Metzler. Confident adaptive language modeling. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL https://arxiv.org/abs/2207.07061.

Yuxuan Tong, Xiwen Zhang, Rui Wang, Ruidong Wu, and Junxian He. DART-Math: Difficulty-aware rejection tuning for mathematical problem-solving. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL https://arxiv.org/abs/2407.13690.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2203.11171.

Yuxiang Wang, Kunyu Feng, Yingda Shen, Haoning Xu, Junyu Wang, and Zhizheng Wu. RecurTrace: Adaptive latent reasoning with loop-time memory. *arXiv preprint arXiv:2609.03379*, 2026. URL https://arxiv.org/abs/2609.03379.

Mingqian Yu, Wenpeng Zhang, and Peilin Zhao. T-LoopFormer: Token-level elastic-depth looped transformers for latent reasoning with dynamic routing. *arXiv preprint arXiv:2609.15160*, 2026. URL https://arxiv.org/abs/2609.15160.

CONTENTS

1 Introduction **1**

2 Setup and Measurement Protocol **2**

3 From Candidate Coverage to Output Accuracy **3**

3.1 Layer edits cover questions missed by ordinary sampling 3

3.2 Direct voting realizes little of the additional coverage 3

4 Structure and Transfer of Layer-Edit Success **5**

4.1 Discovery validity narrows evaluation-prompt coverage 5

4.2 Constituent regularity and successful exceptions 5

5 Prediction Signals for Detection and Selection **6**

6 Limitations **7**

7 Related Work **7**

8 Conclusion **8**

A Evidence Guide and Scoring Conventions **14**

B Routing Protocols and Conditional-Transfer Controls **14**

B.1 Candidate formation, fitting, and final decisions 14

B.2 Fixed-output statistics, supervision, and sampling 15

C Prediction Features and Latent-Score Simulations **17**

C.1 The prefill probe: features, model class, and folds 17

C.2 Three loophole checks on the question-conditioned ranker 19

C.3 Correctness detection and candidate ranking 19

C.4 Latent-score simulations under explicit scoring conventions 20

D Composition and Descriptive Residual Estimation **21**

D.1 Replication and enrichment of FF→✓ events 21

D.2 Composition: the two nulls in full 22

D.3 What the four prompt conditions do—and do not—identify 23

D.4 The variance-component estimator, derived 25

D.5 Estimator sensitivity and registered checks 27

E Reproduction, Generalization, and Registration **29**

E.1 Data and generation 29

E.2 Execution costs and available artifacts 30

E.3 Accounting invariants 30
E.4 Model and task scope 30
E.5 Registration and deviations 31

F Comparison with Related Routing Methods **31**

G Detailed Measurement and Statistical Records **32**
G.1 Formal output accounting and oracle definitions 33
G.2 Coverage counts and intervals 34
G.3 Historical shortlist and conditional-transfer diagnostics 34
G.4 Complete prediction statistics and inspected workpoints 34

A Evidence Guide and Scoring Conventions

The main text contains the definitions and comparisons needed to read its results. This supplement supplies the corresponding protocols, controls, and derivations. The detailed companion provides extended analyses (Appendix C, *Detailed Protocols and Supporting Analyses*), an English presentation of the historical registration and deviations (Appendix I, *Pre-registration and Disclosed Deviations*), and correction history (Appendix K, *Consolidated Correction Log*); it is not required to interpret the main-text results. Companion section numbers below refer to that PDF; other appendix references refer to this PDF. The offline supplement (companion Appendix B) gives the new coverage and decision experiments in the companion; the routing sections below document the original protocols.

Main-text object	Supporting material
Candidate identity and label access	Routing protocol, Appendix B
New coverage and direct voting	Offline protocols, companion Appendix B
Original output gains and SC comparison	Frozen replay, Appendix B.2
Width and supervision diagnostics	Conditional-transfer controls, Appendix B
Correctness detection and path ranking	Features and populations, Appendix C.1
Constituent successes and residuals	Composition and estimator, Appendix D
Generation, costs, and registration	Reproduction and scope, Appendix E

Source identifiers. RouteA, routeB, and routeC mean factorized single-path selection, conditional classification, and structural voting. B1–B4 identify four analysis products: fixed-output statistics, candidate-width controls, supervision controls, and voting/self-consistency controls. A *global* set is formed across the stated question population; a *nested* cover and selector are fitted inside each outer training fold. These fitting scopes do not specify whether the final decision reads a target question's labels.

Refund and no-refund normalization. Accounting (a) is post-hoc KEEP refund. Accounting (b) credits the selected output without a correctness-based refund. Let $A_q = C_{q,\text{KEEP}}$ and let V_q^T be correctness under $T \in \{C, R\}$ for a selected edit, with $V_q^T = A_q$ whenever KEEP is emitted. For a stated positive reference G,

$$c_{\text{fb}}^T = \frac{N^{-1}\sum_q \max(A_q, V_q^T) - L_0}{G}, \qquad c_{\text{dep}}^T = \frac{N^{-1}\sum_q V_q^T - L_0}{G}, \quad L_0 = N^{-1}\sum_q A_q.$$

The source identifier c_{dep} denotes no-refund accounting; target label access must be checked separately. C grades rewrite answers, while $R = D \wedge C$ additionally requires discovery validity. The primary output metric uses C. For example, the difficulty-5 structural voter's C-scored refund and output ratios against $G = 125/565$ are $61/125$ and $12/125$. That discovery-constrained reference is not an upper bound on all C-scored gains.

B Routing Protocols and Conditional-Transfer Controls

B.1 Candidate formation, fitting, and final decisions

The archived factorized single-path selector ranks the full 1,261-path pool with predicted constituent log-odds. SUB scores sum the SKIP and REP terms; marginal and KEEP paths have their own scores. The frozen Qwen2.5 construction maps a winning KEEP action to edit index 1, yielding a forced-action decision. These implementation choices are retained when its emitted paths are scored.

The archived structural voter uses the same ranking to form a fresh top-57 list for each question. It counts nonempty extracted answer strings, inserting the deterministic KEEP answer first and only once, followed by non-KEEP candidate answers. Plurality ties prefer the earliest inserted answer. It emits the first candidate path matching the winner; if there is no match, the helper draws a candidate at random. Thus a winning KEEP answer outside the candidate list need not be emitted. The scoring ledger records the emitted path. There are 57–58 generations, plus prefill feature acquisition.

The conditional-transfer classifier fits a baseline-correctness gate and a path classifier within question folds. The gate uses three discovery-prefill copies, predicts 0-shot KEEP correctness, and uses threshold 0.5. The final eligibility test additionally requires the target's 0-shot KEEP to be wrong and some edit to be D-valid. It is therefore a conditional-transfer diagnostic. The fold-specific cover greedily covers D-valid paths on baseline-correct training questions, whereas selector labels are the first cover-valid paths on baseline-wrong training questions. At difficulty 5 the cover sizes are $16, 19, 20, 16, 20$. There are 452 outer-training questions per fold before label filtering; exact usable-label counts are not preserved. The frozen decision vector contains 101 edits and 464 KEEP returns.

A separate global-cover classifier uses a four-condition cover on 129 0-shot-baseline-wrong, R-recoverable questions, with 103 or 104 independent selector-training questions per fold. Its net $+2$ at difficulty 5 and -3 at difficulty 3 refer to that separate construction. The width 57 originates in this global cover and is retained as a specified budget, rather than an independently selected optimum. The nested classifier's net is zero at difficulty 5. These counts and candidate identities must remain separate.

The globally selected fixed path maximizes R rate on all target questions; its difficulty-5 path index is 154. Measured-marginal argmax bins each target question's observed SKIP and REP discovery rates into $\{0, 1/3, 2/3, 1\}$ and uses question-fold-fitted bin scores to select a SUB. Both are label-assisted references. Uniform random single-path selection draws from all 1,261 paths; random 57-vote draws distinct paths without an extra forced KEEP ballot. Their C-output gains are distinct quantities.

B.2 Fixed-output statistics, supervision, and sampling

Fixed-output scores and uncertainty. For all $N = 565$ questions, define $X_q = C_q(\text{output}) - C_q(\text{KEEP})$. Recovered and destroyed counts are the numbers of $+1$ and -1 contributions, net is their difference, and $\Delta\text{Acc} = 100 \sum_q X_q/N$ percentage points. The corrected 95% interval uses 2.5/97.5 percentiles of the empirical question-level paired-bootstrap distribution, conditional on fixed outputs. The statistics product records both Monte Carlo estimates ($B = 10{,}000$, seed 20260913) and exact quantiles from the multinomial distribution on $\{-1, 0, 1\}$; we use its exact columns for the original fixed-output arms. "Exact" refers to computation of this empirical distribution, not guaranteed finite-sample coverage. The original 5/95 intervals are 90%, not 95%. Neither calculation includes probe fitting, candidate selection, generation randomness, or selection of a working point after inspecting the test results. Random-draw percentiles are reported separately, and an interval spanning zero is not an equivalence test. The original fixed-output comparisons use the restored arm-level summary of recoveries, destructions, and confidence intervals.

Candidate families and allowed information. The candidate-width scan uses global discovery-rate top-K candidates selected using all questions' discovery outcomes. Its learned, random, fixed and SUB variants retain target eligibility $\neg C_{0\text{shot},\text{KEEP}} \wedge \exists p\, D_{qp}$. The "ungated" variant removes the fitted baseline gate but retains that eligibility restriction; it does not act unconditionally on all questions. These are exploratory condition-transfer diagnostics, not independent-question deployment evaluations. The nested-cover family is constructed inside the outer folds, using the baseline-correct cover cohort described above. Its diff-5 sizes are $[16, 19, 20, 16, 20]$ and diff-3 sizes $[8, 10, 10, 11, 9]$; neither is a fixed 57-element cover.

The matched C-only oracle uses KEEP $\cup\, S_q$ and final rewrite correctness; the matched constrained reference additionally requires D_{qp}. They are label-using references for a specified candidate family. The historical H-rate top-57, global D-rate top-57 and the archived voter's per-question predicted-margin lists are different. At diff-5, the latter two have C-only/constrained recoveries $198/75$ and $196/61$, respectively; the nested covers give $157/50$. At diff-3 the nested covers give $72/28$. These references separate candidate coverage from selection within those candidates without equating 125 with a universal C-scored oracle bound.

The positive points in this scan are retained: diff-5 reaches $+11$ at $K = 57$, and the eligibility-restricted ungated diff-3 variant reaches $+11$ at $K = 20$ and $+13$ at $K = 30, 57$. These do not validate a full-population forced-action policy or identify an optimal K. The SUB-only candidate

Table 6: Rewrite-scored diagnostics across shortlist sizes.

K	diff-3			diff-5		
	rec/dst	net Q	95% CI	rec/dst	net Q	95% CI
1	0/0	0	$[0, 0]$	11/14	-3	$[-13, 7]$
5	3/3	0	$[-5, 5]$	13/14	-1	$[-11, 9]$
10	3/3	0	$[-5, 5]$	14/14	0	$[-10, 10]$
20	5/4	+1	$[-5, 7]$	17/11	+6	$[-4, 16]$
30	4/3	+1	$[-4, 6]$	16/12	+4	$[-6, 14]$
57	6/4	+2	$[-4, 8]$	22/11	+11	$[0, 22]$

Notes. The learned gated selector uses target-discovery eligibility and a global D-rate ranking, unlike frozen nested conditional classification. Intervals are exact fixed-output 95% bootstrap quantiles, not simultaneous scan intervals.

Table 7: Rewrite-scored net recovery by supervision type.

Training pool	K	First-valid	Multi-positive	Utility
diff4+diff5	10	+2	+4	+8
diff4+diff5	57	-2	-1	+2
diff5-only	10	+1	+4	+2
diff5-only	57	0	+2	+2

Notes. Full selector-training fraction, fixed folds, and target-label assistance. The five seed outputs coincide because subsampling seeds have no effect at fraction one.

controls match single-path layer counts, but retain the discovery-label and full-pipeline cost differences.

Supervision and independent training questions. First-valid uses one path label per discovery-recoverable training question; multi-positive fits one-vs-rest path predictors using discovery-validity labels; utility regresses $D_{qp} - D_{q,\text{KEEP}}$, with ridge penalty 1, and requires positive predicted utility to edit. This utility is not the single-rewrite accuracy difference. All three inherit the archived baseline-correct cover cohort and target-discovery eligibility. Outer test folds have 113 questions; the diff-5 selector training pool has 452, or 1,017 after adding all 565 diff-4 questions. Fractions $0.25/0.5/0.75/1$ give $113/226/339/452$ and $254/508/763/1017$ independent training questions. The cached gate uses the full fold training pool, so these are selector-size curves, not total-label-budget curves. Positive question-path cells in the fold diagnostic are not counts of independent positive questions.

The retained supervision comparison reports marginal arm intervals but no per-question picks or paired supervision-difference interval. Full-fraction seed repetition is not independent evidence of training stability. The recorded utility improvement in the pooled $K = 10$ configuration is a point comparison, not a significant superiority or a demonstrated scaling law.

KEEP self-consistency and archived structural voting. The sampled KEEP baseline uses temperature 0.7, top-$p = 1$, maximum 2,048 output tokens, and 57 responses per question (seed base 20260913 plus the sample index). Prefixes of size $K = 1, 5, 10, 20, 57$ are voted separately. The vote counts nonempty extracted answer strings, breaking ties by first occurrence; semantic equivalence is checked only after choosing the winner. An all-empty vote is incorrect. The archived voter instead includes a KEEP ballot plus non-KEEP shortlist answers and maps its winner back to a candidate path; its tie and fallback behavior is specified in Appendix B. Random 57-path voting uses no extra forced KEEP ballot. These algorithms therefore do not have identical voting rules.

The current replay regraded all three raw SC files with the original evaluator, reproduced the published rec/dst counts, and computed paired SC-minus-voter intervals from fixed per-question outputs ($B = 10{,}000$, seed 20260913). Sixteen retained parsing exceptions and one 60-second timeout are treated as incorrect, as in the recorded evaluation. Alignment follows ascending raw sample indices, matching the archived ordering; an independent source-question-ID manifest is unavailable. Qwen3 ground-truth strings agree in all 565 rows between its cached labels and sorted raw samples. The

Table 8: Self-consistency versus the archived structural voter.

Model / level	SC rec/dst	SC net	95% CI	SC−vote	Paired 95% CI
Qwen2.5 / d5	78/18	+60	[42, 79]	+48	[29, 67]
Qwen2.5 / d3	61/3	+58	[43, 73]	+27	[16, 38]
Qwen3 / d5	29/11	+18	[6, 30]	+23	[11, 35]

Notes. $K = 57$; vote denotes the archived structural voter. Net gains and differences are in questions. SC intervals use exact empirical-bootstrap quantiles; paired differences use 10,000 question-bootstrap draws. All intervals are 95%, conditional on fixed generations and recorded sample alignment.

Table 9: Detector audit of exported output counts.

Level	Protocol	τ	rec/dst	net Q	95% CI
d5	0-shot	0	31/6	+25	[14, 37]
d5	0-shot	(0, 1]	0/0	0	[0, 0]
d5	mixed rewrite	0	49/64	-15	[−36, 6]
d5	mixed rewrite	(0, 1]	27/48	-21	[−38, −4]
d3	0-shot	0	17/5	+12	[3, 21]
d3	0-shot	(0, 1]	0/0	0	[0, 0]
d3	mixed rewrite	0	35/52	-17	[−35, 1]
d3	mixed rewrite	(0, 1]	23/41	-18	[−34, −2]

Notes. Exported counts, not a new evaluation. Intervals exclude probe and threshold selection. At 0-shot diff-5, $\tau = 0$ gives positive net gain but damage $6/565 = 1.06$pp exceeds the historical 1pp cap.

results below are conditional on that alignment and on the recorded generated answers and archived voter decisions, with no new sampling or fitting.

SC@57 exceeds the archived voter for all three archived comparisons, and the conditional paired intervals exclude zero. This supports superiority for these fixed outputs under a nominal completion budget, not an explanation of the causal source of the archived voter's gain. The archived voter can require 58 generations when KEEP is absent from its top 57; structural token/runtime costs were not measured. SC output token totals are 25,197,288 (Qwen2.5 diff-5), 18,573,365 (Qwen2.5 diff-3), and 43,696,170 (Qwen3 diff-5), with recorded batch times of approximately 1,321/1,025/3,155 seconds. Those are batch timings, not per-request latency; the memory reservation estimate is not measured incremental peak memory. Random-vote mean nets are $-2.54/+18.18/-12.80$, with random-candidate-draw 2.5/97.5 percentile ranges $[-13, 8]/[10, 26]/[-20, -5]$; those ranges are not question-bootstrap intervals. The archived voting comparison "diff-3 nested conditional classification" row reuses diff-5 picks and is excluded from valid diff-3 results.

Detector workpoints: retain the prompt boundary. Table 9 reports the exported counts and exact fixed-count empirical-bootstrap intervals. It does not replay probe selection. The original 0-shot workpoints used the final fold's exported probe set on all questions; their net gains are exploratory rather than independent held-out efficacy estimates. The mixed rewrite construction chooses a different, rewrite-named probe set, uses 0-shot agreement to decide, and scores the selected path on rewrite. It is neither a pure rescoring of the original fixed probe set nor a rewrite-only gate. No comparison attributes changes between those rows solely to prompt transfer.

C Prediction Features and Latent-Score Simulations

C.1 The prefill probe: features, model class, and folds

The prefill detection and selection probes in Table 10 use the following features, model class, and question-fold construction. The post-generation detection signals in Table 5 and the latent-score simulation in Table 11 are separate analyses.

Features. The probe reads the residual-stream hidden states of the *base* model (Qwen2.5-3B-Instruct), unedited path, *prefill only*: features are computed from the prompt forward pass, before any output token is generated. A single layer is used: the output of layer 18 of the 36 transformer layers,

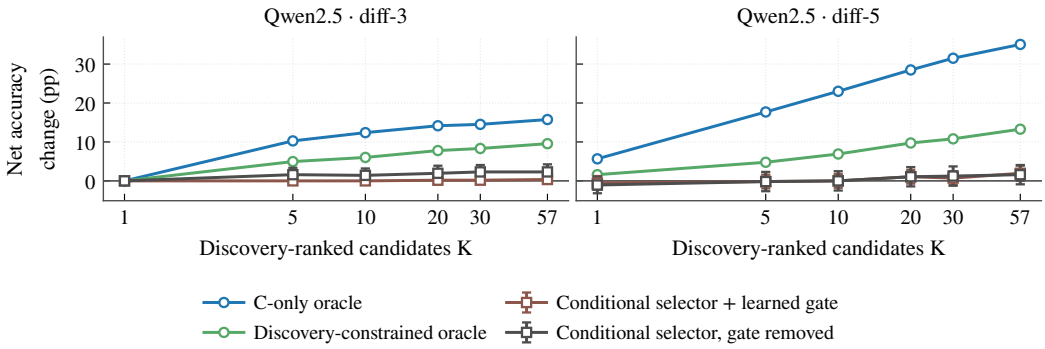

Figure 2: C-scored candidate-width diagnostics, using globally discovery-ranked candidates. Both oracle lines and both selectors use target labels as described in the text. Selector error bars are fixed-output 95% empirical-bootstrap intervals. The learned gate and the discovery eligibility filter are different; the latter remains in both selector curves.

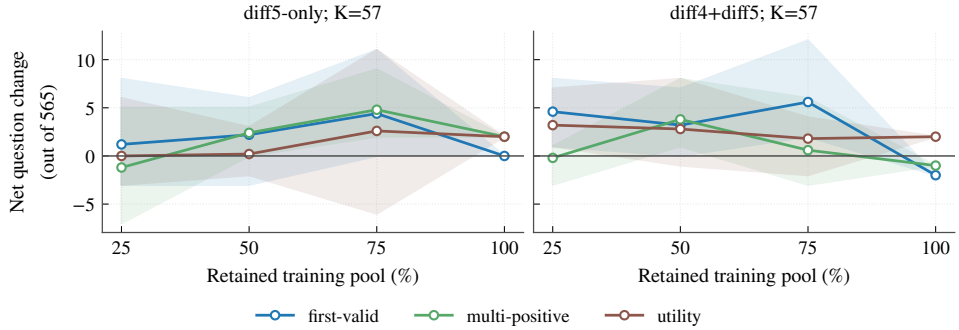

Figure 3: Selector training-fraction diagnostics at cover cap $K = 57$. Lines are means over five subsampling seeds; shading is their range, not a confidence interval. The full-fraction fits coincide by construction. Gate fitting uses the full pool throughout, and all arms retain target-discovery eligibility.

counting the embedding output as layer zero. Pooling is the mean over the user-body token span—the tokens of the user message as rendered by the chat template, excluding the system block and the assistant preamble—giving one 2048-dimensional vector per question. (The extraction pipeline also produces layers 9 and 27 and last-token pooling as sensitivity configurations; the single-layer probe in Table 10 uses layer-18 mean.) The P14 ranker in Table 5B is a different model: it learns token-attention pooling and layer weights across all 36 layers, and scores the resulting question vector against factorized path embeddings. Its reported AUROC uses the question–path interaction score. The post-generation signals in Table 5A are defined separately.

Reduction and model class. Within each training fold the features are standardized (per-dimension mean/scale computed on that fold) and reduced by PCA to 256 components using a full singular-value decomposition (fit on the training fold). The scoring model is a per-target penalized binary logistic regression, $C{=}1/\lambda$ with $\lambda{=}10^{-3}$, optimized by L-BFGS for at most 1,000 iterations, outputting a binomial out-of-fold probability. For the selection target there is one probe per single-edit action—70 in total (34 SKIP + 36 REPEAT), each labeling the validity of that edit on the question—plus a KEEP probe; targets that are single-class within the training fold (a layer that is nearly always valid or never valid) are skipped and replaced by the constant base-rate prior. The regularization constant is a pre-registered fixed value, not tuned; in particular it was never selected on the evaluation split, which would have undone the dissociation of §5.

Labels. Two targets. *Detection*: $y_{\text{KEEP}}(q) = \mathbf{1}[\text{unedited answer on } q \text{ is correct}]$, scored pooled over questions. *Selection*: $y_p(q) = \mathbf{1}[p \text{ is valid on } q]$, with validity the single-edit held-out criterion (discovery under D, scored on the held-out condition), restricted to non-degenerate targets (base rate in $[0.1, 0.9]$; 64 of the 70) and evaluated within question.

Folds. Grouped 5-fold cross-validation by question identity (question md5 prefix), fold assignment fixed once; out-of-fold probabilities on the held-out fold. The *same folds serve both targets*—the dissociation argument of §5 is only valid if the two rows of Table 10 share a fold partition, and they do.

The question-independent reference. For either target the appropriate reference is the best predictor that ignores the question. For selection this is the global per-path validity rate, $\hat{r}_p$ = fraction of questions on which path p is valid, scored as a fixed path score. The in-sample value (0.682) fits rates on all 565 questions, including the scored questions, and evaluates on 125 baseline-error questions with non-degenerate R labels. The historical K1 cross-fit value (0.6770) uses training-fold rates from all questions and excludes KEEP from its 1,260-path evaluation pool. These are distinct from the unified P14 prior (0.6765) in Table 5, whose 1,261 paths include KEEP. Neither "all questions" nor "error-only" in that table names a test population: the terms specify the ranker's training scope. The 0.500 detection reference is chance.

The n=125 within-question set. The selection row of Table 10 is evaluated within question on the n=125 held-out *recoverable* questions at diff-5: the rewrite baseline is incorrect and the evaluated non-KEEP candidate pool contains both valid and invalid R labels. The single-edit probes supply scores to this 1,260-candidate evaluation, which includes substitutions. This is the same 125-question recoverable set that defines G_{heldout} (§4.1), and it is *narrower* than the baseline-wrong population over which the route-probe AUROCs of §3 are computed. The selection entry of Table 10 is therefore a within-question estimator on the recoverable subset to which the pooled route probe is not directly comparable; the dissociation argument of §5 rests on scoring both of the table's rows on their respective best question-independent references (row 2: 0.631 vs 0.682), not on comparing the two entries of the table against each other across populations.

Alternative probe classes. On the 125-question within-question evaluation, concatenating layers $\{9, 18, 27\}$ yields AUROC 0.631, last-token pooling 0.577, and a tested 32-unit nonlinear head 0.553. An external Qwen3-Embedding-0.6B representation under grouped five-fold PCA/logistic fitting reaches 0.622. These are outcomes of the tested representations and fitting procedures. A lower AUROC alone does not isolate overfitting, sample size, or capacity as its cause. The results do not rule out better selection with other training designs or representations.

C.2 THREE LOOPHOLE CHECKS ON THE QUESTION-CONDITIONED RANKER

Three historical auxiliary checks characterize the tested ranker. They retain their original fits and structural references; their values are not combined with the P14 paired intervals in Table 5. Zeroing its question–path interaction changes within-question AUROC by -0.003 ($[-0.021, +0.013]$), providing little evidence of an improvement from that interaction in this fit. A rank-4 layer-pair term yields AUROCs 0.660 in the all-question regime and 0.633 in the error-only regime; both remain below the specified structural reference. The error-only increment interval is $[-0.074, -0.014]$.

The shuffled question–path control gives an error-only increment of -0.033 rather than the registered zero point. Relative to this control, the two regimes have increments $+0.021$ and $+0.040$; these are secondary comparisons with a different reference. Their signs alone neither diagnose leakage nor identify an overfitting mechanism. The main comparison now uses the single P14 implementation in Table 5: the all-question-trained ranker scores 0.6653 against a paired prior of 0.6765, with increment -0.0112 and interval $[-0.0352, 0.0135]$. The error-trained ranker scores 0.6279, with increment -0.0486 and interval $[-0.0800, -0.0183]$. Both are evaluated on the same 125 rewrite-baseline-error questions with non-degenerate R labels, using 1,261 paths including KEEP; the prior uses all training-fold questions. The all-question interval includes zero. These results delimit the tested features, labels, and function classes; they do not establish a general pre-generation selection ceiling.

C.3 CORRECTNESS DETECTION AND CANDIDATE RANKING

Structural agreement, prompt-condition agreement, temperature agreement, and answer log-probability in Table 5 predict 0-shot KEEP correctness. The prompt-condition arm compares available KEEP answers from the four prompt conditions with the 0-shot answer; it does not inject

Table 10: Historical prefill detection and candidate-ranking diagnostics.

Target	Question-indep. ref.	Prefill probe	Increment	Paired 95% CI
Detection (y_{KEEP})	0.500	0.740	$+0.240$	—
Selection ($y_p(q)$, within-q., n=125)	0.682	0.631	$\mathbf{-0.051}$	$[-0.078, -0.024]$

Notes. Prefill detection and candidate-ranking diagnostics using the same feature family and question folds. Detection is pooled over baseline outcomes and compared with chance. Selection is within-question R-validity AUROC over 1,260 non-KEEP candidates on 125 rewrite-baseline-error questions with non-degenerate labels. Its 0.682 reference fits per-path rates on all 565 questions, including evaluated questions: it is an in-sample all-question prior, not a prior trained only on errors. The selection increment uses a paired question bootstrap; no such interval is released for detection. The P14 comparison in Table 5 instead uses a cross-fitted prior and 1,261 candidates including KEEP, so its prior and intervals are not interchangeable with this diagnostic.

activation noise. The temperature arm uses five sampled answers. The source audit records zero frequency of all five samples being identical. Parse-failure rates on incorrect versus correct questions are 0.174/0.160 at difficulty 3 and 0.123/0.121 at difficulty 5. These checks examine saturation and correctness-dependent extraction failure; they do not identify the cause of the agreement signal.

Detection AUROC concerns one baseline outcome per question. Selection AUROC concerns valid versus invalid paths within questions, with both kinds of path required to define that AUROC. Table 10 uses 125 baseline-error, recoverable questions and a matched in-sample reference; Table 5 compares all-question-trained and error-trained rankers on a common 125-question evaluation set. These values do not constitute normalized output gains. The accounting corrections and auxiliary ranker checks are detailed in the detailed companion, Sections C.1.6 (*Three accounting choices in the dissociation*) and C.3.2 (*Three loophole checks on the question-conditioned ranker*), respectively. The former is a historical diagnostic, distinct from the current P14 comparison.

The four post-generation detection arms and their fixed-signal, question-resampling intervals are reported in Table 5. Each arm uses its available records; these intervals do not include signal or threshold selection.

The unified P14 ranker comparison is in Table 5. Historical alternative fits and controls are described in Appendix C.2; their AUROCs and references are kept separate from the P14 paired intervals.

C.4 LATENT-SCORE SIMULATIONS UNDER EXPLICIT SCORING CONVENTIONS

The simulation assigns scores using ground-truth four-condition validity plus noise on the difficulty-5 grid. Its KEEP reference also uses true baseline correctness. This label-assisted construction illustrates how candidate width, top-1 selection, and refund accounting affect a normalized gain. Its scores are not features available to a learned selector at inference.

Refund simulation. The saved full-pool curve uses accounting (a), post-hoc KEEP refund, and normalizes recovery by $125/565$. At target AUROCs 0.55, 0.75, 0.90, and 0.95, the saved normalized gains are 0.032, 0.240, 0.464, and 0.672, respectively (Table 11). Thus the curve reaches gains above one half of this reference within the scanned range; 0.90 is not an exact threshold. These values come from the saved refund-simulation results, which also record the achieved AUROC. No new simulation is run for this presentation.

No-refund width scan. The separate width scan uses globally discovery-ranked edits, a noisy latent validity score, and a KEEP threshold computed from true baseline correctness. Its no-refund score credits R for an edit and C for an emitted KEEP. At $K = 1260$, 50% is not reached within the scanned AUROC range up to 0.95; at $K = 300$, it is reached near 0.947 (Table 12). Hence the statement that half the reference is never reached at any width would be false. The source computes its denominator as $\text{mean}(\max_p R_{qp}) - L_0$, an edit-validity reference, whereas the refund curve uses the KEEP-preserving oracle reference. These are separately saved scans with different candidate constructions and stated references, rather than a paired C-output policy comparison.

Interpretation and populations. Of the 257 rewrite baseline errors, 125 have a discovery-valid, rewrite-correct edit. The conditional path-validity density used in the conversion analysis is 0.0133

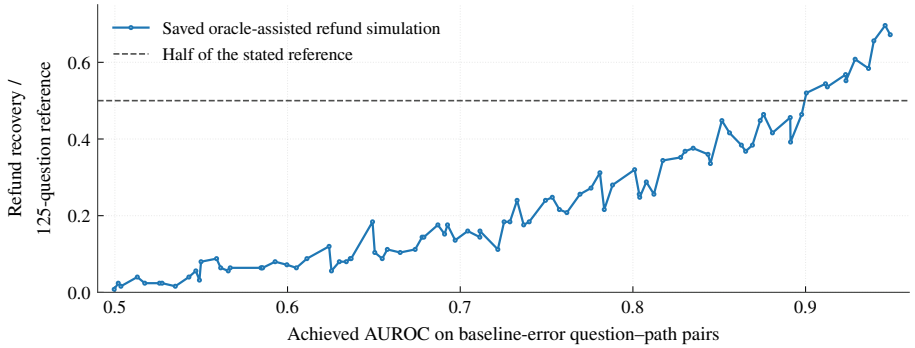

Figure 4: Saved latent-score simulation on difficulty 5. Scores use ground-truth R-validity plus noise; the vertical axis is post-hoc refund recovery normalized by $125/565$. The curve displays all 91 saved rows against achieved baseline-error-subset AUROC. It contains no measured routing points and supplies no deployment threshold. Table 11 reports selected source rows; the separate no-refund width scan is in Table 12.

Table 11: Selected results from the R-based refund simulation.

Target AUROC	Achieved AUROC	Normalized refund gain	Error-subset top-1 hit
0.55	0.549	0.032	1.56%
0.68	0.678	0.144	7.00%
0.75	0.749	0.240	11.67%
0.90	0.898	0.464	22.57%
0.95	0.949	0.672	32.68%

Notes. Selected saved rows of the R-based refund simulation, normalized by the KEEP-preserving reference $125/565$. Target and achieved AUROC differ because latent scores are sampled. The hit rate is over all 257 rewrite baseline-error questions. These are simulation outputs, not observed selector performance or AUROC requirements.

over question–candidate pairs on those baseline errors; the unconditional all-path validity rate 0.113 is a different population. The source bridge values 0.040 (the archived single-path selector) and 0.072 (the archived voter) use R-based refund accounting, rather than the current C-scored net gains. They are preserved as separate source records and are not plotted as realized accuracy. For example, the current difficulty-5 structural voter has C-refund recovery $61/125$ but C-output normalized net $12/125$. The measured nets in Table 2 are the output-level evidence; these latent-score simulations establish no AUROC requirement for an observable selector.

D Composition and Descriptive Residual Estimation

This section retains the constituent-event definitions, response-grid assumptions, and complete residual estimator supporting §4. The residual is descriptive, and the registered behavioral check fails. These analyses supply no upper bound on routing gain.

D.1 Replication and enrichment of FF→✓ events

Here strict FF means that both SKIP and REPEAT are wrong on each of the three discovery prompts. A discovery additionally requires SUB to be correct on all three; its unit is a question–path event, not a distinct question. Table 4 gives the pooled counts and rates. Scoring discovered FF→✓ events on the untouched *rewrite* condition gives $p_{\text{rep}} = 0.45$–0.51 across strata, with $p_{\text{rep}} = 0.4498$ at diff-5 on 438 discovery events. Against a comparator formed from *all* strict-FF events regardless of discovery status—the conditioning set that holds the FF condition fixed and varies only discovery—the enrichment is 9.93 / 11.16 / $15.55\times$ at strata 1 / 3 / 5, with an event-resampled bootstrap interval of $[13.89,\ 17.14]$ at diff-5 (discovered and matched rewrite outcomes are resampled as events, rather than by cell or question cluster; Appendix E). The matched pool includes the discovered events. Figure 5(b) displays the pooled replication and matched-comparator rates. The

Table 12: Latent-score width scan under no-refund and refund accounting.

K	AUROC for 30%	AUROC for 50%
1	—	—
5	—	—
30	—	—
57	0.95	—
300	0.91	0.95
1260	0.92	—
Refund scan, K=1260	0.80	0.88

Notes. Latent-score width scan using no-refund R-accounting, with true baseline correctness used in the KEEP threshold. Entries are interpolated achieved baseline-error-subset AUROCs for a specified normalized gain; "—" means not reached over target AUROC ≤ 0.95. This scan uses $\mathrm{mean}(\max_p R_{qp}) - L_0$ as its reference. The last row is a separate refund scan with a KEEP-preserving reference $125/565$; the rows are separate saved simulations, rather than a paired C-output policy comparison.

saved comparison excluding discovered events yields ratios of approximately 10–16×; we retain the all-strict-FF comparator because it fixes the marginal failure condition without conditioning on discovery failure. At diff-5 the strict-FF discovery rate is $438/235{,}873 = 0.1857\%$. The rewrite replication rate $p_{\text{rep}} = 197/438 = 0.44977$ has a different denominator; dividing it by the matched all-strict-FF rewrite success rate 0.02893 gives $15.55\times$. The historical 1.3% instead summarizes $5{,}625/444{,}133 = 1.2665\%$ in a different quadrant: neither component is D-valid, and the SUB outcome is R-valid. A non-D-valid component need only fail one discovery prompt, so this quadrant is not strict FF.

The reassignment null of §4.2—removing the FF-marginal pool's question labels in place and permuting which FF-marginal questions are marked as discovered, thereby breaking only the discovery–replication association—places the observed p_{rep} far above the null at every stratum: observed 0.51 / 0.46 / 0.45 against a null 99th percentile of 0.111 / 0.069 / 0.048 at diff-1/3/5, a permutation $p < 2 \times 10^{-4}$ at each stratum.

The independent reimplementation that reproduces both quantities to four decimal places was not free of information. The two implementations initially disagreed by a factor of 2.4 because the FF marginal condition—"valid on *zero* of the three discovery conditions"—had been re-read as "not valid on all three," which is nearly vacuous. That is the class of divergence an independent reimplementation exists to surface.

D.2 Composition: the two nulls in full

TT requires both components to be D-valid (three-prompt D-validity); the SUB outcome is R-validity (four-prompt R-validity). The diff-5 conditional failure rate $1 - 0.5765 = 0.4235$ admits two readings. Against perfect compositionality it is a 42.35% shortfall. Against independence, the relevant unconditional SUB-only R-success rate is 0.1070, so conditioning raises success to 0.5765, an enrichment of $5.39\times$ (Table 4). The historical all-path mean 0.113 mixes KEEP, SKIP, REPEAT, and SUB and is not this SUB base rate.

The saved decoupled permutation diagnostic permutes SUB outcomes against fixed marginals. Its 99th percentiles of conditional failure are $0.58/0.88/0.99$ at diff-1/3/5, compared with observed failure rates $0.1878/0.3213/0.4235$ (Table 13). Thus the registered criterion requiring observed failure to exceed that null was not met. Composition is informative without guaranteeing success: neither the conditional failures nor the enrichment alone establishes a mechanism of destructive interference.

The failure rate rises monotonically with stratum. We do *not* claim that interference increases with intrinsic difficulty, because the TT quadrant is itself a selection event whose frequency also varies with stratum: conditioning changes both the estimand and the subpopulation, and the selection displacement $\Delta\theta$ is directly measurable and varies by a factor of 3.5 across strata (Appendix D.5). A fully defensible test would regress Y_{ijq} over *all* cells with the measured $\hat{a}_i(q)$ and $\hat{b}_j(q)$ as continuous covariates, so that no conditioning on TT is required; that regression is not part of this audit. Two descriptive constructions corroborate the direction without conditioning on TT: $\Delta\theta$ *decreases*

Table 13: Decoupled-null diagnostic for conditional SUB failure.

Stratum	Observed $P(R_{\text{SUB}} = 0 \mid \text{TT})$	Decoupled-null p_{99}
diff-1	0.1878	0.58
diff-3	0.3213	0.88
diff-5	0.4235	0.99

Notes. Saved decoupled-null diagnostic for conditional SUB failure. TT uses D-valid components (three discovery prompts), and SUB failure is $1 - R_{\text{SUB}}$ over all four prompts. The null permutes SUB outcomes against fixed component marginals. Observed rates do not exceed the saved null 99th percentiles, so the registered gate fails. SUB-only base rates and success enrichment appear in Table 4.

with difficulty ($+0.885$, $+0.480$, $+0.255$) while the mixed-pool second-order correlation *increases* (raw 0.465, 0.817, 0.970, Table 16).

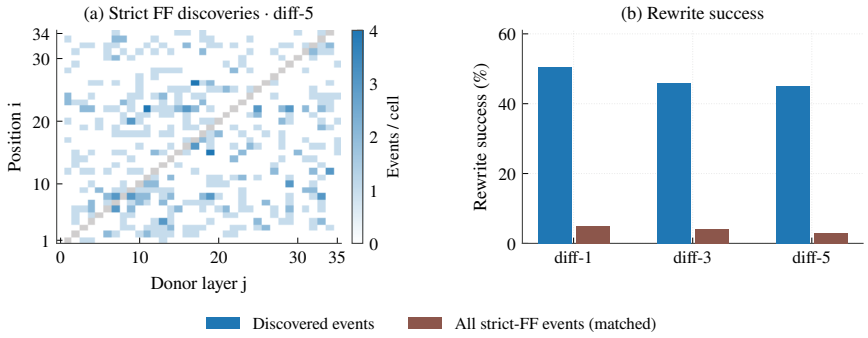

Figure 5: Strict-FF discovery events and rewrite replication. (a) Raw discovery-event counts per eligible (i, j) cell at diff-5; these counts are not normalized by the number of strict-FF opportunities. (b) Pooled rewrite success among discoveries and among all matched strict-FF events, including discoveries, at diff-1/3/5. Strict FF requires both components to fail every discovery prompt; discovery additionally requires SUB to succeed on all three. The replication counts are $50/99$, $134/292$, and $197/438$, and the matched rates are 5.086%, 4.113%, and 2.893%. These are question–path events, not independent-question counts (Appendix D.1).

D.3 What the four prompt conditions do—and do not—identify

This appendix states the best identification case the design can make and where it stops. The four-condition design gives more moments per cell than a single draw, but the conditions are distinct prompt templates—not independent replications—so it does *not* identify a separable interaction σ^2_δ out of the pooled residual of §4.2. We report the variance ratio only as a measured, descriptive quantity on this dataset; the estimator of Appendix D.4 is written for that design.

(i) **One condition does not identify the interaction.** Write the model of Eq. (D.4) one level down, adding the condition index explicitly,

$$Y^{(n)}_{ijq} = \mu + \alpha_i + \beta_j + \eta_q + \gamma_{ij} + a_i(q) + b_j(q) + \delta_{ijq} + \varepsilon^{(n)}_{ijq}, \tag{2}$$

with $\varepsilon^{(n)}_{ijq}$ the mean-zero Bernoulli noise of a single draw. With a single prompt condition there is exactly one observation per (i, j, q) cell. The model places no restriction on δ_{ijq}, so the residual of any fit is a single number equal to $\delta_{ijq} + \varepsilon_{ijq}$, and no observation tells the two apart: a question-specific interaction is indistinguishable from the outcome of a coin flip, in exactly the sense that σ^2_δ and σ^2_ε are jointly unidentifiable from the cell's one draw. Any "interaction" reported against a single condition is therefore a statement about the residuals, i.e. about noise, and is tautological in the direction we flag in §4.2.

(ii) n **conditions per cell give moment equations that separate the two.** When the same cell is evaluated under n conditions, summarize it by the empirical rate $\bar{Y}_{ijq} = n^{-1} \sum_n Y^{(n)}_{ijq}$. Two moments are now available. *Within-cell*: the sample variance of the n draws $v_{ijq} = \frac{1}{n-1} \sum_n (Y^{(n)}_{ijq} -$

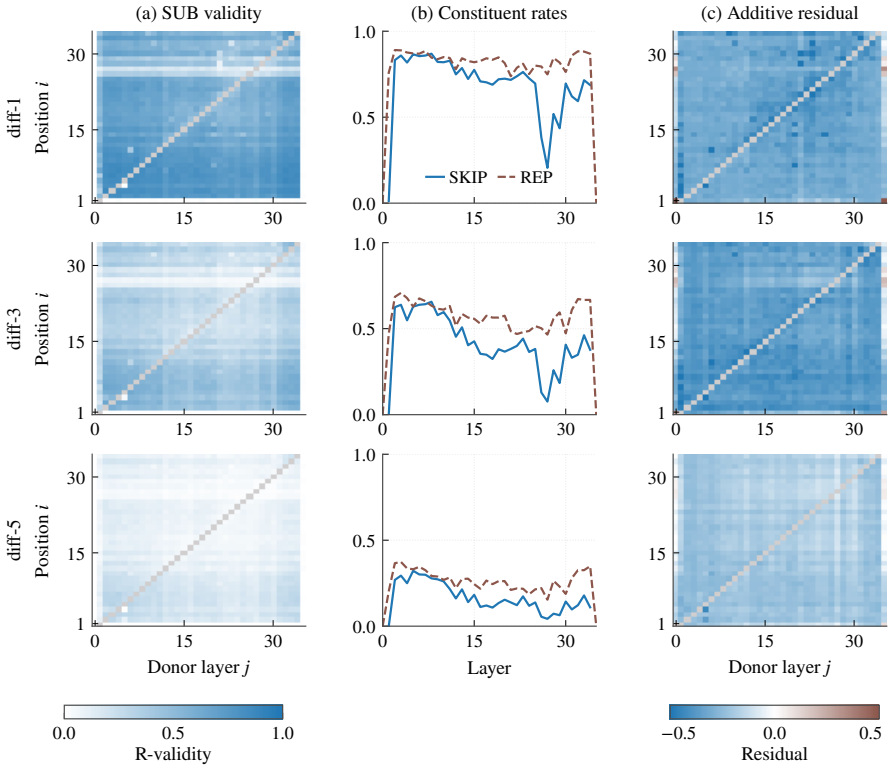

Figure 6: Aggregate prediction from measured constituent success rates. Each row is one difficulty stratum (diff-1, diff-3, diff-5). Left: mean SUB R-validity M_{ij} on the 1,190 eligible cells. Middle: measured SKIP rates $\hat{a}_i^{\mathrm{SKIP}}$ and REP rates $\hat{b}_j^{\mathrm{REP}}$, with correlation between M_{ij} and $\widehat{M}_{ij} = \hat{a}_i^{\mathrm{SKIP}} + \hat{b}_j^{\mathrm{REP}} - \bar{M}$ ($r = 0.950/0.943/0.912$). Right: the descriptive residual $M_{ij} - \widehat{M}_{ij}$, shown with a shared symmetric color scale across strata. It is not the fitted ANOVA row/column residual of Eq. (5) and need not have zero mean. The figure describes aggregate association and residuals, not a causal mechanism or a routing bound.

$\bar{Y}_{ijq})^2$ estimates σ_ε^2 (the Bernoulli draw variance, $p(1-p)$) with no contamination from δ_{ijq}, which is constant across the conditions of a cell. *Between-cell*: the variance of the extracted residuals $\hat{e}_{ijq}$ over cells is

$$\mathbb{E}\big[\mathrm{Var}\big(\hat{e}_{ijq}\big)\big] = \sigma_\delta^2 + \frac{\sigma_\varepsilon^2}{n}, \tag{3}$$

because a cell-rate with n draws has sampling variance σ_ε^2/n under conditionally independent, equal-variance noise (the interaction is a property of the question, not of the condition draw, so averaging over conditions reduces the noise floor by the factor n). The four-condition design is what makes the right-hand side interpretable: the within-cell moment pins σ_ε^2, and n in the denominator is design-supplied rather than assumed. The estimator of Appendix D.4 operates on the rate grid $\bar{Y}_{ijq}$ and reports the between-cell variance directly. That variance still carries the within-cell sampling floor of Eq. (3), σ_ε^2/n. This floor is measured directly: the within-cell moment gives $\hat{\sigma}_\varepsilon^2 \in [0.065, 0.094]$, so $\hat{\sigma}_\varepsilon^2/4 \in [0.016, 0.023]$—on the order of a third of the raw $\hat{\sigma}_e^2$ on each row of Table 14. That table reports the quantity twice: the raw $\hat{\sigma}_e^2$ (a measured residual variance whose composition is not identified) and the floor-corrected ratio, which subtracts this floor. Subtracting the floor *raises* every ratio (by $\approx 1.4\times$ in the least favourable cell, $0.0403 \to 0.0547$ at diff-1), yet all reported corrected cells still clear the 0.10 gate, with the maximum rising to 0.055; the verdict therefore does not turn on the subtraction either way. Two facts make the raw column the conservative direction. First, the floor is the variance of the n-condition average in Eq. (3): a single-condition router faces the un-averaged floor σ_ε^2 (≤ 0.25, Bernoulli), four times larger, which is exactly the factor $n{=}4$ buys (point (iii) below). Second, the residual is not purely floor: the discriminating test shows it reproduces across prompt conditions (Appendix D.4, Split B): $\hat{e}$ reproduces across the two disjoint condition halves ($r_B{=}0.53$–0.64, far above the permutation-null p_{99} of ≈ 0.004–

0.005), which independent sampling noise would not. As §4.2 stresses, that reproducible residual still pools per-question marginal structure with any genuine three-way interaction—a faithful simulator with per-question marginals and zero interaction reproduces a passing Split B gate—so it shows a question-specific *residual* is reproducible, not that a separable δ_{ijq} interaction is identified; the raw column is a measured residual variance of unspecified composition, and §4.2 reports it only in those terms.

(iii) **Why n=4 specifically.** The discriminating test of Appendix D.4 (Split B) re-derives $\hat{e}$ on two *disjoint halves* of the condition axis, $\{0\text{-shot}, 1\text{-shot}\}$ against $\{2\text{-shot}, \text{rewrite}\}$, and correlates the two estimates over all (i, j, q) triples. The split is only a split if each half is non-empty, so n=4 is the smallest even n that gives two disjoint pairs; n=2 would force 1 vs 1, leaving each half's $\hat{e}$ as a single-condition residual (point (i)). Correlation across those residuals can still detect shared structure under independent noise, but neither half supplies a within-cell variance estimate. With n=4, each half contains two observations, allowing the within-half moment calculation in point (ii) and reducing the sampling variance of each half's mean. The choice of the particular split $\{0, 1\}$ vs $\{2, \text{rewrite}\}$ is deliberately asymmetric: it holds the two k-shot conditions together and separates the paraphrase (*rewrite*) from the template family, so that Split B's reproducibility measures agreement across the two construction families rather than across two instances of the same family.

(iv) **Why this does not identify a separable interaction.** Eq. (3) separates σ^2_ε from σ^2_δ only if the n conditions were conditionally independent, identically distributed Bernoulli draws within each cell. They are not draws: the four conditions are distinct prompt templates and the paraphrase (*rewrite*) differs in kind, so conditional independence and the $1/n$-averaging reduction of Eq. (3) are not established. Consequently the four-condition design does *not* by itself identify a separable interaction, and we make no such claim: the variance ratio is a measured descriptive quantity whose composition (marginal vs. interaction) is not resolved, and the routing conclusions of §3 do not depend on resolving it.

D.4 The variance-component estimator, derived

This is the estimator behind every number in Table 14, written so that a reader can reimplement it from the equations alone; the independent-reimplementation story of Appendix E.3 is only credible if a third party can run it.

Data matrix. For a fixed criterion, form the rate grid $G_{ijq} = n^{-1} \sum_n Y^{(n)}_{ijq}$: the validity rate of the edit $\text{SUB}(i \leftarrow j)$ on question q across the n=4 prompt conditions (Appendix D.3). The grid is $[n_q \times 34 \times 36]$ because the 34 eligible replacement positions $\times$ 36 eligible donor layers define the cells; its 34 diagonal cells i=j and any cell without a path in the 1,261-path pool are structural missing values and are excluded from every mean and every variance below.

Model and side conditions. The descriptive specification, discussed in §4.2, is:

$$G_{ijq} = \underbrace{\mu}_{\text{global}} + \underbrace{\alpha_i}_{\text{delete-}i\text{ main}} + \underbrace{\beta_j}_{\text{repeat-}j\text{ main}} + \underbrace{\eta_q}_{\text{question}} + \underbrace{\gamma_{ij}}_{\text{layer-pair interaction}} + \underbrace{(\alpha\eta)_{iq}}_{a_i(q)} + \underbrace{(\beta\eta)_{jq}}_{b_j(q)} + \delta_{ijq} + \varepsilon_{ijq}, \tag{4}$$

with $i \in \{1..34\}$, $j \in \{1..36\}$ (layer indices; two boundary positions are excluded from deletion, §2), $q \in \{1..n_q\}$. Side conditions: $\sum_i \alpha_i = \sum_j \beta_j = \sum_q \eta_q = 0$; γ is centered over the eligible grid; the two question$\times$margin terms $a_i(q)$ and $b_j(q)$ are *not* restricted, which is the point of the design. The directly measured SKIP/REPEAT marginal profiles of §2 serve as the design's interpretive anchor—they are what makes reading the $a_i(q)$, $b_j(q)$ terms as per-question marginals meaningful—but they are *not* the estimator's input: in Eq. (5) below, $(\widehat{\alpha\eta})_{iq}$ and $(\widehat{\beta\eta})_{jq}$ are estimated as SUB-grid row and column means, not read off the SKIP/REPEAT arms. The interaction γ_{ij} is pooled over questions; δ_{ijq} is the per-question residual interaction, the quantity of interest.

Stage 1: moment equations for the components. With $\mathbb{E}_{\cdot}$ denoting the mean over the indicated index set (over finite cells only), the standard ANOVA-mean-difference moments are

$$\hat{\mu} = \mathbb{E}_{ijq}[G], \qquad \hat{\alpha}_i = \mathbb{E}_{jq}[G] - \hat{\mu}, \tag{5}$$

$$\hat{\beta}_j = \mathbb{E}_{iq}[G] - \hat{\mu}, \qquad \hat{\eta}_q = \mathbb{E}_{ij}[G] - \hat{\mu}, \tag{6}$$

$$\hat{\gamma}_{ij} = \mathbb{E}_{q}[G] - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j, \qquad (\widehat{\alpha\eta})_{iq} = \mathbb{E}_{j}[G] - \hat{\mu} - \hat{\alpha}_i - \hat{\eta}_q, \tag{7}$$

$$(\widehat{\beta\eta})_{jq} = \mathbb{E}_{i}[G] - \hat{\mu} - \hat{\beta}_j - \hat{\eta}_q, \tag{8}$$

with missing cells dropped before each marginal average and set to zero in the two-way quantities $\hat{\gamma}$ (their count is negligible and they carry no information). The $\hat{\gamma}_{ij}$ of Eq. (5) is an ANOVA row/column residual fitted from the SUB grid. Figure 6 instead subtracts a prediction made from the measured SKIP and REP success rates, $\hat{a}_i^{\mathrm{SKIP}} + \hat{b}_j^{\mathrm{REP}} - \bar{M}$. These estimators are different: the latter descriptive residual need not be centered (its diff-5 mean is approximately -0.2041). The ANOVA moment interpretation and its centering constraints therefore do not transfer to that figure; neither residual by itself identifies a latent mechanism.

Stage 2: the two variances. The residual is the quantity of interest:

$$\hat{e}_{ijq} = G_{ijq} - \big[\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j + \hat{\eta}_q + \hat{\gamma}_{ij} + (\widehat{\alpha\eta})_{iq} + (\widehat{\beta\eta})_{jq}\big], \qquad \hat{\sigma}_e^2 = \mathbb{E}_{ijq}[\hat{e}_{ijq}^2], \qquad \hat{\sigma}_\gamma^2 = \mathbb{E}_{ij}[\hat{\gamma}_{ij}^2]. \tag{9}$$

$\hat{\sigma}_e^2$ is the residual variance after removing the estimated marginals and the layer-pair interaction on the rate grid; $\hat{\sigma}_\gamma^2$ is the variance of the estimated layer-pair interaction itself. Their ratio is the descriptive quantity assessed against the registered 0.10 threshold. Two caveats govern its reading. First, the reported cross-fit column does *not* remove each question's own marginals: as the cross-fitting paragraph below explains, those terms are replaced by cross-half averages (essentially their zero means), so the cross-fitted residual pools per-question marginal structure with any genuine three-way interaction. A faithful simulator with per-question marginals and *zero* interaction reproduces a passing Split B gate, so the ratio is a measured, descriptive quantity—not an identified interaction component and not a bound on how much of the headroom a fixed table could represent (Appendix D.3). Second, by Appendix D.3 Eq. (3), the residual also carries a σ_ε^2/n sampling floor of the same order as the measurement itself (≤ 0.06 against the reported $\hat{\sigma}_e^2 \approx 0.04$–$0.06$). We do *not* subtract it, for the two reasons given in Appendix D.3: the floor is the variance of the n-condition average (a single-condition router faces it un-reduced), and Split B shows the residual is not pure floor, since $\hat{e}$ reproduces across disjoint condition halves (r_B=0.53–0.64) which sampling noise would not. So $\hat{\sigma}_e^2$ is reported raw as a measured residual variance of unspecified composition; this is a descriptive statement, not a bound on an identified interaction or on fixed-table representability. Subtracting the floor raises the ratio when its positive numerator is held fixed; the floor-corrected sensitivity values are reported in Table 14.

Cross-fitting. The two question×margin terms make $\hat{\sigma}_e^2 > 0$ tautological unless the fit is prevented from using a question's own marginals to predict its own substitutions. We therefore split the questions once, uniformly at random and seed-fixed (seed 0), $Q_1, Q_2 \subseteq Q$, $Q_1 \cap Q_2 = \emptyset$, with $|Q_1| = |Q_2| = \lfloor n_q/2 \rfloor$, and compute the residual of a question in Q_1 against the per-layer offsets $(\widehat{\alpha\eta})$, $(\widehat{\beta\eta})$ estimated from Q_2 only, and conversely:

$$\begin{aligned} \hat{e}^{\mathrm{cf}}_{ijq} &= G_{ijq} - \big[\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j + \hat{\eta}_q + \hat{\gamma}_{ij} + \bar{a}_i^{(\bar{q})} + \bar{b}_j^{(\bar{q})}\big], \\ \bar{a}_i^{(\bar{q})} &= \mathbb{E}_{q' \in Q_{\bar{q}}}[(\widehat{\alpha\eta})_{iq'}], \quad q \in Q_q, \;\; \bar{q} = \text{the other half}. \end{aligned} \tag{10}$$

$\hat{\eta}_q$ is *not* cross-fitted (difficulty is a property of the question; there is no other-half estimate of it), and $\hat{\gamma}$ is estimated on the full grid (it is the object of Split A below, not of the residual). The cross-fit and in-sample columns differ on two distinct axes. The residual $\hat{\sigma}_e^2$ is genuinely out-of-fold in the cross-fit column—the per-layer offsets no longer adapt to the held-out questions—and on Qwen2.5 diff-5 it rises $0.0424 \to 0.0583$ ($1.37\times$). The numerator $\hat{\sigma}_\gamma^2$, in contrast, is *not* refit under cross-fitting ($\hat{\gamma}$ stays on the full grid in both estimators); its apparent cross-fit/in-sample spread $0.0005 \to 0.0009$ ($1.71\times$) is a difference of rate definition between the two columns—the cross-fit estimator collapses the four prompt conditions by the *mean* validity rate while the in-sample (V7) estimator uses the *robust4 AND* (binary) grid—not an effect of the out-of-fold replacement (verified numerically to four decimals in Appendix E). This joint comparison does not isolate the effect of in-sample fitting.

Cross-fitting is the historical pre-registered estimator (Appendix D.5) and the in-sample column of Table 14 is a joint response-grid/estimator sensitivity, not an isolated cross-fitting comparison.

Estimator conventions. For odd n_q, the two equal-sized subsets do not partition all questions; the remaining question retains zero cross-fitted margin offsets and is still included in the residual array, and the same implementation keeps $\hat{\gamma}$ fixed across the margin-offset replacement. Consistent with the cross-fitting paragraph, the retained implementation shows $\hat{\sigma}^2_\gamma$ unchanged by the out-of-fold replacement; the $1.71\times$ cross-fit/in-sample numerator spread in Table 14 is the rate-definition difference (4-noise mean vs robust4 AND), not the replacement. Reconstructing the two rate grids matches the retained $\hat{\sigma}^2_\gamma$ values to four decimals. The implementation entrypoints and model-specific estimator records are in the detailed companion, Section C.5.5 (*The variance-component estimator, derived*).

Split A (reproducibility of γ). With the question split of the cross-fitting held fixed, compute $\hat{\gamma}^{(A_1)}$ and $\hat{\gamma}^{(A_2)}$ on the two halves (Eq. (5) applied to each half), and set

$$r_A = \mathrm{corr}\big(\hat{\gamma}^{(A_1)}_{ij}, \hat{\gamma}^{(A_2)}_{ij}\big) \text{ over the eligible grid,} \tag{11}$$

averaged over 100 random question splits. Because each $\hat{\gamma}_{ij}$ averages ≈566 question-cells, r_A measures the *precision* with which γ is estimated, not its *magnitude*: any systematic γ drives r_A toward 1 at this cell count regardless of its variance—the deviation-1 argument of Appendix E.5. $r_A = 0.841$ (Qwen2.5, 0.84–0.90 across strata) and 0.885 (Qwen3) therefore say that γ is reproducibly nonzero and precisely estimated, which is exactly the reading that does *not* bear on the gate; the gate-bearing object is $\hat{\sigma}^2_\gamma/\hat{\sigma}^2_e$ (Eq. (9)), which is small.

Split B (reproducibility of measured residual e). With the same question halves, split the *condition* axis as $H_1 = \{\text{0-shot, 1-shot}\}$, $H_2 = \{\text{2-shot, rewrite}\}$, re-derive the rate grid on each half, run the full Stage-1–2 pipeline (including cross-fit) on each to obtain $\hat{e}^{H_1}$, $\hat{e}^{H_2}$ over all (i, j, q) triples, and set

$$r_B = \mathrm{corr}\big(\hat{e}^{H_1}_{ijq}, \hat{e}^{H_2}_{ijq}\big) \text{ over all triples,} \qquad \text{PASS if } r_B \geq 0.05 \text{ and } r_B > \mathrm{null}_{p99}, \tag{12}$$

with the null obtained by permuting question labels against fixed cell structure: independently permuting the question axis of G^{H_1} and G^{H_2} and re-running the residualization. The permutation preserves both the cell structure (which (i, j) cells exist) and the question marginal structure (each question's own rate profile), and breaks only the question×cell association—the association $\hat{e}$ is meant to capture—so the null is the distribution of r_B when no question-specific interaction exists. Observed $r_B = 0.530$ (Qwen2.5) and 0.637 (Qwen3) sit far above their permutation null p_{99} (0.004 for Qwen2.5, 0.005 for Qwen3); the gate passes at all measured cells.

Why Split B is a reproducibility check, not an identification test. The two splits answer different questions because of what each $\hat{\gamma}$ and each $\hat{e}$ is built from. $\hat{\gamma}_{ij}$ pools ≈566 question-cells per entry, so its sampling noise is small, its two half-estimates agree if γ is merely nonzero, and r_A is saturated above the preregistered 0.15 threshold for any model with any nonzero interaction—the Deviation-1 failure. $\hat{e}_{ijq}$ is a *single-question* cell residual built from four draws; its two condition-half estimates reproduce a question-specific *residual* across prompt families. This reproducibility is genuine, but it is *not* evidence for a separable δ_{ijq}: because the cross-fitted residual still pools each question's marginal terms, a faithful simulator with per-question marginals and no interaction also reproduces a passing Split B gate (a synthetic false positive). So Split B shows a question-specific residual is reproducible across prompt conditions—the reading §4.2 adopts—rather than identifying a three-way interaction; Split A is retained as a precision report and as the source of the deviation disclosed in Appendix E.5.

D.5 Estimator sensitivity and registered checks

The mean-response/cross-fit and robust4-response/in-sample columns change both the input response grid and the estimator. Their numerical differences cannot be attributed to cross-fitting alone. All 18 reported variance ratios are below 0.10, but the residual contains question-specific marginal terms, interaction, and noise. The Bernoulli-floor subtraction assumes a sampling model for fixed prompt templates. A zero-three-way-interaction construction can still pass Split B, since

Table 14: Historical sensitivity of descriptive variance ratios.

Model	Stratum	Mean / cross-fit	In-sample variant	Floor sensitivity
Qwen2.5-3B-Instruct	diff-1	0.0403	0.0397	0.0547
Qwen2.5-3B-Instruct	diff-2†	0.0301	0.0283	0.0425
Qwen2.5-3B-Instruct	diff-3	0.0264	0.0237	0.0374
Qwen2.5-3B-Instruct	diff-4†	0.0224	0.0190	0.0322
Qwen2.5-3B-Instruct	diff-5	**0.0158**	0.0127	0.0237
Qwen3-4B	diff-5	**0.0219**	0.0530	0.0290

Notes. Historical descriptive variance-ratio sensitivities $\hat{\sigma}^2_\gamma/\hat{\sigma}^2_e$, retained with all reported numerical cells. The mean-grid/cross-fit and in-sample columns differ in response-grid and estimator choices, so their difference is not an isolated cross-fitting effect. The residual e pools question-specific margins, interaction, and condition noise; subtracting a Bernoulli floor is an assumption-dependent sensitivity, not identification of a pure interaction. All 18 ratios are below 0.10 (maximum 0.055). The dagger marks the historical diff-2/4 confirmatory protocol (Appendix D.5); remaining cells are exploratory. These ratios do not explain or bound routing accuracy.

Table 15: Variance decomposition and its three diagnostic checks.

(a) Split-half diagnostics and gates Quantity	Value (range over strata)	Pre-registered gate
r_A (Split A, persistent γ)	0.84–0.90	< 0.15 **(misspec.; §E.5)**
$\hat{\sigma}^2_\gamma/\hat{\sigma}^2_e$	< 0.10 throughout (Table 14)	< 0.10 ✓
r_B (Split B, question-specific residual e)	0.53–0.61	≥ 0.05 ✓
null p_{99}	≈ 0.004	—
(b) Response-grid and estimator sensitivity, difficulty 5	mean grid / cross-fit	robust4 grid / in-sample
$\hat{\sigma}^2_\gamma$	0.000921	0.000539
$\hat{\sigma}^2_e$	0.0583	0.0424
$\hat{\sigma}^2_\gamma/\hat{\sigma}^2_e$	0.0158	0.0127
(c) Behavioural consistency, gate ±10pp — ***failed*** Stratum	naive $1-(\mathbb{E}[p])^4$	corrected $1-\mathbb{E}[p^4]$ / measured (resid.)
diff-1	0.508	0.468 / 0.188 (+28.0pp, **miss**)
diff-3	0.599	0.542 / 0.321 (+22.1pp, **miss**)
diff-5	0.653	0.606 / 0.424 (+18.3pp, **miss**)

Notes. The variance decomposition (Qwen2.5-3B-Instruct) and its three checks, in one place. **(a)** Split-half diagnostics with their pre-registered gates; the r_B null permutes question labels of the substitution grid against fixed question marginals, preserving both cell structure and question marginal structure and breaking only the question×cell association. With $\approx 6.7 \times 10^5$ triples per stratum, the analysis emphasizes effect size: the pre-registered gate is the effect-size floor $r_B \geq 0.05$, not a p-value. **(b)** Historical columns change both the response grid and the residual estimator (difficulty 5); their differences are not attributable to cross-fitting alone. The measured residual e pools question-specific marginals and any remaining interaction/noise. **(c)** Behavioural-consistency check (pre-registered, **failed**): predicted TT→ × rate against the direct measurement; the order correction moves every stratum in the right direction by 4–6pp and no stratum reaches the ±10pp gate.

that residual retains question-specific margins. Difficulty 2/4 entries follow the recorded confirmatory protocol; the output rescoring and width/supervision controls are exploratory.

The registered TT-failure prediction misses the observed rates at all three strata. Correcting the order of expectation reduces the errors but leaves residuals of 28.0, 22.1, and 18.3pp, outside the ±10pp gate (Table 15). The original Split-A threshold $r_A < 0.15$ also fails; reproducibility of a small component does not measure its magnitude. The null constructions and branch records are detailed in the detailed companion, Sections C.4.2–C.4.3 (composition nulls and the failed behavioural check) and Appendix I (*Pre-registration and Disclosed Deviations*); Section C.5.3 summarizes the separate confirmatory protocol on difficulty 2/4.

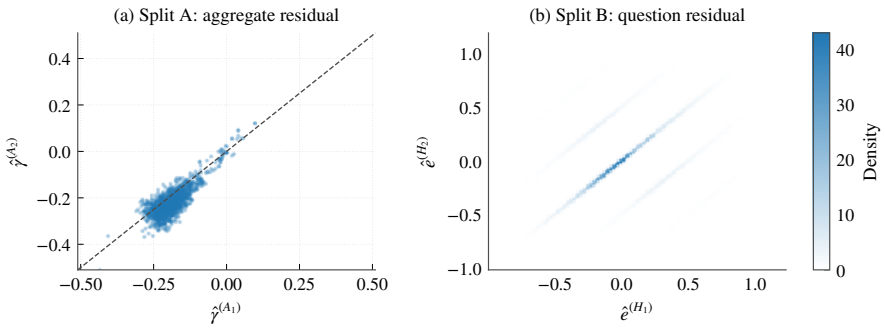

Figure 7: Reproducibility is not magnitude. Left: Split A, questions split 50/50 (one fixed split, seed 0), $\hat{\gamma}^{(A_1)}$ versus $\hat{\gamma}^{(A_2)}$ across 1,190 eligible cells ($r_A = 0.822$). Table 15 reports results over split resamples. Right: Split B, $\hat{e}^{(H_1)}$ versus $\hat{e}^{(H_2)}$ over question–cell observations, with $H_1 = \{0\text{-shot}, 1\text{-shot}\}$ and $H_2 = \{2\text{-shot}, \text{rewrite}\}$; shading shows binned density. The residual arrays and fixed split follow the archived plotting estimator; the historical $\hat{\delta}$ label is updated to the descriptive $\hat{e}$. The correlation-null threshold is reported in Table 15, not on the residual-value axis. These panels describe reproducibility, not an identified pure three-way interaction (§4.2).

E Reproduction, Generalization, and Registration

E.1 Data and generation

DART-Math supplies the five level labels and ground-truth answers. The evaluation uses the first 565 retained questions per level, the common minimum across the available per-stratum grids; this is a comparability truncation, not a random population sample. The source problem pool includes public MATH/GSM8K material. Phrase overlap documents reuse of public questions but does not identify overlap with a model's unavailable pretraining corpus. The separate MMLU analysis (Hendrycks et al., 2021) uses its 14,042-question test set and 57 subjects. Models and datasets remain subject to their original license terms.

Generation and scoring configuration. The evaluation is deterministic greedy decoding of the base instruction-tuned model over the CoA path-edited checkpoints (Appendix C.1 defines the features; this is the decoder). *Decoding.* Greedy decoding uses temperature zero, one response per prompt, the default nucleus threshold of 1.0, and the default random seed of zero. Generation is capped at 2,048 tokens without an additional early-stop rule; the maximum sequence length is 8,192 tokens. Serving uses the vLLM V1 engine (v0.24.0 for the Qwen2.5 grid, v0.21.0 for Qwen3), one tensor-parallel replica per model in its own subprocess, bfloat16 numerical precision and an accelerator-memory allocation fraction of 0.8, with prefix caching and chunked prefill enabled; per batch, 64 questions $\times$ 4 conditions $=$ 256 prompts are generated in one call. *The chain-of-thought template* (0-shot condition), verbatim, invoked as the user turn under the chat template (system: "You are a helpful assistant."), with an assistant-generation prefix appended:

```
Solve the following math problem step by step.  Put your
final answer in \boxed{ANSWER}.
### Problem Start
{q}
### Problem End
```

1- and 2-shot exemplars. Exemplars are correct chain-of-thought solutions generated ahead of time (greedy, by the same base model, random seed 42, with at most 1,024 generated tokens for diff-1/2/3 and 2048 for diff-4/5), 10 per stratum, drawn from a stratum-matched pool restricted to *correct* solutions; the 1-/2-shot conditions prepend one/two of them in template 0, de-duplicated against the target question. The exemplar set is fixed once per stratum and held *identical across all* 1,261 *paths* of that stratum, so it cannot create an across-path difference. *The rewrite condition. rewrite* is a *fixed, hand-written paraphrase template*—template 1 of the same family ("Work through the math problem below step by step. Place your final answer in \boxed{ANSWER}…")—with zero few-

shot exemplars. It is *not* produced by a model, is not randomized, and *does not alter the question text or any mathematical content*: only the surface template framing changes. This defines a surface-template transfer axis, not independent resampling of a latent question. Residual and floor-correction claims remain conditional on Appendix D.3's assumptions; surface paraphrases need not produce independent errors. *Scoring.* Correctness is a deterministic extraction-and-equivalence judgement using the original DART-Math evaluator, not a model judge: the boxed final answer is extracted strictly, without falling back to the last number in the response, normalized, and compared to the ground truth by literal string, by numeric closeness (relative tolerance 10^{-9}, absolute 10^{-8}), or by symbolic equivalence (SymPy, with an antlr4 parser for LaTeX formulas); percentage forms and OlympiadBench originals are handled by the same pipeline. A *parse failure* is a response from which no comparable boxed answer can be extracted; these are counted and reported (§C.3) rather than silently admitted.

E.2 Execution costs and available artifacts

SUB preserves the number of layer executions per token. A repeated layer still needs the appropriate execution-instance cache state; changing the schedule can require prefill and cache recomputation. Computing layer-18 features adds prefill work. The classifier uses three discovery-prefill observations, the structural voter uses 57–58 full generated answers, and the same-prompt detector uses KEEP plus five probes. The mixed rewrite detector additionally needs the selected rewrite output. These are part of the system cost even if a table scores one final answer.

The SC logs record batch generation times of 1,321, 1,025, and 3,155 seconds for Qwen2.5 difficulty 5, Qwen2.5 difficulty 3, and Qwen3 difficulty 5, with output-token totals 25,197,288, 18,573,365, and 43,696,170. These are full-run records, not per-request latency. Device labels and the 68 GiB allocation estimate are generator-entered metadata rather than independent hardware or peak-memory measurements. Structural timing and token logs are unavailable.

The retained analysis archive includes correctness matrices, prefill features, source snapshots, selected frozen path decisions, analysis-product summaries, and raw SC responses. Figure generation records input hashes. Rescoring a frozen decision, replaying cached answers, refitting a selector, and regenerating model responses require different artifacts. For the original width and supervision controls, complete fitted selector parameters and per-question candidate/decision ledgers are unavailable. Merged raw detector/structural responses and complete structural cost logs are also unavailable in the retained archive. These are limitations of the original runs, distinct from the new offline experiments, which save fitted probe parameters, rankings, shortlists, and answer ledgers (companion Appendix B.5). The restored Qwen2.5 answer caches support conditional voting replay, subject to the provenance and checker limitations in companion Appendix B.4. The SC replay preserves recorded parsing-error/timeout handling and row alignment (Appendix B.2). Further artifact and execution notes are in the detailed companion, Appendix D (*Engineering Notes and Reproducibility*) and Section B.5 (*Artifact map and reproduction limits*).

E.3 Accounting invariants

An all-KEEP decision must have zero net; an oracle must agree with its stated candidate set and correctness constraint; refund recovery and no-refund net must equal rec and rec$-$dst. KEEP encodings (sentinel, negative index, or canonical index zero) must produce identical scores. An executed C-correct answer earns credit even if D is false. Eligibility using target discovery correctness must be exposed independently of the score. These checks are applied before interpreting the frozen output counts. A discarded convention that scored an emitted KEEP as an incorrect edit violates the all-KEEP invariant. Scoring and interval corrections are documented in the detailed companion, Appendices J (*Verification Protocol and Diagnosed Defects*) and K (*Consolidated Correction Log*), with source data retained.

E.4 Model and task scope

Qwen3-4B difficulty-5 C-scored nets are -46 for factorized single-path selection, $+5$ for its global-cover conditional classifier, and -5 for structural voting; their fixed-output intervals are $[-63, -29]$, $[-1, 11]$, and $[-20, 10]$. The full-pool C-only/constrained oracle recoveries are $155/85$, and the

Table 16: Second-order composition in a mixed edit pool.

Stratum	Raw r ($N = 314$)	Logit R^2	Logit N
diff-1	0.465	0.090	306
diff-3	0.817	0.521	274
diff-5	0.970	0.856	251

Notes. Second-order composition in a mixed pool: double skip, double repeat, skip–repeat, substitution pairs, and randomized/cell-compound variants. Raw $r = \mathrm{corr}(M_{\text{obs}}, M_{\text{recon}})$ uses all 314 paths per stratum. M_{obs} is the fraction of 300 questions correct in both 0-shot and rewrite; M_{recon} is the logit-additive prediction from measured marginal rates. The separate logit regression excludes paths with $M_{\text{obs}} < 0.02$, $M_{\text{obs}} > 0.98$, or a 95% Wilson interval wider than 0.2; its retained N and R^2 are shown separately. Both summaries show stronger association at harder strata in this mixed pool; neither establishes calibrated prediction or a law for substitution pairs alone. An attenuation-corrected correlation is omitted because reconstruction-side reliability was not independently measured.

global R-rate top-57 constrained oracle recovers 55. The comparison covers two related Qwen instruct models and does not isolate an architectural or generational cause.

The separate MMLU grid reports a cross-fit residual ratio 0.0049, a prompt-transfer oracle gap 34.4pp, and FF replication of 0.486 on the held-out 0-shot arm, with 4.2-fold enrichment. Discovery uses a 5-shot arm; there is no rewrite condition. These are descriptive task-transfer observations. Task-fixed routing and second-order composition use different candidate spaces and are retained, with their full tables and estimator qualifications, in the detailed companion, Sections C.6.3 (*Between-task variance and a task-fixed router*) and E.1 (*Scope: a second model generation, second-order composition, static proxies*). They are not additional deployment replications of the math selectors.

E.5 Registration and deviations

The original math registration was frozen on 2026-08-08; the MMLU record followed on 2026-08-23. An English presentation with condensed connective prose, historical source-hash prefixes, and disclosed deviations is retained in the detailed companion, Appendix I (*Pre-registration and Disclosed Deviations*). Companion Section C.5.3 separately summarizes the later confirmatory protocol and gives its full hash; that protocol's full text remains in the retained archive. These nonpublic records and their hashes do not independently establish registration timing; the evidence assessed here is the protocol and results reported in the paper and companion. Three outcomes are material to the interpretation here. The Split-A correlation threshold failed and was subsequently replaced as the primary summary by a co-registered variance-ratio criterion, a disclosed post-observation change. The decoupled TT-failure null also failed: observed failure was below its 99th percentile, indicating enrichment of success rather than the registered failure claim. The behavioral-consistency prediction remained outside its ± 10pp gate after correcting the order of expectation. The constructive-synergy branch passed its recorded Case-1 criterion. Output rescoring, candidate-width, supervision, and self-consistency analyses are exploratory follow-ups and are not retroactively preregistered.

F Comparison with Related Routing Methods

Dr.LLM, published at ICLR 2026, learns skip/execute/repeat decisions and grades the generated routed answer. CoLa uses per-input MCTS path search; its label-guided search is an existence/search comparison. PoLar predicts contiguous segment programs from an external question representation; its single-decode and pass@5 results have different generation budgets. These protocols differ from selection over a frozen, finite single-edit pool. The audited descriptions use no post-hoc refund of failed edits with a correct default answer. We therefore make no claim that our accounting findings invalidate their reported gains. We have not reproduced the original Dr.LLM or PoLar systems. Their decision information also differs from our target-label-assisted conditional classifier. The new offline selector is label-free but retains our factorized probe family. Coverage/output distinctions are established in broader test-time scaling evaluation (Hariri et al., 2026); adaptive verification on shared pools (Lin et al., 2026) is a relevant untested alternative to our plurality vote. Detailed

Table 17: Evaluated constructions and evidence boundaries.

Object	Target correctness labels	Interpretation
Full and shortlist oracles	Evaluation, plus discovery when constrained	Existence coverage
Factorized single-path	None	One emitted C-scored answer
Structural voter / SC	None	Post-generation C-scored output
Conditional classifier and controls	Discovery labels	Conditional transfer to C
Training-fold fixed path	None at prediction	Held-out C-scored output
Original global fixed path	Discovery and evaluation labels	In-sample selected reference
Latent-score simulations	True validity and baseline correctness	Specified score-model behavior

Notes. Interpretation is independent of whether one or many answers are generated.

protocol comparisons are in the detailed companion, Appendix F (*Protocol Audit of the Closest Prior Work*); Appendix G (*Extended Related Work*) places them in the broader literature.

Looped computation at fixed and adapted checkpoints. Recurrent architectures and theoretical studies motivate depth reuse for reasoning (Dehghani et al., 2019; Saunshi et al., 2025). Recent methods distinguish the recurrence update from the choice of execution depth. Training-Free Looped Transformers (Chen et al., 2026) applies recurrence to frozen checkpoints, using damped residual updates motivated by numerical integration. Its block iteration and state mixing are outside our ordinary single-layer REPEAT and skip-and-repeat SUB family. Even without weight training, a different recurrence operator can yield different answers. LoopUS (Park et al., 2026) converts pretrained models into encoder–recurrent-block–decoder architectures through post-training, with selective gates, random depth supervision, and confidence-based exit. These procedures adapt the computation available to a decision rule. Our grid holds the checkpoints fixed and measures repetition as one of the two constituents of SUB, enabling comparison of the composite with its parts. It neither implements these methods nor tests their reported gains.

Selective iteration and decision-time information. TaH (Fu et al., 2026) studies how additional iterations can repair wrong token predictions or corrupt correct ones. It adapts the backbone before fitting a token-level iteration decider. RecurTrace (Wang et al., 2026) freezes the base model, trains added recurrence and memory components, and distills a halting head from depth-dependent loss improvements. Its decisions read the evolving loop state. Both use oracle-derived training supervision; this is distinct from consulting the evaluation question's correctness labels at inference. Their learned policies therefore must not be grouped with our label-assisted oracle references or conditional-transfer diagnostics. LoopFormer (Jeddi et al., 2026) aligns trajectories of different lengths for budget-conditioned computation, while T-LoopFormer (Yu et al., 2026) combines token-level depth routing with recursion-specific KV caches. These methods select computation using different states, action spaces, and training procedures from our question-level path rankers and answer voters. Our findings are limited to the evaluated selectors; they provide no overall performance assessment of looped models. Loop count, emitted-token count, and per-token layer executions also remain distinct from measured end-to-end cost.

Mechanistic context. Kapl et al. (2026) compare depth-grown and looped models and report related signatures of iterative computation. Blayney et al. (2026) examine cyclic latent-state trajectories and layer-specific fixed points in the recurrent models they study. These analyses motivate questions about internal dynamics. Our constituent and substitution outcomes are behavioral measurements: a substitution's success despite failed constituents does not identify a recurrent fixed point, an iterative algorithm, or the mechanism responsible for that success.

G DETAILED MEASUREMENT AND STATISTICAL RECORDS

The main text retains the comparisons needed to interpret the findings. This section records the formal accounting and complete numerical displays.

Table 18: Open comparisons and extensions.

Comparison	Additional evidence needed
Independent label-free confirmation	New questions under a frozen protocol; current offline decisions reuse inspected questions
Stronger router or answer verifier	Original-method reproduction or shared-pool selector with measured cost
Question-clustered FF confirmation	Question-level resampling and independent event replication
Equal end-to-end inference cost	Structural and sampling token/runtime measurements, with all feature costs
Supervision or representation effect	Matched decision information and paired output comparisons
Cross-family deployment replication	Additional base checkpoints and independent routing evaluation
Model adaptation or distillation	Training and generation budgets, final-output evaluation

Notes. These are future controls, not completed evaluations.

G.1 Formal output accounting and oracle definitions

Let Y_{nqp} be correctness for prompt condition n, question q, and path p. Discovery validity is $D_{qp} = \bigwedge_{n\in\{0,1,2\text{-shot}\}} Y_{nqp}$, evaluation correctness is $C_{qp} = Y_{\text{rewrite},qp}$, and four-condition validity is $R_{qp} = D_{qp} \wedge C_{qp}$.

Definition 1.

An output evaluation specifies candidate sets S_q, a decision rule g, an answer selector or aggregator s, their access to target labels, and the number of generated outputs. If they emit answer a_q, its question contribution is $X_q = C(a_q) - C_{q,\text{KEEP}}$. Recovered (rec) and destroyed (dst) questions have $X_q = +1$ and $X_q = -1$, respectively. The net and accuracy change are

$$N(S,g,s) = \sum_{q=1}^{565} X_q = \text{rec} - \text{dst}, \qquad \Delta\text{Acc} = \frac{N(S,g,s)}{565}. \tag{13}$$

Candidate identity and label access are part of the evaluation; a shared cardinality $|S_q| = K$ specifies only the shortlist size.

Oracle references. For a stated candidate set S_q, define

$$O_C(q;S_q) = \max_{p\in S_q\cup\{\text{KEEP}\}} C_{qp}, \tag{14}$$

$$O_{D\wedge C}(q;S_q) = C_{q,\text{KEEP}} \vee \mathbf{1}[\exists p \in S_q : D_{qp} \wedge C_{qp}]. \tag{15}$$

Both preserve a correct KEEP answer. The first measures evaluation-prompt coverage; the second restricts recovery to discovery-valid paths. Both inspect evaluation labels to establish existence. A decision fixed from the target question's discovery correctness and then scored on C is a *conditional-transfer diagnostic*. A label-free decision uses features or generated answers without consulting that target's correctness labels.

Output scoring and post-hoc refund. Writing $A_q = C_{q,\text{KEEP}}$ and $V_q = C(a_q)$, the two accounting rules are

$$s_q^{\text{refund}} = \max(A_q, V_q), \tag{16}$$

$$s_q^{\text{output}} = V_q \quad (\text{with } V_q = A_q \text{ when KEEP is emitted}). \tag{17}$$

Our primary metric uses output scoring. Refund scoring removes all destruction after correctness is known: for the new difficulty-5 direct voter, 60 recoveries and 57 destructions give net 3, whereas refunding gives 60. Appendix A defines normalized refund and no-refund gains with an explicit scoring criterion and reference. A no-refund score alone does not establish that the decision can be made without target labels.

G.2 Coverage counts and intervals

Structural-only coverage U counts questions with at least one correct edit and no correct SC sample. Table 19 gives all counts used in Figure 1a. The full edit pool and SC have different sizes; no matched-budget gain follows from these counts.

Table 19: Candidate coverage and selected answers beyond SC@57.

Model / stratum	SC vote	SC coverage	SC failure	U [95% CI]	In top 57	Selected
Qwen2.5 / diff-3	519	548	29	15 [8,23]	3/15	0/15
Qwen2.5 / diff-5	368	505	137	42 [30,54]	15/42	3/42
Qwen3 / diff-5	370	451	81	56 [42,70]	16/56	0/56

Notes. $N = 565$ per row. SC failure: correct candidate, incorrect vote. U: uncovered by 57 SC samples, covered by 1,260 edits; brackets: 95% question-bootstrap intervals. In top 57: U questions with a nonempty matrix-correct shortlist answer; Selected: those the new voter answers correctly. Original-checker scores and unequal-pool oracle comparisons, not matched-cost gains.

The paired direct-vote minus SC differences in Figure 1b are -26 $[-37, -15]$, -57 $[-76, -38]$, and -23 $[-35, -11]$ questions, in table order. These intervals condition on the saved fits and outputs.

G.3 Historical shortlist and conditional-transfer diagnostics

The main-text selectors use the current decision rules in §2. Earlier factorized single-path selection and structural voting are archived as routeA and routeC; target-discovery-assisted classification is routeB. The original Qwen2.5 single-path implementation mapped a winning KEEP to edit index 1. Its voter forced KEEP into the ballot and could return a different shortlisted path when KEEP's answer won. The current fits retain a winning KEEP action and return the vote winner directly. They are distinct protocols, not a relabeling of the historical decisions. Paired KEEP-mapping checks apply to the current predictions, not unmatched historical action files (companion Appendix B.3). The closest audited routing reports use no post-hoc KEEP refund; these scoring corrections concern our own evaluated constructions.

At difficulty 5, the full C-only oracle recovers 238 KEEP errors; the original voter's shortlist oracle recovers 196. Their 42-question difference is lost coverage; the $196 - 12 = 184$ net-output gap includes missed recoveries and destruction. Neither is the 42-question SC-relative increment or guaranteed deployable gain.

The conditional classifier's fold covers recover 157 errors under a C-only oracle, but its selected outputs net zero. The global discovery-ranked width scan reaches $+11$ at $K = 57$ ($[0, 22]$); removing the fitted gate at difficulty 3 increases net from $+2$ to $+13$ while retaining target-label eligibility. Supervision controls show no monotone gain with selector-training fraction. These conditional-transfer results do not establish label-free deployment or a cause of the new selectors' failures (Appendix B).

G.4 Complete prediction statistics and inspected workpoints

Table 20: Detection and ranking AUROCs with uncertainty estimates.

A. Post-generation detection: 0-shot KEEP correctness		
Signal	diff-3	diff-5
Structural-probe agreement ($K = 5$)	0.972 [0.960, 0.983]	0.929 [0.909, 0.947]
Prompt-condition agreement	0.911 [0.875, 0.944]	0.886 [0.857, 0.912]
Temperature agreement ($n = 5$)	0.433 [0.374, 0.495]	0.603 [0.554, 0.651]
Answer log-probability	0.471 [0.405, 0.537]	0.527 [0.477, 0.578]

Table 20 (continued): Detection and ranking AUROCs with uncertainty estimates.

B. Prefill ranking: R-validity at difficulty 5			
Predictor / training population	AUROC	Increment	Paired 95% CI
Cross-fitted structural prior	0.6765	—	—
Ranker, all questions	0.6653	−0.0112	[−0.0352, 0.0135]
Ranker, baseline errors only	0.6279	−0.0486	[−0.0800, −0.0183]

Notes. A: Detection AUROC with 95% question-bootstrap intervals ($B = 1000$), conditional on selected signals. Prompt agreement compares KEEP answers across 0/1/2-shot and rewrite with 0-shot; temperature sampling uses $T = 0.7$. B: Mean within-question R-validity AUROC; paired 95% intervals measure increments over the same cross-fitted prior ($B = 2000$). Neither measures emitted-answer accuracy.

At inspected 0-shot threshold $\tau = 0$, rec/dst are 17/5 and 31/6 at difficulties 3/5: nets $+12/+25$, with fixed-count intervals $[3, 21]/[14, 37]$. Difficulty-5 damage is $6/565 = 1.06$pp, exceeding the 1pp cap. Probe and threshold inspection makes these workpoints exploratory. Changed rewrite probes with a 0-shot gate give $-17/-15$ under a separate cross-condition procedure; Appendix B.2 specifies both procedures and costs.