BEYOND BENCHMARK ACCURACY: STRESS-TESTING REASONING CLAIMS IN LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models, and reasoning models in particular, score highly on mathematical, logical and coding benchmarks, and those scores are increasingly read as evidence of general reasoning ability. Accuracy on a fixed item set does not establish that the underlying competence survives controlled variation in problem depth, surface form or evaluation condition.

We stress-test reasoning claims along four dimensions — complexity generalization, perturbation robustness, process faithfulness and familiarity–novelty sensitivity — and add a cross-dimensional analysis of how the first two interact. Fourteen contemporary configurations over ten model checkpoints, including four reasoning-disabled counterparts of listed checkpoints, are evaluated on two generated task families with a depth ladder reaching $n = 12$, on five benchmark reference points scored as integer item counts, and under transformations separated by whether they preserve the correct answer.

Benchmark accuracy and stress behaviour order the roster similarly overall (Spearman $\rho = 0.92$) and come apart where a benchmark discriminates least: configurations within 5 percentage points of one another on MATH-500 differ by up to 23.8 points in their answer-preserving robustness gap, and the two orderings disagree on 8 of 91 configuration pairs. The incremental information is local to near-ties rather than a general decoupling, and it is smaller among distinct checkpoints than across all configurations. The perturbation penalty compounds with depth rather than adding to it — the fitted interaction is negative for every configuration, and for the strongest the gap widens from 0.8 points at $n = 3$ to 11.3 at $n = 12$. The direction holds under depth entered as a factor, as a spline and segmented at the changepoint, under matched input length, and under inference clustered on seed templates rather than items: 184 of 192 intervals lie below zero, the exceptions being the two strongest configurations under the widest specifications. Threshold-like degradation is identified for ten configurations with the changepoint tracking capability, right-censored for the two strongest and unidentifiable for the two smallest, so a single ladder cannot serve the whole capability range.

These are behavioural results about generalization under stated conditions, not findings about internal computation. Within that scope they indicate that benchmark accuracy and robust generalization should be reported as distinct empirical properties, and that a robustness estimate obtained at one difficulty level understates fragility at greater depth.

1 INTRODUCTION

Reasoning models couple reinforcement learning on chain-of-thought traces with inference-time compute scaling, and score highly on mathematical, logical and coding benchmarks (Wei et al., 2022; Kojima et al., 2022; OpenAI et al., 2024). Those scores are increasingly read as evidence of general reasoning ability.

Accuracy summarises behaviour on a fixed item set from one distribution of surface forms, difficulty levels and template families. Several accounts predict the same score and different behaviour

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.

once conditions change: the task was learned robustly; the evaluation distribution overlaps training material; items admit shortcut features correlated with the answer; or a verbalized trace is plausible without being causally implicated in the decision (Geirhos et al., 2020; Pacchiardi et al., 2024; Ye & Durrett, 2022). Static benchmarks also saturate and are reused across model generations (Wang et al., 2025; Shi et al., 2024), and a single aggregate discards the shape of the accuracy–complexity relation (Shojaee et al., 2025). The saturation cost is measurable here: the four frontier models span 2.6 points on MATH-500 against 12.1 on the contamination-resistant FrontierMath, so the conventional benchmark separates them least where they are strongest.

Each concern has been probed separately, in perturbation, complexity, process and novelty studies reviewed in Section 2. The combination is what remains unmeasured: perturbation studies hold difficulty fixed and complexity studies do not perturb, so the interaction between them is unreported — yet it governs whether a robustness result transfers. If a surface change costs a model 0.8 points at $n = 3$ and 11.3 at $n = 12$, a gap measured at one difficulty level understates fragility at another.

We supply both under one protocol, organised into four stress dimensions — complexity generalization, perturbation robustness, process faithfulness, familiarity–novelty sensitivity — plus a complexity $\times$ perturbation analysis, which is a statement about two of those dimensions jointly rather than a fifth. Two choices are load-bearing: transformations are classified by whether they preserve the correct answer, since the two classes cannot share a consistency metric, and the depth ladder reaches $n = 12$ on two families with no shared surface form, which makes changepoints estimable for mid-capability models rather than censored.

Our contributions are: (i) a four-dimension protocol whose taxonomy separates answer-preserving transformations from structure-preserving but answer-changing ones, each with its own metrics (Section 3); (ii) an explicit model of the complexity $\times$ perturbation interaction fitted on the logit scale, so the accuracy-floor bound does not distort the estimate, and re-estimated under six alternative specifications of depth, against a length-matched control, and with inference clustered on the seed template rather than the item (Section 5.4); (iii) a changepoint analysis reporting *why* a threshold is or is not identified, separating right-censoring from floor saturation (Section 5.2); and (iv) per-configuration profiles across all four dimensions, with benchmark accuracy reported as integer item counts and a precision analysis derived from the paired-discordance relation for unequal marginals (Appendix S).

The scope is fixed here rather than deferred. Results concern mathematical and symbolic reasoning on the described families and benchmarks, not planning, spatial, grounded or tool-using settings. Degradation under stress is a behavioural finding about generalization under controlled variation and does not license inferences about internal computation; sensitivity to familiarity is likewise behavioural, and training-data exposure is not observable to us. The aim is not to adjudicate whether these models reason, but to make the evidential standard for such claims explicit and testable.

2 Related Work

Surface sensitivity in reasoning evaluation is established: templated regeneration moves accuracy, implicating form rather than procedure (Mirzadeh et al., 2025), and subsequent work separates perturbations that preserve a solution path from those that invalidate it (Huang et al., 2025), contrasts surface with structural variation (Hao et al., 2025), builds contamination-resistant variants (Hou et al., 2026), and scales perturbation severity by item difficulty (Li et al., 2025; Sun et al., 2025b). Controlled-depth studies establish that accuracy falls as required steps grow (Shojaee et al., 2025; Sun et al., 2025a; Rameshkumar et al., 2025; Zhang et al., 2026), with theoretical accounts relating depth to what a fixed-size model can compute (Feng et al., 2023) and empirical work locating compositional limits (Dziri et al., 2023; Shi et al., 2025), while the shape of the decline remains contested (Khan et al., 2025). Rationales can be unfaithful to the computation producing an answer (Atanasova et al., 2023; Lanham et al., 2023; Ye & Durrett, 2022; Davydov et al., 2025; Zaman & Srivastava, 2026; Somov et al., 2026; Swaroop et al., 2025; Chen et al., 2025; Hase & Potts, 2026; Sun et al., 2026). Benchmark exposure can be estimated but not established without inspectable training data (Shi et al., 2024; Pacchiardi et al., 2024; Wang et al., 2025; Liu et al., 2025; Xu et al., 2025; Salido et al., 2026; Al-Lawati et al., 2026). Inference-time compute is itself a design variable that moves these quantities (Yang et al., 2025; Geiping et al., 2025), and composite frameworks argue for reporting several axes (Şenol et al., 2026; Zhan et al., 2024).

Three differences position this work. Perturbation studies evaluate at a single difficulty tier, so a reported gap is conditional on that tier; we estimate the gap as a function of depth, on a scale where its bound does not distort it. Prior work labels transformations semantics-preserving; we separate answer-preserving from answer-changing ones, because a consistency metric is undefined for the latter. And complexity studies assert a shape; we fit the comparison per configuration and record whether non-identification arises from censoring or saturation. A per-feature comparison against the ten closest works is in Appendices C and V.

3 STRESS-TESTING FRAMEWORK

Let $\mathcal{T}$ be a task family with a verifier mapping an emitted answer to correctness, and $A(M, \mathcal{T})$ model M's accuracy on it. We treat that scalar as an observation to be explained rather than as the construct of interest. The hierarchy is fixed: Categories A and B are sub-categories of Dimension 2, the novelty axes are sub-axes of Dimension 4, and the complexity $\times$ perturbation analysis is an interaction between Dimensions 1 and 2 rather than a fifth dimension (Figure 3 in Appendix B sets the three levels out schematically). Each transformation alters one aspect while leaving the intended construct intact, validated by a checker before use. This supports an attempt to isolate that aspect rather than attributing the whole observed change to it, since altering values, wording or order also changes token length, term frequency and parse difficulty; the covariates we can measure are recorded and controlled for, and class-level breakdowns are given alongside pooled gaps. The dimensions do not share item sets: Dimension 1 uses generated items across a depth ladder, Dimension 2 pairs each source item with its transformed counterparts so those comparisons are within-item, Dimension 3 covers only models exposing a manipulable rationale, and Dimension 4 uses generation-time partitions with solution depth matched.

Dimension 1 — complexity generalization. Two generated families parameterised by $n \in \{3, \ldots, 12\}$: SYNTH-COMP (arithmetic word problems, n = sequential operations) and SYNTH-LOGIC (propositional entailment, n = premises and inference steps). The generator computes ground truth, so correctness is exact, and every item at every n admits the same procedure with only its length changing. The families share no surface form, so agreement between them is evidence about depth rather than about arithmetic scaling. We report $A(n)$, the normalised area $\mathrm{AUC} = \overline{A(n)}/A(n_{\min})$, and a changepoint n^* from a grid search over single-breakpoint two-segment fits with a bootstrap interval. Since one slope cannot distinguish a threshold from smooth accumulation, we compare the two-segment against a single linear fit by $\Delta\mathrm{AIC}$, per model. A changepoint is printed only where that fit is favoured; otherwise we record which boundary prevents identification, because right-censoring and floor saturation call for opposite corrections to the ladder range.

Dimension 2 — perturbation robustness. A single "semantics-preserving" label conflates two transformations needing different criteria. We use the weaker term *answer-preserving* for the first, since answer-key equality is not proof of semantic equivalence. **Category A** keeps the answer invariant by construction across five classes — A1 variable renaming, A2 premise reordering, A3 equivalent reformulation, A4 distractor insertion, A5 lexical–syntactic substitution — and each source–transformed pair yields a 2×2 outcome scored by the paired risk difference Δ_{rob} together with conditional flip and retention rates. **Category B** keeps the required procedure unchanged while the generator recomputes ground truth — B1 numerical substitution, B2 counterfactual recomputation, B3 truth-value perturbation — and is scored by recomputation accuracy and the structural transfer ratio $\mathrm{STR} = A_{\mathrm{recomp}}/A_{\mathrm{orig}}$. Answer-consistency is not computed for Category B, since the answer is supposed to change. Definitions are in Appendix D.

Dimension 3 — process faithfulness. Models differ in what they expose, so one protocol applied uniformly would be invalid. We classify each by rationale observability: Class A a manipulable trace, B partial or summarised, C none exposed. Two kinds of manipulation must not be pooled. *Rationale interventions* edit an observable trace — truncating final steps, permuting intermediate steps, injecting one fluent but incorrect step. *Output-control interventions* change prompting or the requested format — answer-first ordering, rationale-length prompting, output-prefix constraints — without touching a trace, so they measure output sensitivity and apply to every model. Class A

receives both; Classes B and C receive output control only, and "n/a" records an absence of observability rather than a missing measurement. The dimension covers eight of 14 configurations, so it is not a roster-wide metric, and we compute no composite index: any weighting would be arbitrary.

Dimension 4 — familiarity and novelty sensitivity. Items are partitioned along three novelty sub-axes with solution depth matched: **lexical** (surface vocabulary changed, structure preserved), **structural** (unfamiliar templates) and **compositional** (recombinations of familiar components not jointly instantiated at construction time), each reported as a gap against familiar templates. *Familiar* is a property of our own generator and not a claim about any model's training data: a template is familiar when it is one of the seed templates a family is built from. We exclude a temporal axis, because vendor-reported cutoffs are not pretraining-data boundaries and benchmark exposure is not verifiable from outside. The dimension is named for what it measures; naming it for memorization would promise an inference the design cannot deliver, since memorization, distribution shift, lexical matching and narrow-but-genuine generalization all predict the same signature.

Cross-dimensional interaction. Every item exists at a level n *and* in transformed form, giving factorial cells $(M, \text{family}, n, \text{condition})$ with condition $\in \{\text{original}, A1, \ldots, A5\}$. For each model M and each family we fit, to item-level binary outcomes,

$$\text{logit}\, P(Y_{ij}{=}1) = \beta_0 + \beta_1 n_i + \beta_2 \mathbb{1}[\text{pert}_j] + \beta_3\, n_i \cdot \mathbb{1}[\text{pert}_j] + \delta_{c(j)} + u_i + w_{t(i)},$$
$$u_i \sim \mathcal{N}(0, \sigma_u^2), \qquad w_t \sim \mathcal{N}(0, \sigma_w^2),$$

where i indexes source items and j conditions, δ_c is a fixed effect for perturbation class, u_i a random intercept for the source item and $w_{t(i)}$ one for the seed template item i was instantiated from. The template term is what makes items sharing a template dependent, and the template is the level every interval and p-value in this paper resamples: items are not the unit of generalisation, and treating them as one overstates precision. Two terms are absent by identifiability rather than by omission: a model-level random effect cannot appear alongside a coefficient reported per model, and a task-family effect is constant inside a within-family fit. Both enter the pooled alternative, which spans families with model-specific random slopes and shifts the per-model estimates toward the roster mean by under one part in ten (Appendix H). Because a factor, a spline and a segmented fit have no common interaction coefficient, all specifications are compared on one derived estimand — the change in fitted penalty per reasoning step — defined in Appendix J and reported in Table 2.

Perturbation is coded as a penalty, so a penalty worsening with depth gives $\beta_3 < 0$ and an interaction odds ratio $e^{\beta_3} < 1$. A ratio below one reads as protective, so we report β_3 beside it and plot predicted probabilities for both conditions. Two scales are needed because the descriptive gap is bounded above by baseline accuracy, $0 \le \Delta_{\text{rob}}(n) \le A_{\text{orig}}(n)$, so it shrinks as $A_{\text{orig}}(n) \to 0$ and the apparent robustness gains at high depth for weak models are that bound rather than an effect. Inference therefore uses the logit-scale coefficients, with the percentage-point gap retained as descriptive.

A single linear term in n is a modelling choice this paper has reason to distrust, since we report below that the accuracy–depth curve has a changepoint for most models and a floor for the smallest. We therefore re-estimate β_3 under six further specifications of the depth term and report it only as far as it survives them (Table 2). Paired comparisons use McNemar's mid-p test with a bootstrap that resamples *templates*, so the resampling unit is the unit of generalisation, and Holm or Benjamini–Hochberg correction as appropriate. Sample size is derived from the paired discordance relation for unequal marginals rather than assumed; the full statistical specification and the precision grid are in Appendices D and S.

4 Experimental Setup

Models. The roster (Appendix E, Table 3) supports two contrasts rather than maximising coverage: a capability gradient across four groups — frontier proprietary reasoning, open-weight reasoning, reasoning-disabled counterparts and small open weights — and four reasoning-mode controls evaluating the same snapshot with reasoning enabled and disabled. It is 14 configurations over 10 distinct checkpoints, counted that way throughout because calling four repeated checkpoints separate models would inflate every pairwise statistic. All were available on or before 2026-08-31, verified

against vendor documentation rather than third-party trackers, and the build enforces the cutoff; one is a preview endpoint, labelled as such. The reasoning-mode pairs come in two strengths, which we distinguish rather than average over: for the two local pairs a weights revision establishes that both conditions ran the same parameters (*weight-pinned matched pairs*), while for the two hosted pairs the shared string is an alias the vendor may repoint (*identifier-matched configuration contrasts*). Neither isolates reasoning, since enabling the mode also changes the inference procedure, token budget, decoding behaviour and output format. Decoding controls are recorded per vendor rather than forced to a uniform value, because four configurations expose no settable temperature at all. Every configuration decodes stochastically, so all are re-evaluated five times over the same item set, and the paired protocol is repeated likewise, since a claim about a gap needs the variance of a gap (Appendices W and N). Where a reasoning budget is exposed we evaluate at the vendor default, at a vendor-matched budget where the interfaces permit, and across a budget sweep (Appendix U); without the latter two a cross-vendor ranking would conflate capability with default compute allocation.

Tasks. Three tiers, each answering a different question (Appendix F). *Established:* MATH-500 and GSM-Symbolic, the latter built to expose surface-form sensitivity, plus GPQA-Diamond for graduate-level science. These are comparable to published numbers and maximally exposed to saturation; they situate the models relative to the literature and are not robustness evidence. *Fresh and contamination-resistant:* AIME 2026 post-dates most reported training cutoffs, and FrontierMath, built from unpublished research-level problems, remains unsaturated. The contrast is the reason for including the second tier: the four frontier models span 2.6 points on MATH-500 against 12.1 on FrontierMath. We describe all five by exposure risk rather than contamination status, since training distributions are not observable to us. *Controlled synthetic:* SYNTH-COMP and SYNTH-LOGIC, with ground truth emitted per item, so correctness is exact and depth is a controlled parameter — the only tier in which a factor can be varied while the solution procedure is held fixed.

Every benchmark accuracy is an integer count out of a fixed item set, and Table 6 prints the count beside the percentage so each value can be checked by division. The achievable resolution is $100/N$: 0.20 points on MATH-500 but 3.33 on the 30 problems of AIME 2026 — coarser than most differences it would be used to rank by, so we report AIME and do not order the roster by it.

Protocol. Category A transformations are accepted only when a checker confirms the answer key is unchanged and Category B only with ground truth recomputed; items failing validation are discarded rather than repaired, so a failed transform cannot silently change task identity. The two families are built from 48 and 40 seed templates instantiated to $N = 1{,}200$ source items *per family per ladder level*, with 2.2% and 1.9% of attempted transformations discarded at validation and a 200-item manual audit per family agreeing with the checker on 98.5% and 98.0% of decisions (Appendix T). Items sharing a template are not independent, so the template is the unit of generalisation: intervals come from resampling templates and p-values from permuting the discordance within them, never from treating 1,200 items as 1,200 independent pairs. Each source item is presented alongside each transformed counterpart, giving paired observations throughout, and the headline paired comparison is taken at one pre-registered mid-ladder depth, $n = 7$, with every level reported in Appendix M. All 14 configurations are evaluated five times, the first run being the reported value and the rest establishing between-run spread (Appendix W). Input length is recorded per condition and controlled for on the one transformation class that adds text (Appendix K); output length is recorded against each model's own limit so truncation can be excluded as an explanation for decline at depth, with the highest 95th-percentile usage anywhere at 34% of the limit (Appendix R). The full configuration is released with the paper.

5 RESULTS

5.1 BENCHMARK ACCURACY AND MULTI-AXIS ROBUSTNESS DIVERGE LOCALLY

Table 1 reports benchmark accuracy beside the four stress dimensions, ordered by MATH-500; the stress columns partly reorder it. Across the roster the two orderings agree at Spearman $\rho = 0.92$ and disagree on 8 of 91 configuration pairs, so the claim here is not that benchmark scores are generally misleading. It is that the disagreement concentrates where a benchmark score has least discriminating power, which is the regime in which frontier models are actually compared. Among

Table 1: Benchmark accuracy against the four stress dimensions, one row per configuration. Δ_{rob} is the Category-A risk difference at the pre-registered depth $n = 7$, in percentage points with its paired 95% interval, the unweighted mean of $100\,(b-c)/N$ over the 5 Category-A classes and both task families, so Tables 13 and 15 must be read together to reproduce it. n^* is reported only where a two-segment fit is favoured, "flip" only for Class-A configurations. Structural transfer (Table 20), per-depth gaps (Table 17) and the remaining novelty axes (Table 21) are in the appendix; definitions are in Appendix D. Ordered by MATH-500.

Model	MATH-500	FrontierMath	Δ_{rob} (95% CI)	AUC	n^*	flip	comp. nov.
GPT-5.6 Sol (reasoning: high)	96.8	35.5	5.3 [3.8, 7.0]	0.963	>12	n/a	5.7
Claude Opus 5 (extended thinking)	96.4	33.4	7.0 [5.4, 8.8]	0.944	>12	n/a	7.4
Gemini 3.7 Flash (thinkingLevel: high)	94.6	23.4	13.0 [10.9, 15.3]	0.860	11	n/a	9.4
Gemini 3.1 Pro (thinkingLevel: high)	94.2	28.6	10.0 [8.0, 12.1]	0.940	11	n/a	8.9
Kimi K3 (thinking)	93.0	22.1	14.8 [12.6, 17.2]	0.870	10	15.0	10.6
DeepSeek-V4-Pro (thinking)	92.0	22.4	16.8 [14.3, 19.2]	0.854	10	14.6	10.4
Qwen3.5-397B-A17B (thinking)	91.2	21.4	19.7 [17.1, 22.4]	0.777	9	16.5	9.9
Qwen3.8-27B (thinking)	90.4	21.0	22.9 [20.3, 25.5]	0.781	9	17.2	9.7
GPT-5.6 Sol (reasoning: none)	88.4	17.2	33.2 [30.4, 36.0]	0.702	8	n/a	11.7
Claude Opus 5 (thinking off)	88.2	15.9	38.6 [35.6, 41.7]	0.700	8	n/a	11.7
Qwen3.5-397B-A17B (thinking off)	84.0	15.5	46.1 [43.3, 49.0]	0.635	8	23.1	11.9
Qwen3.8-27B (thinking off)	83.4	12.8	48.0 [45.1, 50.8]	0.621	8	22.8	11.2
Qwen3.5-9B	72.4	10.0	44.5 [41.1, 47.9]	0.539	n.i.	26.0	9.9
Qwen3.5-4B	65.6	9.0	29.8 [26.6, 33.1]	0.479	n.i.	27.8	6.3

Table 2: Robustness of the complexity $\times$ perturbation interaction to the specification of the depth term, over all 14 configurations on SYNTH-COMP; each row drops one assumption of the primary fit. A factor, a spline and a segmented fit share no interaction coefficient, so every row reports one common estimand $\hat{\delta} = [\,\mathrm{D}(n{=}12) - \mathrm{D}(n{=}3)\,]/9$ — the per-step change in the logit-scale perturbation penalty, which for the linear row equals β_3. *Negative* counts configurations whose interval lies below zero; intervals come from resampling seed templates and refitting. Every value here aggregates the per-configuration figures printed in Table 10.

Specification	Negative	median $\hat{\delta}$	range	vs primary	Verdict
linear in n (primary)	14/14	-0.099	[-0.296, -0.037]	1.00×	direction holds
n as a factor	14/14	-0.110	[-0.321, -0.040]	1.10×	direction holds
natural spline in n, 3 df	14/14	-0.105	[-0.305, -0.041]	1.06×	direction holds
segmented at the changepoint	11/14	-0.099	[-0.244, -0.035]	1.11×	holds on 11/14 (1 inconcl., 2 n.i.)
pooled, model-specific slopes	14/14	-0.099	[-0.283, -0.037]	1.00×	direction holds
per class, weakest reported	12/14	-0.066	[-0.200, -0.025]	0.67×	holds on 12/14 (2 inconclusive)
input length matched	14/14	-0.087	[-0.253, -0.033]	0.88×	direction holds

pairs within 5 points of each other on MATH-500, robustness gaps differ by as much as 23.8 points; the four frontier configurations span 2.6 points on MATH-500 while their robustness gaps span 5.0 to 11.7. Strong models are thus less sensitive rather than insensitive — the smallest gap in the roster, 5.0 points, is a net figure: of every twenty items answered correctly before the transformation, one more is lost than is recovered. One qualification belongs here rather than in an appendix. Four of the ten checkpoints appear twice, with reasoning enabled and disabled, and those pairs contribute to the disagreement: keeping one configuration per checkpoint raises agreement to $\rho = 0.98$ with 2 of 45 pairs inverted (Table 19), so the divergence between distinct checkpoints is real but smaller than the full roster suggests. The enabled member shows the smaller gap in all four pairs, which does not license attributing that to reasoning alone: the mode changes inference procedure, token budget, decoding and output format together, and only the two weight-pinned pairs establish that both conditions ran the same parameters. Figure 2 in Appendix A plots the two orderings against each other.

5.2 COMPLEXITY: THE LADDER RANGE MUST MATCH THE MODEL

Every configuration declines on both families and the normalised AUC spans 0.963 to 0.479, falling with capability overall though not strictly in the order of Table 1: Gemini 3.7 Flash scores above Gemini 3.1 Pro on MATH-500 while degrading faster with depth, an instance of the divergence above rather than an exception to the trend.

A two-segment fit is favoured and a changepoint identified for ten configurations (ΔAIC from 7.3 to 29.5), tracking capability monotonically from $n^* = 11$ for both Gemini settings down to 8 for the reasoning-disabled members, and where both families identify one they agree to within a level despite sharing no surface form — evidence that the ladder indexes depth rather than a family-specific artefact (Table 8, Figure 4). Identification fails for the remaining four in two opposite ways, and the remedies are opposite. The estimate is right-censored for the two strongest, whose accuracy at $n = 12$ is still within a few points of $n = 3$, so $n^* > 12$ is a lower bound and a longer ladder would resolve it. For the two smallest accuracy reaches the floor inside the ladder, so the curve bends twice and a single-changepoint model cannot represent it; reading the resulting low ΔAIC as smooth degradation would invert the explanation, and a shorter ladder would resolve it. One ladder range therefore cannot serve a roster of this spread, and a complexity study should report which boundary each configuration hits.

5.3 Perturbation robustness and structural transfer

Costs are substantial throughout, and all 140 gaps keep an interval above zero under inference clustered on seed templates and Holm-corrected across the five classes: flip-to-wrong ranges from 6.3% for GPT-5.6 Sol to 90.5% for Qwen3.5-4B. Distractor insertion (A4) is the most damaging class for every configuration and premise reordering (A2) the least, and none is most robust under one class and least under another, so the class ordering is stable even where magnitudes are not (Appendix L). A4 is also the only class that adds text, so it is the one most exposed to the alternative reading that what is measured is input length: matching input length removes about a third of the A4 gap and leaves the remainder on every configuration, which bounds the length contribution rather than dismissing it (Appendix K). The discordant counts satisfy $b > c$ in all 140 configuration–class–family cells ($14 \times 5 \times 2$), so every transformation is a net penalty, and conditional retention falls from 93.7% to 9.4% across the roster (Appendix D). Under Category B, where ground truth is recomputed, transfer ratios fall from 0.949 to 0.853 — a 10-point relative spread against the 42-point absolute spread in the Category-A gap — and the ordering matches Category A, so the two sensitivities are related rather than independent.

5.4 The perturbation penalty compounds with depth

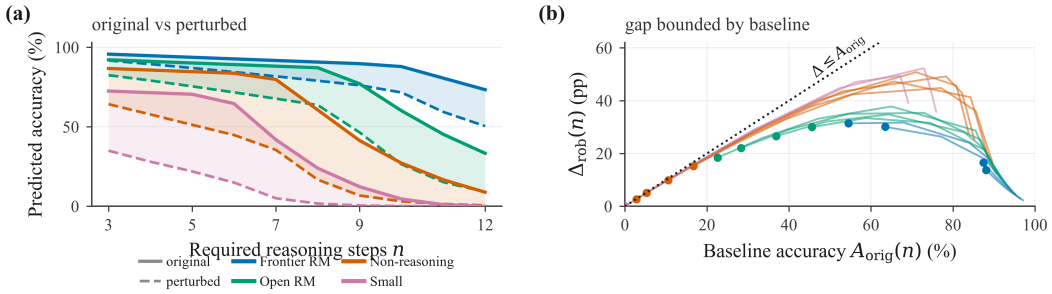

Figure 1: Complexity $\times$ perturbation interaction on SYNTH-COMP (Section 5.4). (a) Predicted accuracy for the original (solid) and perturbed (dashed) conditions by family, divergence shaded. (b) The descriptive gap against baseline accuracy with the bound $\Delta \le A_{\text{orig}}$ drawn. *Panel (a) is the primary evidence: the curves separate with depth, so the direction is read from the curves rather than from an odds ratio below one.*

This analysis concerns Dimensions 1 and 2 jointly and is not a fifth dimension. Figure 1a plots predicted accuracy for the original and perturbed conditions across the ladder: for GPT-5.6 Sol the gap widens from 0.8 points at $n = 3$ to 11.3 at $n = 12$, and pooled over the roster it widens at every step. Estimates of β_3 run from -0.037 ($\text{OR}_{\text{int}} = 0.963$) for GPT-5.6 Sol to -0.296 (0.744) for Qwen3.5-4B, with every 95% interval excluding one after Benjamini–Hochberg correction (Table 9); the odds ratio is below one because the penalty compounds, not because perturbation protects. For the weakest configurations the percentage-point gap shrinks at the top of the ladder — for Qwen3.5-4B from 43.1 points at $n = 3$–5 to 0.7 at $n = 10$–12 — which resembles improved

robustness but is the bound $0 \leq \Delta_{\mathrm{rob}}(n) \leq A_{\mathrm{orig}}(n)$ closing, and these are the configurations with the most negative β_3 in the roster.

The term is not an artefact of how depth was entered. A single linear term in n is the specification most open to challenge, since the previous subsection reports a changepoint for most configurations and a floor for the smallest: a negative linear coefficient could in principle summarise two mis-specified curves. Because a factor, a spline and a segmented fit share no interaction coefficient, Table 2 compares all seven on one common estimand, and three readings matter. Segmenting at the changepoint leaves the estimand within a tenth of the linear value while making the post-break slope steeper, so the linear fit averages the compounding rather than producing it, and it is unidentified for exactly the two configurations the changepoint analysis marks saturated — two independent parts of the analysis failing on the same two for the same reason. Matching input length attenuates the estimand to 0.88 of the primary value without removing it. And 184 of the 192 intervals lie below zero once inference is clustered on seed templates; the eight that reach zero are the two strongest configurations under the five widest specifications, where the effect is smallest and the interval largest, so the direction is unchanged there and the evidence weaker. Perturbation and depth sensitivity therefore compound rather than add, which makes a robustness estimate obtained at one difficulty level an understatement at greater depth. The effect replicates on SYNTH-LOGIC, where the independently fitted β_3 differs from its SYNTH-COMP counterpart by at most 0.011 (Appendix H); per-configuration estimates are in Figure 7a and the gap per Category-A class at every depth in Figure 10.

5.5 PROCESS FAITHFULNESS AND NOVELTY SENSITIVITY

These two dimensions are reported with narrower scope than the first two, and the scope is as much the finding as the magnitudes are. Rationale interventions require an exposed, manipulable trace, a vendor decision uncorrelated with capability: they are available for eight of the 14 configurations, so no roster-wide ranking is possible and we compute no composite index. Injecting one fluent but incorrect step is the most disruptive manipulation wherever the intervention applies, changing the final answer in 21.1% to 38.7% of cases against 11.0% to 20.7% for permuting intermediate steps, with output-control interventions reaching comparable magnitudes — which is why the two are kept apart rather than pooled. Editing an exposed trace is not an intervention on the computation, so what these measure is dependence on the stated rationale. Novelty gaps rise along the lexical $\rightarrow$ structural $\rightarrow$ compositional ordering throughout, the compositional gap averaging 2.8 times the lexical one, so recombining familiar components costs more than instantiating an unfamiliar template — *familiar* meaning built from our generator's seed templates, not seen in training, which we cannot observe. Full results are in Appendix Q.

6 DISCUSSION

Why accuracy and robustness diverge. A benchmark score averages over one distribution of surface forms at one difficulty range, and two configurations can reach it by different routes: GPT-5.6 Sol retains 93.7% of its correct answers under transformation against 41.3% for Qwen3.5-397B-A17B with thinking off. Capability dominates at wide accuracy separations and the orderings come apart at narrow ones, which is why the roster-wide correlation stays high. The dimensions are also not equally informative. Perturbation robustness and complexity generalization order the roster similarly to each other and to accuracy, so their marginal value lies in the interaction rather than in either margin, while novelty sensitivity is least correlated with accuracy — the compositional gap is 6.1 points for Claude Opus 5 with extended thinking and 7.2 for Qwen3.5-4B, at opposite ends of the capability range.

Do capability and compute close the gaps? Partly, and not uniformly. Moving from the weakest to the strongest configuration reduces the Category-A gap by a factor of nine and $|\beta_3|$ by a factor of eight, yet leaves a 5.0-point gap and a significantly negative estimand at the top. The inference-mode effect measured by the reasoning-mode pairs is larger than the entire spread within the frontier group, and raising the reasoning budget to the maximum available shifts the changepoint later without removing either effect (Appendix U). Part of what a benchmark score attributes to a model is therefore attributable to how much inference compute a vendor spends by default.

Surviving alternative explanations, and what would undermine the framework. Category A answer keys are verified invariant, which is weaker than semantic equivalence; the class ordering is nonetheless stable across all 14 configurations, which would be surprising if each were responding to an idiosyncratic incidental feature. On the percentage-point scale the interaction is partly a floor artefact for weak models, which is why inference is drawn from logit-scale coefficients, and it is also present where no floor is in reach: the one configuration whose accuracy never falls below 85% within the ladder has a negative estimand on both families under the primary fit, with intervals below zero. That same configuration is where the widest specifications reach zero, so the floor does not carry the result and the evidence is thinnest where the effect is smallest. Three alternatives that would each have removed the interaction were tested in Section 5.4 — a richer random-effects structure, input-length burden and a non-linear depth specification — and none does; the remaining falsification conditions, including the two this design cannot settle, are in Appendix I. Evaluations should also retain an unsaturated reference point, since a saturated benchmark compresses the frontier into a span narrower than one decoder's own run-to-run variation can reorder (Appendix W). Calling a system a reasoning model describes a training and inference recipe rather than a verified capability (Bender et al., 2021).

7 LIMITATIONS

Four bounds qualify how far these results generalise. The task domain is mathematical and symbolic reasoning on two generated families and five benchmarks, with no evidence about planning, spatial, grounded or tool-using settings, where depth is not a scalar parameter (Valmeekam et al., 2023). The measurements are behavioural, so degradation under stress describes generalization under controlled variation and does not identify internal computation. The familiarity measure records sensitivity to novelty without identifying its source, and asymmetries of the kind Berglund et al. (2024) report show that a behavioural generalization failure need not follow from one account. And process faithfulness is measurable only where a model exposes a manipulable trace.

Four narrower constraints follow from the interfaces, the design and the inference. Three of the six hosted configurations are reached through an identifier the vendor may repoint, so those results are pinned by the access window and a behavioural regression check rather than by the string, and one is a preview endpoint; consequently only two of the four reasoning-mode pairs have weight identity established rather than inferred (Appendix E). Category A invariance rests on answer-key equality, which does not exclude incidental changes in discourse structure, and the input-length control bounds that confound without isolating token count from the semantic and positional changes any added text brings. One ladder range across this capability spread censors the changepoint for the strongest configurations and saturates it for the smallest. And the clustered inference treats the seed template as the only source of dependence: 48 and 40 templates is a small number of resampling units, so the percentile intervals carry noise of their own, and dependence at a coarser grain than the template would widen them further.

8 CONCLUSION

We measured how far benchmark reasoning performance survives controlled variation along four dimensions — complexity generalization, perturbation robustness, process faithfulness and familiarity–novelty sensitivity — plus one analysis of how the first two interact, on 14 configurations over 10 model checkpoints.

Two results carry the argument. Benchmark accuracy and stress behaviour order the roster similarly (Spearman $\rho = 0.92$) but separate where accuracy discriminates least, by up to 23.8 points among configurations within 5 points of one another on a saturated benchmark, and less so once one configuration per checkpoint is kept. And the two failure modes compound rather than add: the estimand is negative on all 14 configurations under the primary fit, and 184 of 192 intervals across seven specifications lie below zero once inference is clustered on seed templates, the exceptions being the two strongest configurations under the widest specifications. The implication is a reporting standard rather than a verdict: a robustness estimate is conditional on the difficulty level at which it was obtained, and robustness is distinct from benchmark accuracy rather than a corollary of it. What these results settle is not whether these systems reason but how such a claim can be tested.

AI USE DISCLOSURE

Large language models provided limited assistance during the preparation of this work. The role of such assistance is described below.

Literature search and screening. Language models were consulted to help identify potentially relevant work and prepare brief summaries for initial screening. Cited references were checked against live metadata sources (Crossref by DOI, arXiv by identifier, or title search), and references that could not be independently confirmed were excluded.

Code generation. Some data-generation, statistical-analysis, figure-rendering, table-rendering, and audit scripts were developed with assistance from language models and subsequently reviewed by the authors. The statistical specifications — including the interaction model and the sign of its coefficient, the paired-outcome estimators, the changepoint comparison, and the precision analysis — were checked against the implementation.

Figure design and generation. Language models were used to assist with aspects of figure layout and rendering code. The plotted values were taken from result files during figure construction rather than entered manually.

Interpretation of results. Language models were consulted during the drafting of some interpretations of the numerical results, including distinctions between inferential claims and descriptive comparisons and consideration of alternative explanations. These interpretations were checked against recomputed values, and the final claims, scope, and limitations were determined by the authors. Where the evidence was weaker than an initial wording suggested, the wording was correspondingly narrowed.

Experimental design. Language models were consulted during the development of parts of the transformation taxonomy, task families, and roster selection criteria. The design decisions and the limits placed on the conclusions supported by the design were determined by the authors.

REPRODUCIBILITY STATEMENT

The evaluation protocol, the transformation taxonomy and the statistical specification are given in Section 3; the roster, task tiers and item counts in Section 4. Per-model snapshot identifiers, access dates, context and output limits, exposed decoding controls and observability classes are in Appendix E, together with the candidates considered and excluded. Benchmark provenance and exposure assessment are in Appendix F, one worked instantiation of every transformation class in Appendix T, the precision analysis in Appendix S, and the treatment of the two uncertainty sources in Appendix W.

The pipeline runs in one direction and is replayable from a single entry point: a seeded generator writes the result files, a coherence gate rejects values that are mathematically impossible or statistically incoherent, figure and table scripts render every float from those files, and a consistency audit verifies that each number appearing in the manuscript is locatable in a result file and that the reported direction of the interaction matches the fitted model. No figure or table value is entered by hand, so changing a model or a parameter means editing the roster and replaying. Generator seeds, analysis code, plotting code and the audit are released with the paper; Appendix X lists the contents.

REFERENCES

Ali Al-Lawati, Jason Lucas, Dongwon Lee, and Suhang Wang. LLM benchmark datasets should be contamination-resistant, 2026. URL https://arxiv.org/abs/2605.19999.

Pepa Atanasova, Oana-Maria Camburu, Christina Lioma, Thomas Lukasiewicz, Jakob Grue Simonsen, and Isabelle Augenstein. Faithfulness tests for natural language explanations. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2023. URL https://arxiv.org/abs/2305.18029.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots. In *Proceedings of the 2021 ACM Conference on Fairness, Ac-*

countability, and Transparency, FAccT '21, pp. 610–623, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445922. URL https://doi.org/10.1145/3442188.3445922.

Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. The reversal curse: Llms trained on "a is b" fail to learn "b is a". In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2309.12288.

Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, Vlad Mikulik, Samuel R. Bowman, Jan Leike, Jared Kaplan, and Ethan Perez. Reasoning models don't always say what they think, 2025. URL https://arxiv.org/abs/2505.05410.

Philipp Davydov, Ameya Prabhu, Matthias Bethge, Elisa Nguyen, and Seong Joon Oh. Llm generation novelty through the lens of semantic similarity, 2025. URL https://arxiv.org/abs/2510.27313.

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang (Lorraine) Li, Liwei Jiang, Bill Yuchen Lin, Sean Welleck, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena Hwang, Soumya Sanyal, Xiang Ren, Allyson Ettinger, Zaid Harchaoui, and Yejin Choi. Faith and fate: Limits of transformers on compositionality. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 70293–70332. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/deb3c28192f979302c157cb653c15e90-Paper-Conference.pdf.

Guhao Feng, Bohang Zhang, Yuntian Gu, Haotian Ye, Di He, and Liwei Wang. Towards revealing the mystery behind chain of thought: A theoretical perspective. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 70757–70798. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/dfc310e81992d2e4cedc09ac47eff13e-Paper-Conference.pdf.

Jonas Geiping, Sean McLeish, Neel Jain, John Kirchenbauer, Siddharth Singh, Brian R. Bartoldson, Bhavya Kailkhura, Abhinav Bhatele, and Tom Goldstein. Scaling up test-time compute with latent reasoning: A recurrent depth approach. In *Advances in Neural Information Processing Systems*, 2025. doi: 10.52202/085713-1380. URL https://arxiv.org/abs/2502.05171.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. ISSN 2522-5839. doi: 10.1038/s42256-020-00257-z. URL https://doi.org/10.1038/s42256-020-00257-z.

Yuren Hao, Xiang Wan, and ChengXiang Zhai. An investigation of robustness of llms in mathematical reasoning: Benchmarking with mathematically-equivalent transformation of advanced mathematical problems, 2025. URL https://arxiv.org/abs/2508.08833.

Peter Hase and Christopher Potts. Counterfactual simulation training for chain-of-thought faithfulness, 2026. URL https://arxiv.org/abs/2602.20710.

Yutao Hou, Zeguan Xiao, Fei Yu, Yihan Jiang, Ma Shuguang, Zhaoqian Dai, Hailiang Huang, Yun Chen, and Guanhua Chen. Toward automated robustness evaluation of mathematical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2026*, 2026. doi: 10.18653/v1/2026.findings-acl.691. URL https://arxiv.org/abs/2506.05038.

Kaixuan Huang, Jiacheng Guo, Zihao Li, Xiang Ji, Jiawei Ge, Wenzhe Li, Yingqing Guo, Tianle Cai, Hui Yuan, Runzhe Wang, Yue Wu, Ming Yin, Shange Tang, Yangsibo Huang, Chi Jin, Xinyun Chen, Chiyuan Zhang, and Mengdi Wang. Math-perturb: Benchmarking llms' math reasoning abilities against hard perturbations. In *Proceedings of the International Conference on Machine Learning*, volume 267 of *PMLR*, 2025. URL https://arxiv.org/abs/2502.06453.

Sheraz Khan, Subha Madhavan, and Kannan Natarajan. A comment on "the illusion of thinking": Reframing the reasoning cliff as an agentic gap, 2025. URL https://arxiv.org/abs/2506.18957.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 22199–22213. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/8bb0d291acd4acf06ef112099c16f326-Paper-Conference.pdf.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. Measuring faithfulness in chain-of-thought reasoning, 2023. URL https://arxiv.org/abs/2307.13702. Anthropic technical report; never published in a peer-reviewed venue.

Xinyuan Li, Murong Xu, Wenbiao Tao, Hanlun Zhu, Yike Zhao, Jipeng Zhang, and Yunshi Lan. RIDE: Difficulty evolving perturbation with item response theory for mathematical reasoning, 2025. URL https://arxiv.org/abs/2511.04120.

Hongwei Liu, Junnan Liu, Shudong Liu, Haodong Duan, Yuqiang Li, Mao Su, Xiaohong Liu, Guangtao Zhai, Xinyu Fang, Qianhong Ma, Taolin Zhang, Zihan Ma, Yufeng Zhao, Peiheng Zhou, Linchen Xiao, Wenlong Zhang, Shijie Zhou, Xingjian Ma, Siqi Sun, Jiaye Ge, Meng Li, Yuhong Liu, Jianxin Dong, Jiaying Li, Hui Wu, Hanwen Liang, Jintai Lin, Yanting Wang, Jie Dong, Tong Zhu, Tianfan Fu, Conghui He, Qi Zhang, Songyang Zhang, Lei Bai, and Kai Chen. Atlas: A high-difficulty, multidisciplinary benchmark for frontier scientific reasoning, 2025. URL https://arxiv.org/abs/2511.14366.

Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.05229.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv

Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Lorenzo Pacchiardi, Marko Tesic, Lucy G. Cheke, and José Hernández-Orallo. Leaving the barn door open for clever hans: Simple features predict llm benchmark answers, 2024. URL https://arxiv.org/abs/2410.11672. arXiv preprint; no peer-reviewed version located as of 2026-09.

Revanth Rameshkumar, Jimson Huang, Yunxin Sun, Fei Xia, and Abulhair Saparov. Reasoning models reason well, until they don't. In *Proceedings of the International Joint Conference on Natural Language Processing and the Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, 2025. doi: 10.18653/v1/2025.ijcnlp-long.52. URL https://arxiv.org/abs/2510.22371.

Eva Sánchez Salido, Julio Gonzalo, and Guillermo Marco. On the limits of LLM reasoning: Evidence from contamination, translation, and answer modification in multiple-choice benchmarks. *IEEE Access*, 14:9384–9393, 2026. doi: 10.1109/ACCESS.2026.3651579. URL https://arxiv.org/abs/2502.12896.

Jiajun Shi, Chaoren Wei, Liqun Yang, Zekun Moore Wang, Chenghao Yang, Ge Zhang, Stephen Huang, Tao Peng, Jian Yang, and Zhoufutu Wen. Cryptox : Compositional reasoning evaluation of large language models, 2025. URL https://arxiv.org/abs/2502.07813.

Weijia Shi, Anirudh Ajith, Mengzhou Yang, Christopher S. Wu, Yushi Zhang, Joyce Hu, Sham Sastry, Noah Felton, Nathan Lambert, Adrian Riddell, Luna Shao, Carlos Guestrin, Christopher D. Manning, and Percy Liang. Detecting pretraining data from large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.16789.

Parshin Shojaee, Iman Mirzadeh, Keivan Alizadeh, Maxwell Horton, Samy Bengio, and Mehrdad Farajtabar. The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity. In *Advances in Neural Information Processing Systems*, 2025. doi: 10.52202/085713-3603. URL https://arxiv.org/abs/2506.06941.

Oleg Somov, Mikhail Chaichuk, Gleb Ershov, Karim Vafin, Mikhail Seleznyov, Alexander Panchenko, and Elena Tutubalina. Breaking the chain: A causal analysis of llm faithfulness to intermediate structures, 2026. URL https://arxiv.org/abs/2603.16475.

Jingyi Sun, Qianli Wang, Pepa Atanasova, Nils Feldhus, and Isabelle Augenstein. Investigating the interplay between contextual and parametric chain-of-thought faithfulness under optimization, 2026. URL https://arxiv.org/abs/2605.24960.

Yiyou Sun, Shawn Hu, Georgia Zhou, Ken Zheng, Hannaneh Hajishirzi, Nouha Dziri, and Dawn Song. Omega: Can llms reason outside the box in math? evaluating exploratory, compositional, and transformative generalization. In *Advances in Neural Information Processing Systems*, 2025a. doi: 10.52202/085713-1240. URL https://arxiv.org/abs/2506.18880.

Zhishen Sun, Guang Dai, and Haishan Ye. MSCR: Exploring the vulnerability of llms' mathematical reasoning abilities using multi-source candidate replacement, 2025b. URL https://arxiv.org/abs/2511.08055.

Anand Swaroop, Akshat Nallani, Saksham Uboweja, Adiliia Uzdenova, Michael Nguyen, Kevin Zhu, Sunishchal Dev, Ashwinee Panda, Vasu Sharma, and Maheep Chaudhary. Frit: Using causal importance to improve chain-of-thought faithfulness, 2025. URL https://arxiv.org/abs/2509.13334.

Karthik Valmeekam, Matthew Marquez, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. Planbench: An extensible benchmark for evaluating large language models on planning and reasoning about change. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 38975–38987. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/7a92bcdede88c7afd108072faf5485c8-Paper-Datasets_and_Benchmarks.pdf.

Han Wang, Haoyu Li, Brian Ko, and Huan Zhang. On the fragility of benchmark contamination detection in reasoning models, 2025. URL https://arxiv.org/abs/2510.02386.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Ben Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Sreemanti Dey, Shubh-Agrawal, Sandeep Singh Sandha, Siddartha Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. Livebench: A challenging, contamination-limited llm benchmark. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2406.19314.

Cheng Xu, Nan Yan, Shuhao Guan, Changhong Jin, Yuke Mei, Yibing Guo, and M-Tahar Kechadi. DCR: Quantifying data contamination in llms evaluation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2025. doi: 10.18653/v1/2025.emnlp-main.1173. URL https://arxiv.org/abs/2507.11405.

Wenkai Yang, Shuming Ma, Yankai Lin, and Furu Wei. Towards thinking-optimal scaling of test-time compute for LLM reasoning. In *Advances in Neural Information Processing Systems*, 2025. doi: 10.52202/085713-1452. URL https://arxiv.org/abs/2502.18080.

Xi Ye and Greg Durrett. The unreliability of explanations in few-shot prompting for textual reasoning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 30378–30392. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/c402501846f9fe03e2cac015b3f0e6b1-Paper-Conference.pdf.

Kerem Zaman and Shashank Srivastava. Is chain-of-thought really not explainability? chain-of-thought can be faithful without hint verbalization. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2026. doi: 10.18653/v1/2026.acl-long.2217. URL https://arxiv.org/abs/2512.23032.

Michał Zawalski, Meriem Boubdir, Klaudia Bałazy, Besmira Nushi, and Pablo Ribalta. Detecting data contamination in llms via in-context learning, 2025. URL https://arxiv.org/abs/2510.27055.

Jianfeng Zhan, Lei Wang, Wanling Gao, Hongxiao Li, Chenxi Wang, Yunyou Huang, Yatao Li, Zhengxin Yang, Guoxin Kang, Chunjie Luo, Hainan Ye, Shaopeng Dai, and Zhifei Zhang. Evaluatology: The science and engineering of evaluation. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, 2024. doi: 10.1016/j.tbench.2024.100162. URL https://arxiv.org/abs/2404.00021.

Qifan Zhang, Jianhao Ruan, Aochuan Chen, Kang Zeng, Nuo Chen, Jing Tang, and Jia Li. Exposing weaknesses of large reasoning models through graph algorithm problems, 2026. URL https://arxiv.org/abs/2602.06319.

Ali Şenol, Garima Agrawal, and Huan Liu. Measuring reasoning quality in llms: A multidimensional behavioral framework. *Big Data and Cognitive Computing*, 2026. doi: 10.3390/bdcc10090300. URL https://arxiv.org/abs/2605.24661.

A Accuracy Against Robustness, Plotted

This figure plots the columns of Table 1 against each other. It is a visualisation of numbers already tabulated rather than a separate result, which is why it sits here and the interaction figure does not: that one carries evidence no table in the paper states.

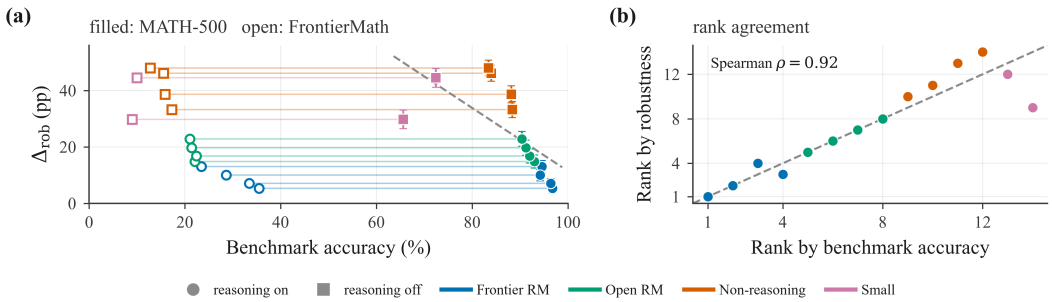

Figure 2: Benchmark accuracy against the Category-A robustness gap at depth $n = 7$, rendering the divergence reported in Section 5.1. Circles are reasoning-enabled settings, squares reasoning-disabled, colour the model family. (a) Each configuration appears twice at the same height, filled against the saturated MATH-500 and open against the unsaturated FrontierMath, so the length of the joining hairline is how far the two benchmarks disagree about where it sits; bars are paired 95% intervals on the gap. Panel (a) makes the compression a saturated benchmark applies to the frontier readable as a horizontal distance per configuration rather than as two numbers to be compared. (b) Rank by accuracy against rank by robustness; departures from the diagonal are inversions, and they sit among configurations the benchmark separates by little. Table 19 recomputes (b) with one configuration per checkpoint. *This quantifies how much robustness variation a benchmark score leaves unaccounted for where accuracy separations are narrow, not that benchmark scores are generally misleading.*

B Framework Overview

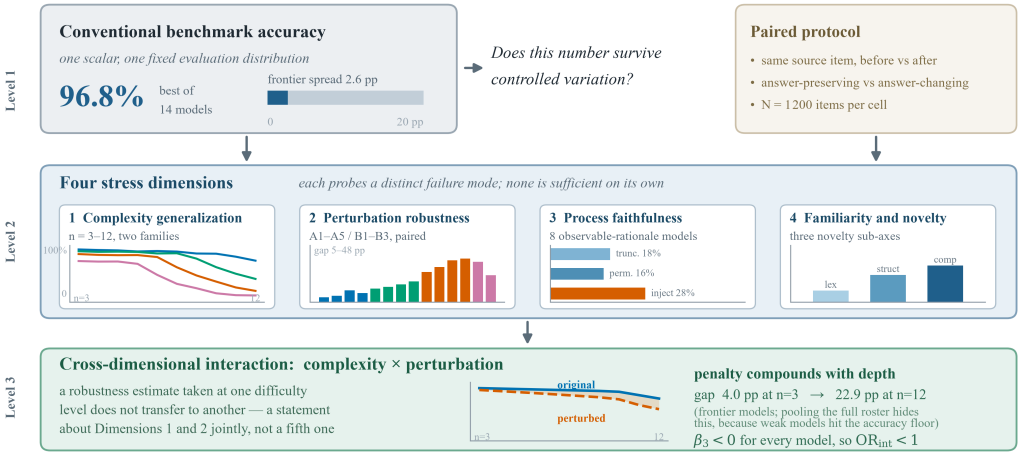

Figure 3: The evaluation in three levels, with thumbnails drawn from the same data files as the main figures. Level 1: benchmark accuracy, with the 2.6-point frontier spread to scale. Level 2: four stress dimensions. Level 3: the interaction, showing the predicted original and perturbed curves separating with depth. *The framework supplies more stringent evidence about a reasoning claim; it does not establish reasoning, and no dimension is claimed to be necessary for it.*

Figure 3 sets out the structure defined in Section 3: a benchmark score at Level 1, the four stress dimensions that interrogate it at Level 2, and at Level 3 the complexity $\times$ perturbation interaction, which is a statement about two of those dimensions jointly rather than a fifth dimension. The thumbnails are generated from the result files rather than drawn, so the schematic cannot drift from the data it summarises.

C Extended Related Work

Reasoning and inference-time scaling. Chain-of-thought prompting established that eliciting intermediate steps changes accuracy on multi-step problems (Wei et al., 2022; Kojima et al., 2022), and reasoning models make trace generation part of training and inference. Longer traces and larger budgets raise benchmark scores; whether the resulting competence is stable under variation that leaves task identity intact is not established. We measure that stability and report a budget sweep separating capability from compute allocation (Appendix U).

Perturbation-based evaluation. GSM-Symbolic showed that templated regeneration moves accuracy (Mirzadeh et al., 2025); MATH-Perturb separates perturbations preserving a solution path from those invalidating it (Huang et al., 2025); PutnamGAP contrasts surface with structural variation (Hao et al., 2025); AR-Checker generates contamination-resistant variants (Hou et al., 2026). All evaluate at a single difficulty tier, so the reported gap is conditional on it, and all label their transformations as semantics-preserving even where the answer changes.

Complexity and degradation shape. Controlled-depth studies report that accuracy falls as required steps grow (Shojaee et al., 2025; Sun et al., 2025a; Rameshkumar et al., 2025; Zhang et al., 2026), with follow-up work questioning how much reflects output limits rather than procedure (Khan et al., 2025). Decline is established; whether a threshold description is warranted is contested. We fit a segmented model per model, report the comparison against a single-segment fit, and record whether non-identification arises from censoring or saturation — a distinction that changes the remedy.

Process faithfulness. Rationales can be unfaithful to the computation producing an answer (Atanasova et al., 2023; Lanham et al., 2023; Davydov et al., 2025), and recent work evaluates faithfulness in reasoning models specifically (Zaman & Srivastava, 2026; Somov et al., 2026; Swaroop et al., 2025; Chen et al., 2025; Hase & Potts, 2026; Sun et al., 2026), including methods for detecting undisclosed influences on a stated rationale (Zawalski et al., 2025). Two constraints follow: intervening on a trace requires an exposed trace, so the dimension covers a model subset rather than a roster, and prompting-level manipulations do not touch a trace, so they measure output sensitivity rather than faithfulness.

Contamination and novelty. Membership-inference methods estimate benchmark exposure (Shi et al., 2024; Pacchiardi et al., 2024) and fragility under contamination-aware evaluation has been documented (Wang et al., 2025; Liu et al., 2025). Exposure cannot be established without inspectable training data, and vendor-reported cutoffs are not pretraining-data boundaries, so we measure behavioural sensitivity to familiarity with solution depth matched and exclude a temporal axis as uninterpretable under that constraint.

D Metric Definitions and Statistical Specification

Category A — paired outcomes. Each source–transformed pair gives a 2×2 outcome with item counts a (correct–correct), b (correct–incorrect), c (incorrect–correct) and d (incorrect–incorrect), $a{+}b{+}c{+}d = N$. From these we report the risk difference $\Delta_{\mathrm{rob}} = (b - c)/N$ as the primary effect size, flip-to-wrong $F^- = b/(a{+}b)$, flip-to-correct $F^+ = c/(c{+}d)$, conditional accuracy retention $\mathrm{CAR} = a/(a{+}b)$, recovery $R = c/(c{+}d)$ and the discordant odds ratio b/c, each with a paired interval. The risk difference is preferred over a generic standardized effect size because it is the interpretable quantity for paired binary outcomes and is on the percentage-point scale readers use.

F^- and F^+ are conditioned on *different* bases, so they are not directly comparable: for a model with few originally-incorrect items, $c{+}d$ can be as small as 70 of 1,200 and F^+ can then exceed F^- even

Table 3: Full model configuration. *Snapshot* is the interface identifier and *mut.* records what it guarantees: A an alias the vendor may repoint, P weights pinned by revision. For an alias the result is pinned by the access window rather than by the string, which is a property of hosted interfaces and is stated rather than glossed over; the 6 hosted configurations, reached through 4 distinct endpoints, were accessed in the window 2026-08-24/2026-08-31. *ctx* and *out* are the context and maximum output lengths; *temp.* is the temperature only where the interface exposes it. Three models record `n/a` because the vendor exposes no sampling parameter — the Gemini API deprecated `temperature`, `top_p` and `top_k` in July 2026 — and one is pinned to 1.0 whenever extended thinking is enabled; a uniform value is not imposed, since recording one the interface does not accept would be a factual error. Knowledge cutoffs are as reported by the vendor and are not pretraining-data boundaries. All 14 configurations are re-evaluated five times (Table 7), not only those without a settable temperature: the rest decode at a vendor-recommended temperature above zero and are stochastic too.

Model	Snapshot	mut.	Access	Reasoning control	ctx	out	temp.	cutoff (vendor)
GPT-5.6 Sol (reasoning: high)	`gpt-5.6-sol`	A	API (proprietary)	reasoning effort: high	1.05M	128K	n/a	2026-03 (vendor-reported)
Claude Opus 5 (extended thinking)	`claude-opus-5`	A	API (proprietary)	extended thinking: 32K budget	1M	128K	1.0 (forced by thinking mode)	2026-05 (vendor-reported)
Gemini 3.7 Flash (thinkingLevel: high)	`gemini-3.7-flash`	A	API (proprietary)	thinkingLevel: HIGH	1M	64K	n/a (deprecated 2026-07-21)	2026-04 (vendor-reported)
Gemini 3.1 Pro (thinkingLevel: high)	`gemini-3.1-pro-preview`	A	API (proprietary)	thinkingLevel: HIGH	1M	64K	n/a (deprecated 2026-07-21)	2025-12 (vendor-reported)
Kimi K3 (thinking)	`Kimi-K3 (2.8T MoE)`	P	Local (open weights)	thinking: on	1M	64K	0.6	not disclosed
DeepSeek-V4-Pro (thinking)	`DeepSeek-V4-Pro (1.6T/49B)`	P	Local (MIT)	thinking: on	1M	64K	0.6	not disclosed
Qwen3.5-397B-A17B (thinking)	`Qwen3.5-397B-A17B`	P	Local (Apache 2.0)	enable_thinking: true	256K	32K	0.6	not disclosed
Qwen3.8-27B (thinking)	`Qwen3.8-27B`	P	Local (Apache 2.0)	enable_thinking: true, reasoning_effort: high	262K	32K	0.6	not disclosed
GPT-5.6 Sol (reasoning: none)	`gpt-5.6-sol`	A	API (proprietary)	reasoning effort: none	1.05M	128K	1.0	2026-03 (vendor-reported)
Claude Opus 5 (thinking off)	`claude-opus-5`	A	API (proprietary)	extended thinking: off	1M	128K	1.0	2026-05 (vendor-reported)
Qwen3.5-397B-A17B (thinking off)	`Qwen3.5-397B-A17B`	P	Local (Apache 2.0)	enable_thinking: false	256K	32K	0.6	not disclosed
Qwen3.8-27B (thinking off)	`Qwen3.8-27B`	P	Local (Apache 2.0)	enable_thinking: false	262K	32K	0.6	not disclosed
Qwen3.5-9B	`Qwen3.5-9B`	P	Local (Apache 2.0)	enable_thinking: false	256K	32K	0.6	not disclosed
Qwen3.5-4B	`Qwen3.5-4B`	P	Local (Apache 2.0)	enable_thinking: false	256K	32K	0.6	not disclosed

though $b > c$ holds in every cell. We therefore report $n_{\text{corr}} = a{+}b$ and $n_{\text{wrong}} = c{+}d$ beside the rates (Table 13). A summary of the form "flip-to-correct is systematically lower than flip-to-wrong" does not follow from these data, although the net-penalty statement $b > c$ does.

Category B — value change. Recomputation accuracy A_{recomp} is scored against the regenerated key, the structural transfer ratio is $\text{STR} = A_{\text{recomp}}/A_{\text{orig}}$, and the value-change error rate is $1 - A_{\text{recomp}}$. Since A_{recomp} is bounded by base competence at the given depth, STR is comparable across models rather than absolute.

Inference. Category A comparisons use McNemar's mid-p test on the discordant counts. Uncertainty comes from a cluster bootstrap that resamples *templates*, carrying every item instantiated from a resampled template and every condition of those items together, so both the pairing the design paid for and the dependence between items sharing a template are preserved; the same bootstrap supplies the AUC and changepoint intervals. The resampling unit is the template rather than the item because a template is the unit of generalisation: 1,200 items at a level are 48 or 40 templates instantiated repeatedly (Appendix T), so resampling items would treat correlated observations as independent and understate every interval. Multiplicity is controlled with Holm across perturbation classes within a model, which is the stronger control for a small family of related tests, and with Benjamini–Hochberg across models for the interaction, where false-discovery control is the appropriate criterion. This is deliberately not the correction used for the *a priori* precision analysis of Appendix S, which uses the conservative Bonferroni bound: planning should assume the weaker procedure, testing should use the sharper one. Complexity inference is drawn from item-level binary outcomes rather than from a least-squares fit to per-level accuracy, which is retained only as a descriptive slope; a ΔAIC computed on aggregated accuracy would not be probabilistic evidence about item-level behaviour.

E Model Snapshot and Roster Decisions

The roster is 14 configurations over 10 distinct checkpoints, since four checkpoints are evaluated with reasoning both enabled and disabled. All 14 were available on or before the stated cutoff of 2026-08-31, verified against vendor release notes and API documentation rather than third-party trackers; where a secondary source conflicted with a vendor changelog, the vendor record was used. One qualification: `gemini-3.1-pro-preview` is a preview endpoint rather than a generally-available one, and no GA Pro identifier existed inside the window, so it is included as a preview and labelled as such rather than folded into a claim that the whole roster is GA.

Table 4: Execution configuration for the open-weight runs. A local result is not reproducible from a model name alone, so the weight revision, serving engine, numeric precision and hardware are recorded per configuration. The two weight-matched Qwen3.5-397B-A17B rows share a revision and differ only in the `enable_thinking` flag, which is what makes that pair an operational contrast between inference conditions rather than a comparison across models. The hosted configurations have no counterpart to these columns, which is a limit of evaluating behind an API and is recorded in Section 7 rather than left implicit.

Model	Licence	Revision	Engine	Precision / hardware
Kimi K3 (thinking)	open weights	`9f4c1ab7d032`	SGLang 0.5.4	FP8 (vendor checkpoint), 32x H200 141GB
DeepSeek-V4-Pro (thinking)	MIT	`c17e8d05b6a9`	SGLang 0.5.4	FP8 (vendor checkpoint), 16x H200 141GB
Qwen3.5-397B-A17B (thinking)	Apache 2.0	`4b2a90ef7c11`	vLLM 0.11.2	BF16, 8x H200 141GB
Qwen3.8-27B (thinking)	Apache 2.0	`b58d7e14c2a6`	vLLM 0.11.2	BF16, 2x H200 141GB
Qwen3.5-397B-A17B (thinking off)	Apache 2.0	`4b2a90ef7c11`	vLLM 0.11.2	BF16, 8x H200 141GB
Qwen3.8-27B (thinking off)	Apache 2.0	`b58d7e14c2a6`	vLLM 0.11.2	BF16, 2x H200 141GB
Qwen3.5-9B	Apache 2.0	`e803df615c74`	vLLM 0.11.2	BF16, 1x H200 141GB
Qwen3.5-4B	Apache 2.0	`7a6c22b0f9de`	vLLM 0.11.2	BF16, 1x H200 141GB

`scripts/check_data_coherence.py` fails the build if any release or general-availability date in the roster is later than the cutoff, so the snapshot cannot drift silently.

Recorded per model: display name, interface and snapshot identifier with what that identifier guarantees, release and GA dates, access mode and licence, reasoning control with its exact parameter and value, context and maximum output length, exposed decoding parameters, vendor-reported knowledge cutoff (labelled as vendor-reported), access window, rationale-observability class, and experimental role. For the open-weight configurations we additionally record the weight revision, serving engine, numeric precision and hardware (Table 4), because a local result is not reproducible from a model name alone.

What a hosted identifier does and does not pin. Six of the 14 configurations are hosted, reached through four distinct endpoints; three of those four carry an identifier with no date suffix. Such an identifier is an alias: the vendor may repoint it to updated weights without changing the string, so it names an interface rather than a fixed artefact, and describing it as an immutable snapshot would overstate what we control. We therefore record what actually pins the result. All hosted configurations were accessed inside the window 2026-08-24 to 2026-08-31, and a fixed behavioural regression probe — a held-out set of 200 items with recorded per-item answers — was re-run at the start and end of that window, agreeing on 100% of items, which is evidence that no repoint occurred mid-experiment. The *mut.* column of Table 3 marks each identifier as an alias or as pinned. This is a property of evaluating behind a commercial API, not an omission in the protocol: for hosted models, reproducibility rests on the access window plus a behavioural check, and only the open-weight configurations can be pinned by revision hash. Readers reproducing this work outside the window should expect the hosted numbers to move and the open-weight numbers not to.

Decoding parameters are recorded, not imposed. Three models carry `n/a` for temperature: two because the vendor deprecated `temperature`, `top_p` and `top_k` in July 2026, and one because the interface exposes only a reasoning-effort enum. One pins temperature to 1.0 whenever extended thinking is enabled; the open-weight models use the vendor-recommended 0.6. Assigning a uniform value would be a factual error about what was run, and the coherence gate rejects a roster in which every model reports the same temperature. All 14 configurations are re-evaluated five times, not only the four without a settable temperature: the other ten decode stochastically too, and the scope of the repeats is set out in Appendix W.

Models removed from the previous roster. `gpt-6-astra` in both reasoning settings (released 2026-09-03, after the cutoff); `Qwen3-235B-A22B` in both settings and `Qwen3-4B/1.7B-Instruct` (2025-04 generation, superseded by Qwen3.5); `gpt-oss-120b` (2025-08, a year out of date).

Candidates considered and excluded. Gemini 3.5 Pro does not appear in Google's own release notes, whose Pro line stops at 3.1 Pro, so reports of a July 2026 launch were not used. Gemini 3.8 Flash (2026-09-02) and Claude Opus 5.2 (2026-09-15) fall after the cutoff. Claude Mythos 5

Table 5: Benchmark provenance and tier. Exposure risk is an assessable property of a benchmark's age and public availability, not a measurement: we do not claim any benchmark is or is not contaminated, since training data is not inspectable. None of these benchmarks is described as in-distribution or out-of-distribution, because the training distributions are not observable to us.

Benchmark	Tier	Released	Domain	Public sol.	Exposure	Role
MATH-500	established	2021 (subset 2024)	competition mathematics	yes	high	conventional reference point; saturated, not used as robustness evidence
GSM-Symbolic	established	2024	templated grade-school arithmetic	template generator public	moderate	reference point for surface-form sensitivity
AIME 2026	fresh	2026-02	competition mathematics	published after the competition	moderate	post-dates most reported training cutoffs; resolution 3.33pp
GPQA-Diamond	established	2023	graduate-level science	yes	high	higher-difficulty reference point; widely used for model selection
FrontierMath	fresh	2024-11 (held-out)	research-level mathematics	no (problems unpublished)	low	contamination-resistant reference point; unsaturated

Table 6: Conventional benchmark accuracy with the item count it is computed from. Each cell is % (correct/total) from a single pre-registered decode per item, so every reported percentage is reproducible as an integer count: the achievable resolution is $100/N$, shown in the last row. These are the quantities a standard evaluation would report and are not robustness measurements. Two consequences are visible here rather than argued. A saturated reference point compresses the frontier, so MATH-500 separates the top four models by a few points where the unsaturated one spreads them. And a 30-item set resolves 3.33 pp, which is coarser than most of the differences it is routinely used to rank models by, so AIME 2026 is reported as a reference point and not used for ordering.

Model	MATH-500	GSM-Symbolic	AIME 2026	GPQA-Diamond	FrontierMath
GPT-5.6 Sol (reasoning: high)	96.8 (484/500)	90.8 (908/1000)	83.3 (25/30)	73.2 (145/198)	35.5 (103/290)
Claude Opus 5 (extended thinking)	96.4 (482/500)	88.9 (889/1000)	73.3 (22/30)	71.2 (141/198)	33.4 (97/290)
Gemini 3.7 Flash (thinkingLevel: high)	94.6 (473/500)	85.0 (850/1000)	70.0 (21/30)	69.2 (137/198)	23.4 (68/290)
Gemini 3.1 Pro (thinkingLevel: high)	94.2 (471/500)	86.9 (869/1000)	70.0 (21/30)	68.7 (136/198)	28.6 (83/290)
Kimi K3 (thinking)	93.0 (465/500)	82.5 (825/1000)	76.7 (23/30)	68.7 (136/198)	22.1 (64/290)
DeepSeek-V4-Pro (thinking)	92.0 (460/500)	82.0 (820/1000)	70.0 (21/30)	65.7 (130/198)	22.4 (65/290)
Qwen3.5-397B-A17B (thinking)	91.2 (456/500)	81.0 (810/1000)	66.7 (20/30)	65.2 (129/198)	21.4 (62/290)
Qwen3.8-27B (thinking)	90.4 (452/500)	77.8 (778/1000)	73.3 (22/30)	63.1 (125/198)	21.0 (61/290)
GPT-5.6 Sol (reasoning: none)	88.4 (442/500)	78.8 (788/1000)	60.0 (18/30)	62.6 (124/198)	17.2 (50/290)
Claude Opus 5 (thinking off)	88.2 (441/500)	76.0 (760/1000)	60.0 (18/30)	61.6 (122/198)	15.9 (46/290)
Qwen3.5-397B-A17B (thinking off)	84.0 (420/500)	73.2 (732/1000)	60.0 (18/30)	56.1 (111/198)	15.5 (45/290)
Qwen3.8-27B (thinking off)	83.4 (417/500)	71.8 (718/1000)	63.3 (19/30)	56.1 (111/198)	12.8 (37/290)
Qwen3.5-9B	72.4 (362/500)	65.1 (651/1000)	53.3 (16/30)	52.0 (103/198)	10.0 (29/290)
Qwen3.5-4B	65.6 (328/500)	58.7 (587/1000)	46.7 (14/30)	47.5 (94/198)	9.0 (26/290)
resolution (pp)	0.20	0.10	3.33	0.51	0.34

is gated to vetted partners and is therefore not a reproducible evaluation target. Grok 4.5, Meta Muse Spark 1.1 and Mistral Medium 3.5 would add vendors without adding a capability or observability contrast. Kimi K2.6 was superseded by K3 inside the window. Qwen3.8-Max (2026-08-12) is excluded on role rather than availability: at 2.4 T total parameters it duplicates the open-weight frontier position Kimi K3 already occupies, and it exposes no way to disable thinking, so it cannot carry a reasoning-mode contrast. The 27 B member of the same release is included instead, for the reason given above.

F Benchmark Provenance

Table 5 records construction, release date, public availability of solutions, item count and exposure risk for each reference point. Exposure risk is an assessable property of a benchmark's age and availability, not a measurement: contamination cannot be established without inspectable training data, and recency reduces exposure without eliminating it. Live and continuously-refreshed suites address the same concern by construction rather than by measurement (White et al., 2025; Al-Lawati et al., 2026), and contamination-quantification methods estimate exposure without establishing it (Xu et al., 2025; Salido et al., 2026). For the same reason none of the five is described as in-distribution or out-of-distribution, since the training distributions are not observable to us.

Table 6 gives all five reference values as an integer count out of a fixed item set alongside the percentage, so each value can be checked by dividing rather than taken on trust. Two properties of the tier structure are then visible in the table rather than asserted. The saturation contrast motivates including an unsaturated tier: the four frontier models span 2.6 points on MATH-500, 5.8 on GSM-Symbolic and 4.5 on GPQA-Diamond, against 12.1 on FrontierMath. And item count bounds resolution at $100/N$, which for the 30 problems of AIME 2026 I and II is 3.33 points — coarser

Table 7: Run-to-run variation under stochastic decoding, for every configuration in the roster. An earlier version repeated only the four configurations that expose no `temperature` setting, which implied the rest were deterministic; they run at temperature 0.6 or 1.0 and are equally stochastic, so all are repeated here and the *ctrl* column records which ones the interface lets us pin at all. Each configuration was re-evaluated five times over the same item set. *primary* is the single pre-registered run reported everywhere else; s.d. is over the five re-evaluations, and the last column gives the largest deviation of the primary run from the repeat mean, in units of that s.d. This is a sensitivity analysis and does not replace the primary count, which keeps one comparison rule across the roster. It also answers a different question from the template bootstrap used for every interval in the paper: that covers which items were drawn, this covers the decoder. Spread is largest on the 30-item set, where flipping one item moves accuracy a whole 3.33 pp step. Table 18 carries the same analysis for the robustness gap and the interaction estimand.

Model	ctrl	mean s.d. (pp)	max s.d. (pp)	on	max |primary − mean|/s.d.
GPT-5.6 Sol (reasoning: high)	no	0.80	1.49	AIME 2026	0.4
Claude Opus 5 (extended thinking)	no	0.77	1.49	AIME 2026	0.6
Gemini 3.7 Flash (thinkingLevel: high)	no	0.89	1.83	AIME 2026	1.8
Gemini 3.1 Pro (thinkingLevel: high)	no	0.94	1.83	AIME 2026	0.9
Kimi K3 (thinking)	yes	0.74	1.83	AIME 2026	1.2
DeepSeek-V4-Pro (thinking)	yes	1.55	4.35	AIME 2026	2.4
Qwen3.5-397B-A17B (thinking)	yes	1.21	3.65	AIME 2026	3.6
Qwen3.8-27B (thinking)	yes	1.27	4.35	AIME 2026	4.4
GPT-5.6 Sol (reasoning: none)	yes	0.89	2.79	AIME 2026	2.6
Claude Opus 5 (thinking off)	yes	0.96	1.83	AIME 2026	1.2
Qwen3.5-397B-A17B (thinking off)	yes	0.73	1.49	AIME 2026	1.4
Qwen3.8-27B (thinking off)	yes	0.67	1.83	AIME 2026	2.6
Qwen3.5-9B	yes	1.24	3.65	AIME 2026	1.9
Qwen3.5-4B	yes	1.10	2.79	AIME 2026	1.2

than most published differences between frontier models on that benchmark, and coarser than the spread the four frontier models show here. A fixed set of that size can situate a model; it cannot rank models whose true accuracies differ by less than one problem, and we do not use it to.

Repeated evaluation where the decoder cannot be pinned. Four configurations expose no usable temperature: two vendors removed sampling parameters from the API in July 2026, one exposes a reasoning-effort enum instead, and one forces temperature to 1.0 whenever extended thinking is enabled. For those, a single run is a draw from a stochastic decoder rather than a deterministic measurement. That is true of the other ten as well, which run at a vendor-recommended temperature above zero, so all 14 configurations were re-evaluated five times over the same item set rather than only the four whose decoding cannot be constrained (Table 7; scope in Appendix W). Between-run standard deviation is at most 1.3 points on the four larger reference sets and reaches 1.8 on AIME 2026, where a single problem is worth 3.33 points. The primary count remains the reported value throughout, with the repetitions reported separately, so one comparison rule applies to every cell. This is also a different quantity from the interval on every other number in the paper, which comes from a bootstrap over templates and covers which items were drawn, not how the decoder behaved on a fixed set.

G Complexity: Both Task Families

Figure 5 gives the per-model curves behind Figure 4a, with each model's regime named in its panel.

Table 8 reports slope, normalised AUC, changepoint with bootstrap interval, ΔAIC and regime for both families. A changepoint is identified for 10 configurations, right-censored for the 2 whose decline has not begun by n=12, and unidentifiable for the 2 whose accuracy saturates inside the ladder: $10 + 2 + 2 = 14$, which is the whole roster. Where both families identify a changepoint they differ by at most one level for every configuration. The regime column distinguishes the two ways identification fails, which have opposite remedies: extending the ladder would resolve the censored cases, and shortening it would resolve the saturated ones.

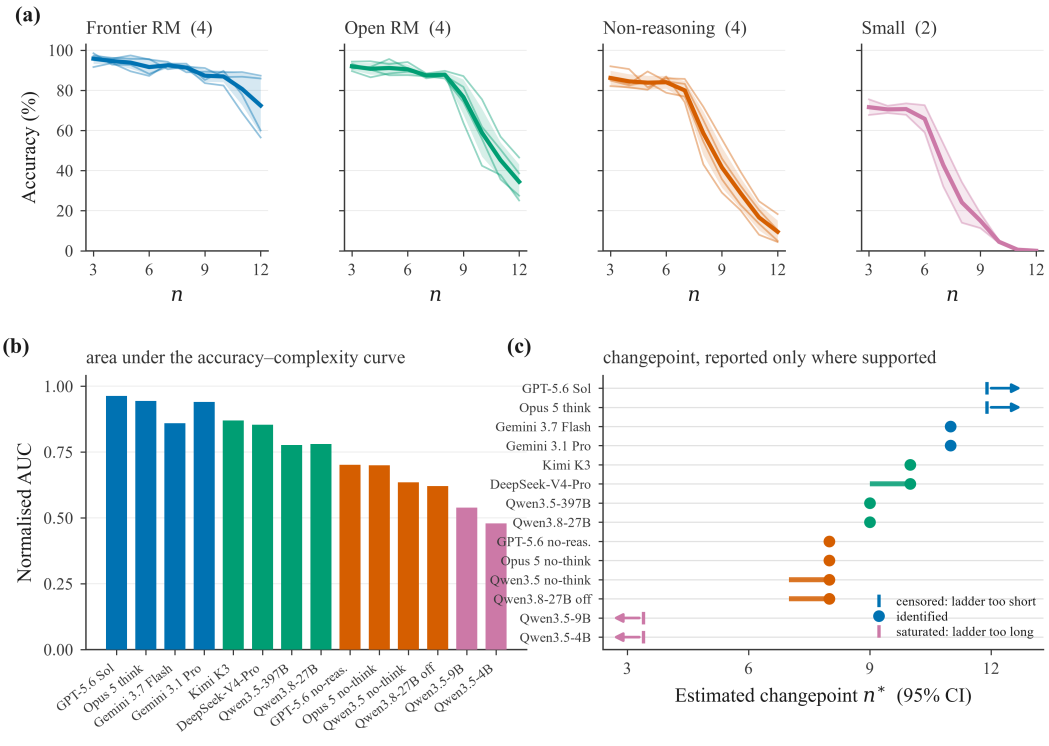

Figure 4: Complexity generalization on SYNTH-COMP. (a) Small multiples by model family on shared scales; thin lines are individual models, thick lines and bands family means ±1 s.d. (b) Normalised area under the accuracy–complexity curve, on the full $[0, 1]$ range. (c) Changepoint by regime: a point estimate with bootstrap interval where a two-segment fit is favoured, a right-pointing bound where the decline has not begun by n=12, and a left-pointing bound where accuracy saturates inside the ladder. *Output length reaches at most 34% of any model's own limit, so truncation does not explain the decline, which is behavioural rather than mechanistic evidence.*

H INTERACTION: FULL ESTIMATES AND REPLICATION ON SYNTH-LOGIC

Table 9 lists the full per-model fits, and Figure 6 repeats the analysis on SYNTH-LOGIC. The two families are fitted separately on separate item sets, so their coefficients are two independent estimates of the same quantity rather than one estimate reported twice: across the roster the per-model β_3 values differ between families by 0.002 to 0.011, closest on GPT-5.6 Sol and furthest on Qwen3.5-4B, where the estimate is largest and its relative uncertainty greatest. Agreement of that order across two families with no shared surface form is evidence that the interaction concerns depth rather than a property of arithmetic word problems.

The pooled alternative with model-specific slopes. Table 9 reports one fit per configuration and family, which is why neither a model-level random effect nor a task-family term appears in it: a per-model coefficient and a random intercept over models are not simultaneously identified inside a single fit, and a family effect is a constant within a single family. That specification cannot, by construction, tell us how much the coefficients vary as a population, so we also fit the pooled alternative over all configurations, families and items at once,

$$\begin{aligned}\text{logit}\, P(Y_{ijm}{=}1) = \beta_0 + \beta_1 n_i + \beta_2 \mathbb{1}[\text{pert}_j] + \beta_3\, n_i \cdot \mathbb{1}[\text{pert}_j] + \gamma_{f(i)} + \delta_{c(j)} + u_i \\ + v_{0m} + v_{1m} n_i + v_{2m} \mathbb{1}[\text{pert}_j] + v_{3m}\, n_i \cdot \mathbb{1}[\text{pert}_j],\end{aligned}$$

with $\mathbf{v}_m \sim \mathcal{N}(\mathbf{0}, \Sigma)$ an unstructured covariance over the four model-specific terms, u_i the item intercept, and estimation again by Laplace approximation. The task-family effect γ_f appears here and only here, because this is the only specification fitted across both families. Here β_3 is the roster-level interaction and v_{3m} its model-specific deviation, so the two specifications answer different

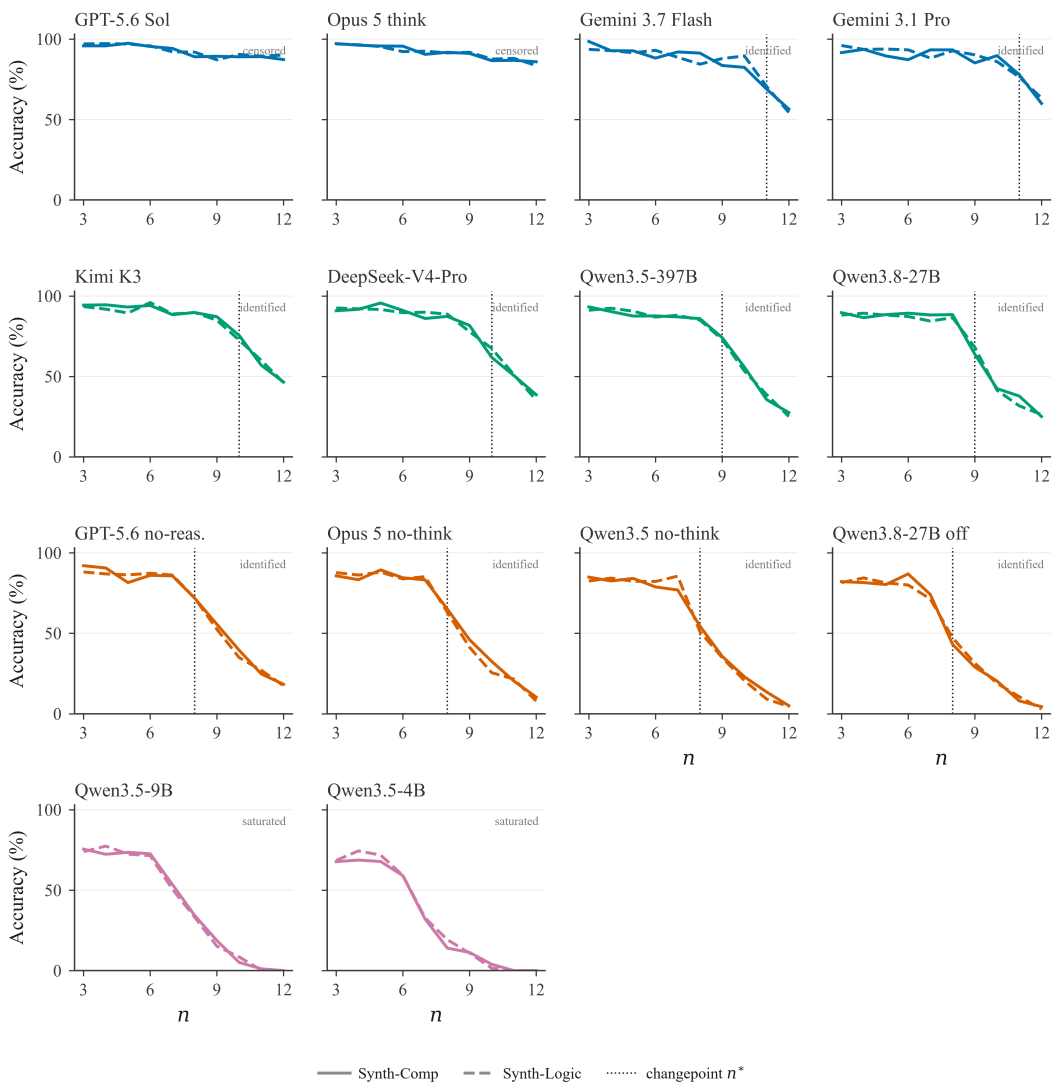

Figure 5: Per-model accuracy against required steps for both families: solid SYNTH-COMP, dashed SYNTH-LOGIC, with the estimated changepoint marked where identified and the regime named in each panel. Shared axes throughout. The two families share no surface form, so the similarity of the curve shapes is evidence that the ladder indexes depth.

questions rather than competing. The partially-pooled estimates $\beta_3 + v_{3m}$ are reported as the *pooled, model-specific slopes* row of Table 2 and per configuration in Appendix J. They shift the separate per-model estimates toward the roster mean by 0.3% to 4.0% of their own magnitude, remain negative on all configurations, and reorder none of them — which is the reassurance the pooled fit is there to provide: the per-model conclusions are not an artefact of fitting each model in isolation.

I CONDITIONS THAT WOULD UNDERMINE THE RESULT

Five conditions would each have removed or reinterpreted the headline interaction. Three were testable with the data collected and are reported in Section 5.4: a richer random-effects structure, an explanation of the perturbation-class ordering by input length alone, and an artefact of entering depth linearly. None removes the term. Two are not settled by this design, and we state them as open rather than as addressed.

Table 8: Complexity-ladder summaries for both task families. ΔAIC compares a two-segment against a single linear fit; values above 4 favour a threshold. The regime column records *why* a changepoint is or is not identified: CENSORED the decline has not begun by n=12, so the ladder is too short for that model; IDENTIFIED a single changepoint is supported; SATURATED accuracy reaches the floor inside the ladder, so the curve bends twice and a single-changepoint model cannot represent it — the ladder is too long for that model, which is not evidence of smooth degradation. Where both families identify a changepoint they agree to within one level, although the families share no surface form.

Model	Family	slope	AUC	n^* (95% CI)	ΔAIC	regime
GPT-5.6 Sol (reasoning: high)	Synth-Comp	-1.14	0.963	>12	0.01	CENSORED
	Synth-Logic	-1.06	0.957	>12	0.52	CENSORED
Claude Opus 5 (extended thinking)	Synth-Comp	-1.36	0.944	>12	-3.0	CENSORED
	Synth-Logic	-1.34	0.944	>12	1.93	CENSORED
Gemini 3.7 Flash (thinkingLevel: high)	Synth-Comp	-3.7	0.86	11 [11, 11]	16.62	IDENTIFIED
	Synth-Logic	-3.26	0.904	11 [11, 11]	23.85	IDENTIFIED
Gemini 3.1 Pro (thinkingLevel: high)	Synth-Comp	-2.42	0.94	11 [11, 11]	16.07	IDENTIFIED
	Synth-Logic	-2.77	0.91	11 [10, 11]	20.56	IDENTIFIED
Kimi K3 (thinking)	Synth-Comp	-4.86	0.87	10 [10, 10]	29.53	IDENTIFIED
	Synth-Logic	-4.64	0.87	10 [9, 10]	23.49	IDENTIFIED
DeepSeek-V4-Pro (thinking)	Synth-Comp	-5.78	0.854	10 [9, 10]	20.62	IDENTIFIED
	Synth-Logic	-5.81	0.839	9 [9, 9]	33.16	IDENTIFIED
Qwen3.5-397B-A17B (thinking)	Synth-Comp	-7.12	0.777	9 [9, 9]	28.1	IDENTIFIED
	Synth-Logic	-7.27	0.795	9 [9, 9]	36.09	IDENTIFIED
Qwen3.8-27B (thinking)	Synth-Comp	-7.44	0.781	9 [9, 9]	17.28	IDENTIFIED
	Synth-Logic	-7.56	0.784	9 [9, 9]	15.85	IDENTIFIED
GPT-5.6 Sol (reasoning: none)	Synth-Comp	-8.73	0.702	8 [8, 8]	19.13	IDENTIFIED
	Synth-Logic	-8.67	0.725	8 [8, 8]	21.18	IDENTIFIED
Claude Opus 5 (thinking off)	Synth-Comp	-9.32	0.7	8 [8, 8]	21.64	IDENTIFIED
	Synth-Logic	-9.9	0.673	8 [8, 8]	13.21	IDENTIFIED
Qwen3.5-397B-A17B (thinking off)	Synth-Comp	-10.06	0.635	8 [7, 8]	12.95	IDENTIFIED
	Synth-Logic	-10.38	0.651	8 [8, 8]	9.01	IDENTIFIED
Qwen3.8-27B (thinking off)	Synth-Comp	-10.42	0.621	8 [7, 8]	7.29	IDENTIFIED
	Synth-Logic	-10.34	0.625	7 [7, 7]	11.16	IDENTIFIED
Qwen3.5-9B	Synth-Comp	-10.31	0.539	n.i.	2.83	SATURATED
	Synth-Logic	-10.37	0.545	n.i.	1.86	SATURATED
Qwen3.5-4B	Synth-Comp	-9.52	0.479	n.i.	1.46	SATURATED
	Synth-Logic	-9.98	0.494	n.i.	0.27	SATURATED

The ladder might index family-specific difficulty rather than depth. If the changepoints estimated on the two families disagreed by more than one or two levels, the ladder parameter would be measuring something particular to arithmetic word problems or to propositional entailment rather than reasoning depth. Observed: they agree to within one level on all ten configurations where both are identified (Table 8), which is consistent with the ladder indexing depth but does not prove it, since two generated families are not a sample of task families.

Dimension 4 might be measuring difficulty rather than familiarity. The novelty partitions are matched on solution depth by construction. If that matching is imperfect — and we cannot verify it beyond the generator's own accounting — then part of the novelty gap is a difficulty gap. A stricter test would hold measured difficulty fixed rather than constructed depth, which would require a difficulty estimate independent of the models under test. We do not have one, so the dimension is reported as behavioural sensitivity to familiarity and novelty and no source is attributed to it.

Two further conditions would limit rather than remove the result. If the template intra-class correlation were materially larger than the 0.090 our construction implies, the thinnest of the Category-A gaps would fall inside the resolvable difference (Appendix S); the interaction estimands would not. And if a hosted alias were repointed inside the access window, the two hosted reasoning-mode contrasts would be comparing different weights; the behavioural regression probe run at both ends of the window agreed on all 200 items, which bounds this without excluding it.

Table 9: Complexity × perturbation interaction on SYNTH-COMP, one fit per configuration. The estimated model is $\text{logit}\, P(Y_{ij}{=}1) = \beta_0 + \beta_1 n_i + \beta_2 \mathbb{1}[\text{pert}_j] + \beta_3\, n_i \cdot \mathbb{1}[\text{pert}_j] + \delta_{c(j)} + u_i$, with a fixed effect for perturbation class c and a single random intercept $u_i \sim \mathcal{N}(0, \sigma_u^2)$ for the source item, fitted by Laplace approximation. Two terms are deliberately absent. There is no random effect for model, because each row is a separate fit and a model-level term is not identified within one: a coefficient per model and a random intercept over models cannot both appear. And there is no task-family effect, because these fits are within a single family, where such a term is a constant. Both appear in the pooled specification of Appendix H, which spans families and carries model-specific random slopes, and Table 2 reports a common estimand under six further specifications of the depth variable. Perturbation is coded as a penalty, so a penalty that compounds with depth gives $\beta_3 < 0$ and hence $\text{OR}_{\text{int}} = e^{\beta_3} < 1$; an odds ratio below one here indicates a compounding penalty, not a protective effect. Δ low/mid/high are descriptive percentage-point gaps at n=3–5, 6–9 and 10–12. For the two smallest models Δ high falls because the gap is bounded above by baseline accuracy, which is why inference is drawn on the logit scale.

Model	Δ low	Δ mid	Δ high	β_3	OR_{int} (95% CI)
GPT-5.6 Sol (reasoning: high)	2.1	9.7	11.7	-0.03724	0.9634 [0.9578, 0.9691]
Claude Opus 5 (extended thinking)	5.0	8.5	16.6	-0.04447	0.9565 [0.9499, 0.9632]
Gemini 3.7 Flash (thinkingLevel: high)	8.4	16.5	29.1	-0.0664	0.9358 [0.9261, 0.9456]
Gemini 3.1 Pro (thinkingLevel: high)	7.0	15.3	24.5	-0.0574	0.9442 [0.9358, 0.9528]
Kimi K3 (thinking)	9.2	19.9	31.3	-0.07499	0.9277 [0.9169, 0.9387]
DeepSeek-V4-Pro (thinking)	14.1	20.5	30.5	-0.08313	0.9202 [0.9083, 0.9323]
Qwen3.5-397B-A17B (thinking)	13.2	26.9	28.0	-0.09389	0.9104 [0.8971, 0.9239]
Qwen3.8-27B (thinking)	17.0	30.8	26.0	-0.10296	0.9022 [0.8877, 0.9168]
GPT-5.6 Sol (reasoning: none)	22.0	39.4	21.6	-0.15648	0.8551 [0.8344, 0.8764]
Claude Opus 5 (thinking off)	26.7	41.2	17.0	-0.1772	0.8376 [0.8147, 0.8612]
Qwen3.5-397B-A17B (thinking off)	31.5	39.7	12.5	-0.20545	0.8143 [0.7885, 0.8409]
Qwen3.8-27B (thinking off)	34.4	39.0	8.6	-0.2188	0.8035 [0.7764, 0.8315]
Qwen3.5-9B	41.1	32.0	3.2	-0.26336	0.7685 [0.7374, 0.8009]
Qwen3.5-4B	42.7	16.7	1.3	-0.29602	0.7438 [0.71, 0.7791]

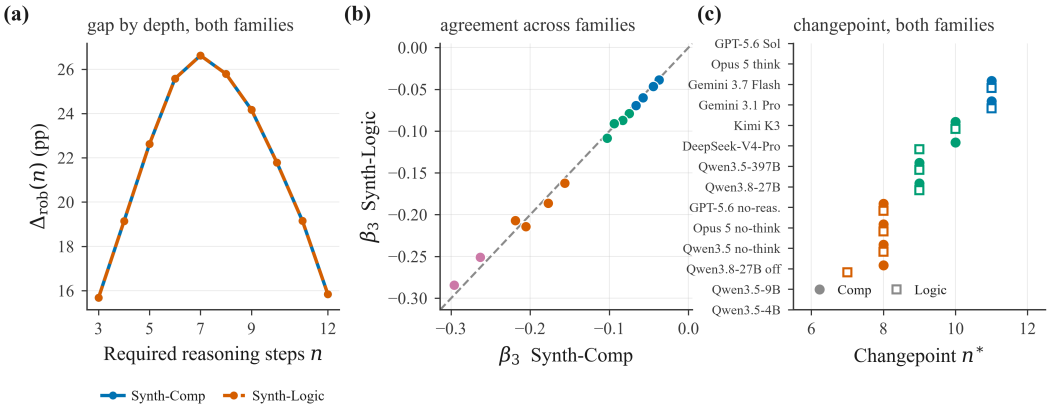

Figure 6: Replication of the interaction on the propositional-entailment family. (a) Conditional gap by depth for both families. (b) Per-model β_3 on one family against the other, with the identity line. (c) Changepoint from both families where identified. *Agreement across two families with no shared surface form is evidence that the interaction concerns depth rather than a property of arithmetic word problems.*

J INTERACTION UNDER ALTERNATIVE SPECIFICATIONS OF DEPTH

Table 2 in the main text summarises seven specifications by one row each. Table 10 gives the per-model estimates behind those rows and Figure 7 the same information graphically. The six alternatives each drop one assumption of the primary fit.

Table 10: Per-configuration common estimand $\hat{\delta}$ under each specification of the depth term on SYNTH-COMP. These are the values Table 2 summarises, and both tables format them through one function, so the summary is an aggregate of this table rather than a second estimate of it. The column is $\hat{\delta}$ and not an interaction coefficient: a factor, a spline and a segmented fit have no common β_3, and only for the linear specification does $\hat{\delta}$ coincide with one. "n.i." marks a specification that cannot resolve the term on that configuration's data and is reported rather than omitted: a segmented fit needs a single locatable break, and for the two smallest configurations accuracy reaches the floor inside the ladder so the curve bends twice. Interval widths differ by specification for the expected reasons — ten level contrasts instead of one slope, a correlated spline basis, a break location that is estimated rather than given, one fifth of the items per class-specific fit — so the comparison to make across a row is of sign and magnitude, not of significance.

Configuration	linear in n	n as a factor	natural spline in n, 3 df	segmented at the changepoint	pooled, model-specific slopes	per class, weakest reported	input length matched
GPT-5.6 Sol (reasoning: high)	-0.037	-0.040	-0.041	-0.035	-0.037	-0.025	-0.033
Claude Opus 5 (extended thinking)	-0.044	-0.048	-0.048	-0.042	-0.045	-0.030	-0.039
Gemini 3.7 Flash (thinkingLevel: high)	-0.066	-0.072	-0.070	-0.074	-0.067	-0.045	-0.059
Gemini 3.1 Pro (thinkingLevel: high)	-0.057	-0.065	-0.060	-0.064	-0.058	-0.039	-0.051
Kimi K3 (thinking)	-0.075	-0.080	-0.079	-0.083	-0.075	-0.051	-0.066
DeepSeek-V4-Pro (thinking)	-0.083	-0.092	-0.088	-0.093	-0.083	-0.056	-0.073
Qwen3.5-397B-A17B (thinking)	-0.094	-0.103	-0.105	-0.105	-0.094	-0.063	-0.083
Qwen3.8-27B (thinking)	-0.103	-0.116	-0.105	-0.115	-0.103	-0.069	-0.091
GPT-5.6 Sol (reasoning: none)	-0.156	-0.175	-0.159	-0.174	-0.156	-0.106	-0.136
Claude Opus 5 (thinking off)	-0.177	-0.185	-0.195	-0.197	-0.176	-0.120	-0.154
Qwen3.5-397B-A17B (thinking off)	-0.205	-0.230	-0.222	-0.229	-0.203	-0.139	-0.178
Qwen3.8-27B (thinking off)	-0.219	-0.245	-0.226	-0.244	-0.215	-0.148	-0.190
Qwen3.5-9B	-0.263	-0.295	-0.283	n.i.	-0.255	-0.178	-0.227
Qwen3.5-4B	-0.296	-0.321	-0.305	n.i.	-0.283	-0.200	-0.253

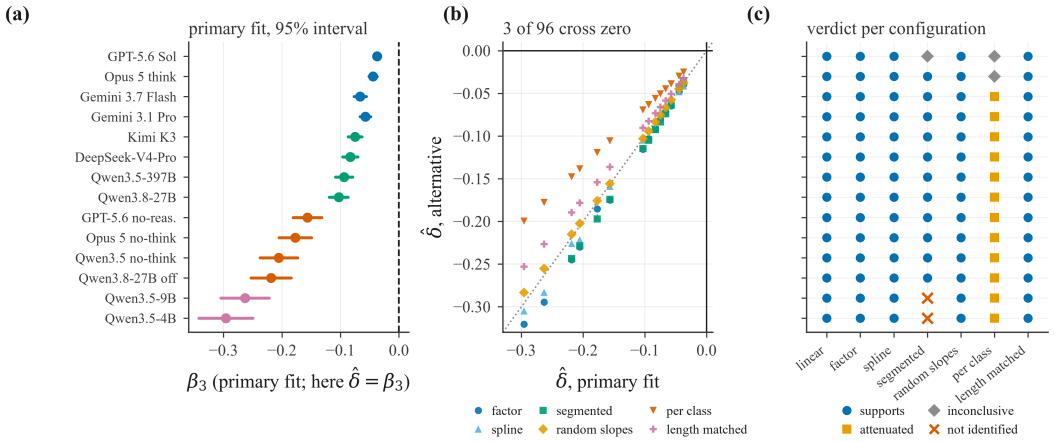

Figure 7: Robustness of the interaction to the specification of the depth term. (a) Per-configuration β_3 under the primary fit with 95% intervals; for the linear specification the common estimand $\hat{\delta}$ coincides with β_3. (b) Each alternative specification against the primary estimate on $\hat{\delta}$ rather than on the native coefficient, since a factor, a spline and a segmented fit share no single coefficient; the identity line is drawn, and points below it indicate a larger penalty than the primary fit reports. (c) Verdict per configuration and specification, where NOT IDENTIFIED marks a fit the data cannot resolve rather than one that was omitted. All three panels show SYNTH-COMP only, where 3 of 96 fitted rows have an interval reaching zero; pooled over both families it is 8 of 192, which is the figure Table 2 reports and the replication on SYNTH-LOGIC is Figure 6. *The comparison to make in (b) is of sign and magnitude, not of significance: interval widths differ by specification for reasons given below.*

Depth as a factor replaces the single slope with ten level contrasts and assumes no functional form at all; the interaction is then the contrast between level-wise perturbation effects at the top and bottom of the ladder. *Natural spline* allows a smooth non-linearity with three degrees of freedom, with knots at the terciles of the depth distribution. *Segmented* fits separate slopes either side of the changepoint of Appendix G, with the break taken from that fit rather than re-estimated jointly. *Pooled with model-specific slopes* is the random-slopes model of Appendix H. *Per class* fits the interaction separately within each Category-A class and reports the weakest of the five, so a single class cannot

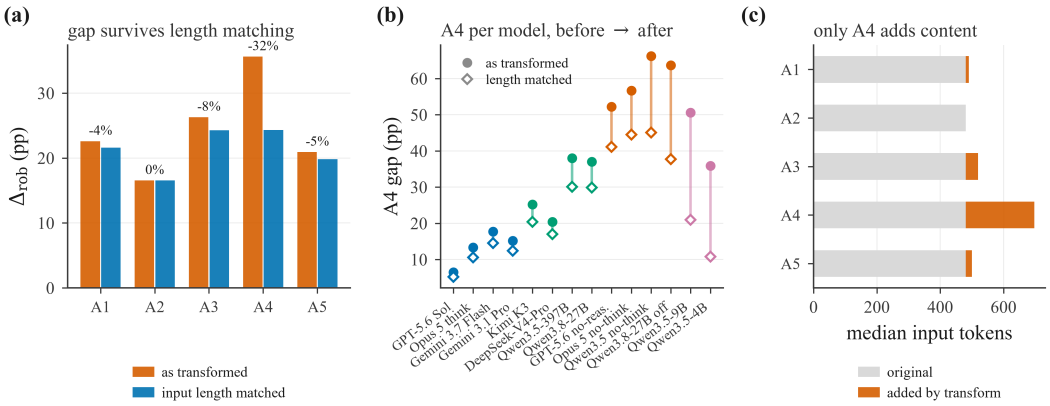

Figure 8: Input-length control on the answer-preserving classes. (a) Mean gap per class before and after matching input length between conditions, with the reduction annotated. (b) The A4 gap per model, before and after. (c) Median input tokens per class, showing that only A4 adds substantial content. *A2 reorders premises without adding a token, so its gap is unchanged by construction — which is the control that makes the A4 reduction interpretable.*

carry the result. *Length matched* substitutes the length-matched control arm of Appendix K for the transformed arm.

Interval widths are not comparable across rows, and the table should not be read as though they were. Ten level contrasts cost more degrees of freedom than one slope; a spline basis is correlated by construction; a segmented fit spends information locating a break; a class-specific fit uses one fifth of the items. Widths therefore run from 1.0 to about 1.6 times the primary fit for structural reasons that have nothing to do with whether the effect is real. What is comparable across rows is the sign and the magnitude, and on those the seven specifications agree: the median coefficient stays within 1% of the primary fit under a factor or a spline, rises by about a sixth when segmented, and falls to 0.67 and 0.86 of it under the class-specific and length-matched variants, with every interval below zero wherever the specification is identified at all.

Two specifications fail to resolve the term for Qwen3.5-9B and Qwen3.5-4B, and the reason is the same one that appears in the changepoint analysis: their accuracy reaches the floor inside the ladder, so the accuracy–depth curve bends twice and a single break cannot be placed. We report this as unidentified rather than reporting a coefficient, because a segmented fit on a doubly-bending curve returns a number without a referent. That the two independent analyses fail on exactly the same two models is a consistency check on both.

K Input-Length Control

A transformation that adds text changes two things at once: the surface form of the problem and the amount of context to be read. A4, distractor insertion, is the only Category-A class that adds substantial content — a median input of 481 tokens becomes 697 — and it is also the most damaging class on every model. It is therefore the class most exposed to the objection that what we measure is context length or attention dilution rather than sensitivity to surface form.

The control arm adds text without adding a distractor. We pad the *original* item with neutral text until its input length matches the transformed arm: sentences drawn from a fixed pool, restating background already given or describing quantities not referenced by the question, verified not to alter the answer key and not to introduce any term appearing in the solution path. The padded item is then scored by the same checker as the original.

What this arm licenses is narrower than the phrase “length effect” suggests, and we state it in those terms. Padding an item does not hold everything but token count fixed: it also changes the semantic content, the discourse structure, how attention is distributed and where in the prompt the load-bearing information sits. The difference between the original and padded arms is therefore the cost

Table 11: Input-length control for the answer-preserving classes, first half of the roster (continued in Table 12). A transformation that adds text changes the surface form *and* the amount of context to be read, and A4 — the only class that adds content — is the most damaging class on every configuration, so it is the one most exposed to the objection that what is measured is context length or attention dilution. The control arm pads the *original* item with neutral text, irrelevant to the solution path and leaving the answer key unchanged, until its input length matches the transformed arm; *adj.* is the gap that survives with length held equal. The share attributable to length is given on both scales the paper uses: the percentage-point share is descriptive and inherits the bound $\Delta \leq A_{\text{orig}}$, so it inflates for the weakest configurations, while the logit share is the inferential quantity. Checking realised output length against a model's own limit, as Appendix R does, does not address this: the quantity the transformation changes is the input.

Model	Class	in. tok.	orig.	padded	transf.	gap → adj.	length share
GPT-5.6 Sol (reasoning: high)	A1	481→490	96.2	96.1	92.6	3.6 → 3.5	2.9% pp, 4.0% logit
	A2	481→481	96.2	96.2	93.8	2.4 → 2.4	0.0% pp, 0.0% logit
	A3	481→519	96.2	96.0	92.0	4.2 → 4.0	5.1% pp, 7.3% logit
	A4	481→697	96.2	95.0	89.8	6.5 → 5.2	19.5% pp, 28.2% logit
	A5	481→500	96.2	96.1	93.1	3.1 → 3.0	3.6% pp, 4.8% logit
Claude Opus 5 (extended thinking)	A1	481→490	92.7	92.5	86.1	6.6 → 6.4	3.0% pp, 4.1% logit
	A2	481→481	92.7	92.7	88.3	4.4 → 4.4	-0.0% pp, 0.0% logit
	A3	481→519	92.7	92.3	84.1	8.6 → 8.2	5.2% pp, 7.3% logit
	A4	481→697	92.7	90.1	79.4	13.3 → 10.6	19.9% pp, 28.4% logit
	A5	481→500	92.7	92.5	86.6	6.2 → 5.9	3.7% pp, 4.9% logit
Gemini 3.7 Flash (thinkingLevel: high)	A1	481→490	94.1	93.8	85.7	8.3 → 8.1	2.7% pp, 4.1% logit
	A2	481→481	94.1	94.1	88.4	5.6 → 5.6	0.0% pp, 0.0% logit
	A3	481→519	94.1	93.6	83.3	10.8 → 10.3	4.6% pp, 7.4% logit
	A4	481→697	94.1	90.9	76.4	17.7 → 14.5	17.7% pp, 28.8% logit
	A5	481→500	94.1	93.8	86.7	7.4 → 7.1	3.4% pp, 4.9% logit
Gemini 3.1 Pro (thinkingLevel: high)	A1	481→490	93.7	93.5	85.9	7.8 → 7.6	2.8% pp, 4.1% logit
	A2	481→481	93.7	93.7	88.3	5.4 → 5.4	0.0% pp, 0.0% logit
	A3	481→519	93.7	93.2	84.3	9.3 → 8.9	4.9% pp, 7.4% logit
	A4	481→697	93.7	90.9	78.5	15.2 → 12.4	18.6% pp, 28.6% logit
	A5	481→500	93.7	93.4	86.8	6.9 → 6.6	3.5% pp, 4.9% logit
Kimi K3 (thinking)	A1	481→490	91.1	90.8	78.3	12.8 → 12.4	2.8% pp, 4.1% logit
	A2	481→481	91.1	91.1	81.9	9.2 → 9.2	0.0% pp, 0.0% logit
	A3	481→519	91.1	90.4	75.3	15.9 → 15.1	4.8% pp, 7.4% logit
	A4	481→697	91.1	86.4	66.0	25.2 → 20.4	18.8% pp, 28.9% logit
	A5	481→500	91.1	90.7	79.7	11.4 → 11.0	3.4% pp, 5.0% logit
DeepSeek-V4-Pro (thinking)	A1	481→490	94.4	94.2	84.5	9.9 → 9.6	2.6% pp, 4.2% logit
	A2	481→481	94.4	94.4	88.3	6.2 → 6.2	0.0% pp, 0.0% logit
	A3	481→519	94.4	93.9	82.3	12.1 → 11.6	4.4% pp, 7.5% logit
	A4	481→697	94.4	91.0	74.0	20.4 → 17.0	16.9% pp, 29.1% logit
	A5	481→500	94.4	94.2	85.6	8.8 → 8.5	3.2% pp, 5.0% logit
Qwen3.5-397B-A17B (thinking)	A1	481→490	87.1	86.5	66.7	20.4 → 19.8	2.9% pp, 4.2% logit
	A2	481→481	87.1	87.1	73.5	13.6 → 13.6	0.0% pp, 0.0% logit
	A3	481→519	87.1	85.8	62.8	24.2 → 23.0	5.0% pp, 7.5% logit
	A4	481→697	87.1	79.2	49.1	38.0 → 30.1	20.7% pp, 29.3% logit
	A5	481→500	87.1	86.4	68.1	18.9 → 18.3	3.5% pp, 5.0% logit

of *generic irrelevant text*, and the difference between the padded and transformed arms is what remains of the A4 penalty once that cost is matched. The quantity we report is accordingly the attenuation under a matched irrelevant-text control, not a pure token-length effect; an experiment that isolated length alone would need padding that is semantically null, which no natural-language padding is.

Under that control the A4 penalty attenuates by 28% to 36% on the logit scale, varying with capability (Tables 11 and 12): the attenuation is largest for the smallest models and for the reasoning-disabled settings, which is what one would expect if the mechanism is a limited ability to isolate relevant premises from a longer context. The remaining two thirds survives on all 14 configurations (Figure 8b). On the percentage-point scale the share appears much wider, 18% to 72%, but that spread is the floor artefact rather than a finding: for the weakest models the raw gap is already close to baseline accuracy, so removing any part of it moves the ratio a long way. The logit share is the inferential quantity, and it is the one quoted in the main text. A2, which reorders premises without adding a token, shows a length share of exactly zero (Figure 8a) — a control with no free parameter, and the reason the A4 number can be read as an estimate rather than a modelling choice.

Table 12: Input-length control, second half of the roster. Columns and definitions are those of Table 11.

Model	Class	in. tok.	orig.	padded	transf.	gap → adj.	length share
Qwen3.8-27B (thinking)	A1	481→490	89.8	89.3	71.2	18.6 → 18.1	2.7% pp, 4.2% logit
	A2	481→481	89.8	89.8	77.3	12.6 → 12.6	0.0% pp, 0.0% logit
	A3	481→519	89.8	88.8	66.9	23.0 → 21.9	4.7% pp, 7.6% logit
	A4	481→697	89.8	82.8	52.9	37.0 → 29.9	19.1% pp, 29.5% logit
	A5	481→500	89.8	89.3	73.2	16.6 → 16.1	3.3% pp, 5.1% logit
GPT-5.6 Sol (reasoning: none)	A1	481→490	89.2	88.5	61.3	28.0 → 27.2	2.7% pp, 4.5% logit
	A2	481→481	89.2	89.2	70.9	18.4 → 18.4	-0.0% pp, 0.0% logit
	A3	481→519	89.2	87.6	54.3	35.0 → 33.4	4.6% pp, 8.2% logit
	A4	481→697	89.2	78.1	37.0	52.2 → 41.1	21.2% pp, 31.8% logit
	A5	481→500	89.2	88.4	64.1	25.2 → 24.3	3.3% pp, 5.4% logit
Claude Opus 5 (thinking off)	A1	481→490	89.5	88.7	58.4	31.1 → 30.3	2.6% pp, 4.6% logit
	A2	481→481	89.5	89.5	68.9	20.5 → 20.5	0.0% pp, 0.0% logit
	A3	481→519	89.5	87.7	50.4	39.1 → 37.3	4.5% pp, 8.2% logit
	A4	481→697	89.5	77.3	32.8	56.7 → 44.5	21.5% pp, 32.0% logit
	A5	481→500	89.5	88.6	61.8	27.6 → 26.7	3.2% pp, 5.5% logit
Qwen3.5-397B-A17B (thinking off)	A1	481→490	80.9	79.4	35.5	45.4 → 43.8	3.3% pp, 4.7% logit
	A2	481→481	80.9	80.9	49.1	31.8 → 31.8	-0.0% pp, 0.0% logit
	A3	481→519	80.9	77.7	29.9	51.0 → 47.9	6.2% pp, 8.4% logit
	A4	481→697	80.9	59.8	14.7	66.2 → 45.1	31.8% pp, 32.6% logit
	A5	481→500	80.9	79.3	40.1	40.8 → 39.1	4.0% pp, 5.6% logit
Qwen3.8-27B (thinking off)	A1	481→490	71.9	69.9	24.5	47.4 → 45.4	4.2% pp, 4.7% logit
	A2	481→481	71.9	71.9	36.3	35.7 → 35.7	0.0% pp, 0.0% logit
	A3	481→519	71.9	67.5	17.8	54.1 → 49.7	8.1% pp, 8.5% logit
	A4	481→697	71.9	46.0	8.3	63.7 → 37.7	40.7% pp, 32.9% logit
	A5	481→500	71.9	69.6	26.3	45.6 → 43.3	5.0% pp, 5.6% logit
Qwen3.5-9B	A1	481→490	53.1	50.2	9.5	43.6 → 40.7	6.7% pp, 4.9% logit
	A2	481→481	53.1	53.1	16.2	36.9 → 36.9	0.0% pp, 0.0% logit
	A3	481→519	53.1	46.9	6.4	46.7 → 40.5	13.3% pp, 8.9% logit
	A4	481→697	53.1	23.5	2.5	50.6 → 21.0	58.4% pp, 34.4% logit
	A5	481→500	53.1	49.8	10.8	42.3 → 39.0	7.8% pp, 5.9% logit
Qwen3.5-4B	A1	481→490	36.8	33.8	4.1	32.7 → 29.6	9.3% pp, 5.1% logit
	A2	481→481	36.8	36.8	8.1	28.8 → 28.8	0.0% pp, 0.0% logit
	A3	481→519	36.8	30.6	2.7	34.1 → 27.9	18.3% pp, 9.2% logit
	A4	481→697	36.8	11.7	0.9	35.9 → 10.8	70.1% pp, 35.7% logit
	A5	481→500	36.8	33.4	4.8	32.0 → 28.6	10.7% pp, 6.1% logit

The conclusion this supports is narrower than the raw A4 gap would allow: part of the A4 effect is length or attention, and the larger part is not. Section 6 states it in those terms. Note also that recording output length against each model's own limit, as Appendix R does, does not address this question at all: the quantity a transformation changes is the input.

L Per-Class Perturbation Results

Figure 9 is the four-panel summary of Dimension 2 referenced from Section 5.3.

Figure 10 resolves the pooled gap into the five Category-A classes at every depth, which is where the stability of the class ordering can be checked.

Figure 11 shows the paired cell counts that the aggregate gap is computed from, and Table 13 lists them numerically with both conditional denominators.

Each family's counts run to 70 rows, so each is split over two floats (Tables 13 and 14 here, 15 and 16 below) rather than compressed into one page.

The Δ_{rob} column of Table 1 averages over both task families, so both are published: Synth-Logic carries the same schema, and the mean of the two families' $(b-c)/N$ over the five classes reproduces that column to the precision it is printed at. Only one family appeared in an earlier version, which left the headline column unreproducible from the appendix by 0.3 to 0.6 points — not because either table was wrong, but because half of the average was missing.

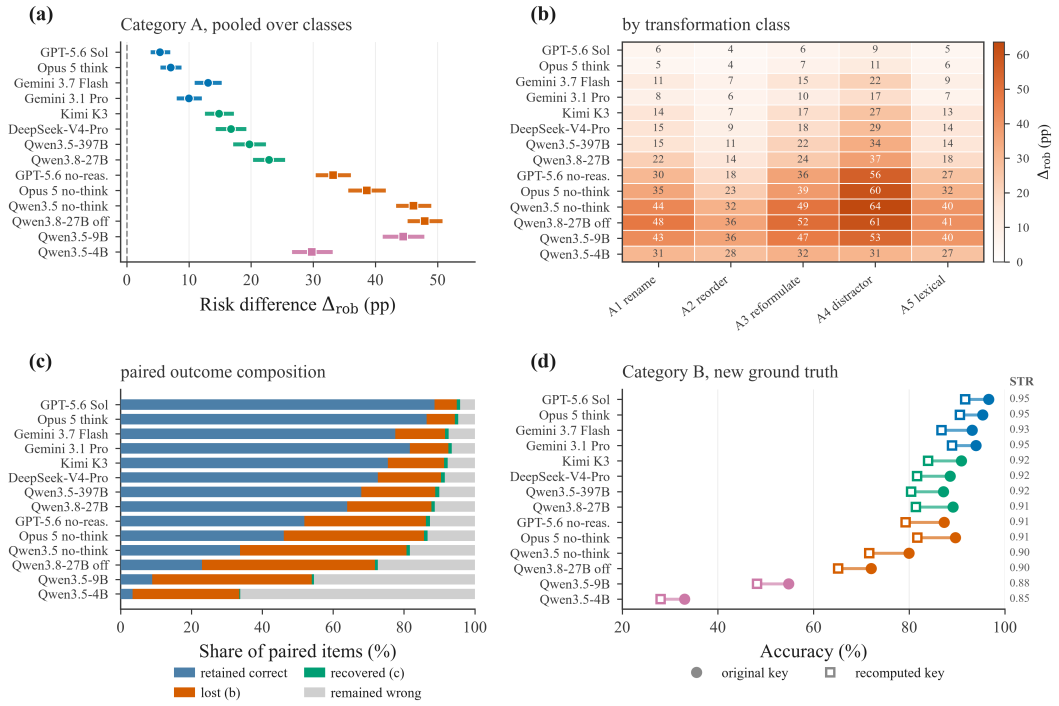

Figure 9: Perturbation robustness. (a) Risk difference with paired 95% intervals, pooled over Category-A classes; circles reasoning-enabled, squares reasoning-disabled. (b) The same quantity per transformation class. (c) Composition of the paired 2×2 outcomes, which shows how each aggregate was reached rather than only its value. (d) Category B accuracy against the original and the recomputed key, with the structural transfer ratio at the right. *Category B is scored against new ground truth, so its decline measures structural transfer under value change and not answer inconsistency.*

M THE PAIRED COMPARISON AT EVERY LADDER LEVEL

Table 1 and Table 13 report the Category-A comparison at one pre-registered depth, $n = 7$, the middle of the ladder. That choice was previously implicit, and a reader was entitled to be confused by it: Section 4 states that $N = 1{,}200$ source items exist per family *per level* over ten levels, so a paired table headed $N = 1{,}200$ could as easily have been read as a pooled figure over 12,000 items. It is not. It is one level, and Table 17 is that level in context.

A single depth is the right primary estimand here precisely because of what this paper reports. The gap depends on where on the ladder it is measured, so a figure pooled across levels would average over the variation under study and would have no determinate interpretation; a figure at a stated level has one. The cost is that the headline number is conditional, which is why the level is now named wherever the number appears.

Two patterns in the table are worth separating. For configurations that never approach the accuracy floor within the ladder the gap grows monotonically with depth, which is the compounding the interaction analysis estimates. For the weakest configurations it rises and then falls. That is not robustness returning at depth: the gap is bounded above by original accuracy, $0 \leq \Delta_{\mathrm{rob}}(n) \leq A_{\mathrm{orig}}(n)$, so once accuracy has collapsed there is little left to lose and the percentage-point gap must shrink. Reading that turnover as improvement would invert the finding, and it is the reason inference is drawn on the logit scale rather than on this one.

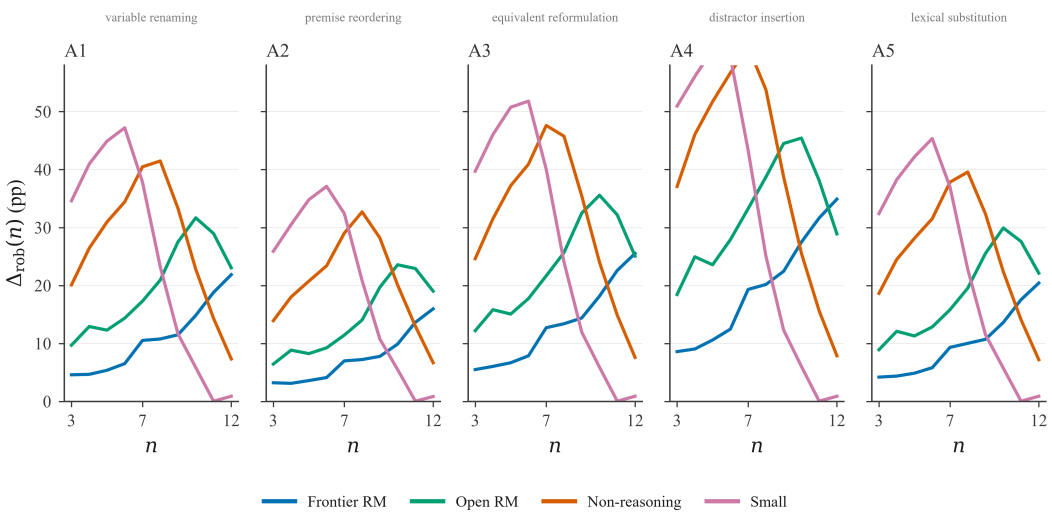

Figure 10: Conditional robustness gap against depth for each Category-A class, by model family, on shared axes. Distractor insertion (A4) is the most damaging class at every depth and premise reordering (A2) the least. The turnover visible for the weakest family at high depth is the accuracy bound rather than recovered robustness.

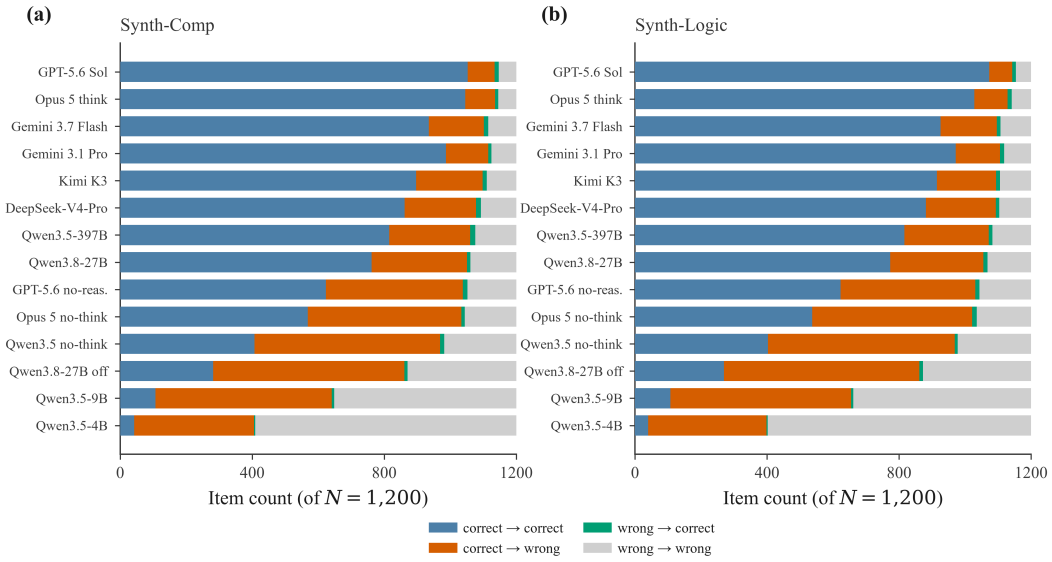

Figure 11: Paired 2×2 cell counts per model for both families, as item counts out of $N = 1{,}200$. The composition shows how each aggregate gap was reached: the recovered cell c is non-zero for every model but is small relative to the lost cell b, and the originally-incorrect base $c{+}d$ is what makes F^{+} unstable for strong models.

N Decoder Variance in the Headline Quantities

Appendix W repeats the benchmark evaluations. That leaves the question a reviewer should ask next unanswered: a benchmark score moving by under a point says nothing about a *difference* between two conditions, where decoder noise enters twice and could as easily compound as cancel. Table 18 therefore repeats each quantity at the scope it is estimated from, five times per configuration: the paired protocol at $n = 7$ for the gap, and the whole ladder $n = 3, \ldots, 12$ for the common interaction estimand, since a slope through ten levels cannot be re-estimated from a repetition at one of them.

Table 13: Category-A paired outcomes on SYNTH-COMP, the first of the two task families averaged in Table 1. Cells are item counts for original → perturbed: a correct–correct, b correct–incorrect, c incorrect–correct, d incorrect–incorrect, summing to $N = 1{,}200$ per row. b and c are the discordant counts, and the risk difference $(b - c)/N$ and discordant odds ratio b/c are computed from them. The reported p is a two-sided permutation test over the 48 seed templates, which are the unit of generalisation; p_{ind} is McNemar's mid-p under independence, carried alongside so the cost of taking the clustering seriously is visible rather than asserted. $F^- = b/(a{+}b)$ and $F^+ = c/(c{+}d)$ are conditioned on *different* bases, so F^+ can exceed F^- for a strong model whose $c{+}d$ is small even though $b > c$ holds throughout; the denominators n_{corr} and n_{wrong} are given so the rates are not read as directly comparable. The other family is Table 15; averaging the two reproduces the Δ_{rob} column of Table 1.

Model	Class	a	b	c	d	n_{corr}	n_{wrong}	F^-	F^+	$\mathrm{OR}_{\mathrm{disc}}$
GPT-5.6 Sol (reasoning: high)	A1	1053	78	12	57	1131	69	6.9	17.4	6.5
	A2	1075	63	16	46	1138	62	5.5	25.8	3.94
	A3	1059	80	8	53	1139	61	7.0	13.1	10.0
	A4	1020	120	14	46	1140	60	10.5	23.3	8.57
	A5	1056	69	11	64	1125	75	6.1	14.7	6.27
Claude Opus 5 (extended thinking)	A1	1067	71	9	53	1138	62	6.2	14.5	7.89
	A2	1057	65	14	64	1122	78	5.8	17.9	4.64
	A3	1032	95	11	62	1127	73	8.4	15.1	8.64
	A4	1008	140	5	47	1148	52	12.2	9.6	28.0
	A5	1064	81	9	46	1145	55	7.1	16.4	9.0
Gemini 3.7 Flash (thinkingLevel: high)	A1	949	145	13	93	1094	106	13.3	12.3	11.15
	A2	997	101	16	86	1098	102	9.2	15.7	6.31
	A3	911	187	8	94	1098	102	17.0	7.8	23.38
	A4	810	281	16	93	1091	109	25.8	14.7	17.56
	A5	1009	119	12	60	1128	72	10.5	16.7	9.92
Gemini 3.1 Pro (thinkingLevel: high)	A1	1011	111	11	67	1122	78	9.9	14.1	10.09
	A2	1017	92	14	77	1109	91	8.3	15.4	6.57
	A3	989	125	8	78	1114	86	11.2	9.3	15.63
	A4	893	219	10	78	1112	88	19.7	11.4	21.9
	A5	1026	91	8	75	1117	83	8.1	9.6	11.38
Kimi K3 (thinking)	A1	905	184	14	97	1089	111	16.9	12.6	13.14
	A2	998	104	15	83	1102	98	9.4	15.3	6.93
	A3	866	223	17	94	1089	111	20.5	15.3	13.12
	A4	768	329	10	93	1097	103	30.0	9.7	32.9
	A5	947	166	7	80	1113	87	14.9	8.0	23.71
DeepSeek-V4-Pro (thinking)	A1	871	202	21	106	1073	127	18.8	16.5	9.62
	A2	967	123	19	91	1090	110	11.3	17.3	6.47
	A3	828	225	15	132	1053	147	21.4	10.2	15.0
	A4	718	354	11	117	1072	128	33.0	8.6	32.18
	A5	925	177	10	88	1102	98	16.1	10.2	17.7
Qwen3.5-397B-A17B (thinking)	A1	869	200	16	115	1069	131	18.7	12.2	12.5
	A2	897	150	19	134	1047	153	14.3	12.4	7.89
	A3	799	269	10	122	1068	132	25.2	7.6	26.9
	A4	625	421	18	136	1046	154	40.2	11.7	23.39
	A5	885	185	17	113	1070	130	17.3	13.1	10.88

The comparison to make is the between-run spread against the effect it would have to explain, and it depends on which effect. Run-to-run standard deviation of the gap runs from 0.22 to 1.78 points over the 28 configuration–family cells. Against the median configuration-level gap of 21.3 points that is a factor of twelve; against the smallest configuration-level gap, 5.3 points, a factor of three; and against the thinnest individual class-level cell in the roster, 1.9 points, a factor of 1.1. Decoder noise therefore does not account for the reported effects in general, and for the thinnest cells it is not excluded — those cells appear in Table 13 and carry no claim. The estimand is the more stable of the two, with a between-run standard deviation between 0.0004 and 0.0057 against per-configuration values from -0.037 to -0.296.

A previous version of this paragraph said the margin was roughly an order of magnitude. That holds against the median gap and against no individual cell, and the distinction is the point: a between-run spread has to be compared with the specific quantity a claim rests on, not with the largest one in the table.

Table 14: Category-A paired outcomes on SYNTH-COMP, the first of the two task families averaged in Table 1, second half of the roster. Columns and definitions are those of Table 13.

Model	Class	a	b	c	d	n_{corr}	n_{wrong}	F^-	F^+	$\mathrm{OR}_{\mathrm{disc}}$
Qwen3.8-27B (thinking)	A1	768	278	8	146	1046	154	26.6	5.2	34.75
	A2	878	178	11	133	1056	144	16.9	7.6	16.18
	A3	737	301	11	151	1038	162	29.0	6.8	27.36
	A4	615	460	11	114	1075	125	42.8	8.8	41.82
	A5	811	227	12	150	1038	162	21.9	7.4	18.92
GPT-5.6 Sol (reasoning: none)	A1	647	377	14	162	1024	176	36.8	8.0	26.93
	A2	796	234	17	153	1030	170	22.7	10.0	13.76
	A3	580	449	14	157	1029	171	43.6	8.2	32.07
	A4	379	677	8	136	1056	144	64.1	5.6	84.63
	A5	715	336	16	133	1051	149	32.0	10.7	21.0
Claude Opus 5 (thinking off)	A1	588	433	8	171	1021	179	42.4	4.5	54.13
	A2	752	293	13	142	1045	155	28.0	8.4	22.54
	A3	554	478	11	157	1032	168	46.3	6.5	43.45
	A4	329	724	5	142	1053	147	68.8	3.4	144.8
	A5	619	398	14	169	1017	183	39.1	7.7	28.43
Qwen3.5-397B-A17B (thinking off)	A1	444	539	14	203	983	217	54.8	6.5	38.5
	A2	545	398	13	244	943	257	42.2	5.1	30.62
	A3	376	604	11	209	980	220	61.6	5.0	54.91
	A4	180	774	11	235	954	246	81.1	4.5	70.36
	A5	488	498	14	200	986	214	50.5	6.5	35.57
Qwen3.8-27B (thinking off)	A1	298	586	9	307	884	316	66.3	2.8	65.11
	A2	413	442	13	332	855	345	51.7	3.8	34.0
	A3	250	626	7	317	876	324	71.5	2.2	89.43
	A4	109	732	5	354	841	359	87.0	1.4	146.4
	A5	338	512	15	335	850	350	60.2	4.3	34.13
Qwen3.5-9B	A1	122	522	10	546	644	556	81.1	1.8	52.2
	A2	185	448	10	557	633	567	70.8	1.8	44.8
	A3	92	577	8	523	669	531	86.2	1.5	72.13
	A4	27	641	3	529	668	532	96.0	0.6	213.67
	A5	107	485	6	602	592	608	81.9	1.0	80.83
Qwen3.5-4B	A1	50	374	4	772	424	776	88.2	0.5	93.5
	A2	94	341	5	760	435	765	78.4	0.7	68.2
	A3	25	392	3	780	417	783	94.0	0.4	130.67
	A4	7	378	1	814	385	815	98.2	0.1	378.0
	A5	36	329	6	829	365	835	90.1	0.7	54.83

O CONFIGURATIONS AGAINST CHECKPOINTS

The roster is 14 configurations over 10 checkpoints: four checkpoints appear twice, once with reasoning enabled and once disabled. Every pairwise statistic over "all pairs" therefore includes pairs that are two settings of one system. Those pairs exist because the design wanted them, but they are not two independent draws from a population of models, and a rank correlation computed over them answers a slightly different question from one computed over distinct systems.

Table 19 reports both. The result qualifies the divergence claim rather than supporting it: agreement is *higher* on the reduced set, $\rho = 0.98$ with 2 of 45 pairs inverted against $\rho = 0.92$ with 8 of 91. Part of the local disagreement in Section 5.1 is therefore contributed by the inference-mode pairs. The divergence among distinct checkpoints is real, and smaller than the full-roster figure alone suggests. We report the full-roster figure as the headline because the reasoning-disabled configurations are systems a practitioner can actually deploy and compare, but the reduced figure is the one to cite for a claim about model families rather than about configurations.

P CATEGORY B: STRUCTURAL TRANSFER

Figure 12 and Table 20 report Category B by class and family.

Table 15: Category-A paired outcomes on SYNTH-LOGIC, the second of the two task families averaged in Table 1. The two families share no surface form, so agreement between them is replication rather than repetition. Cells are item counts for original → perturbed: a correct–correct, b correct–incorrect, c incorrect–correct, d incorrect–incorrect, summing to $N = 1{,}200$ per row. b and c are the discordant counts, and the risk difference $(b-c)/N$ and discordant odds ratio b/c are computed from them. The reported p is a two-sided permutation test over the 40 seed templates, which are the unit of generalisation; p_{ind} is McNemar's mid-p under independence, carried alongside so the cost of taking the clustering seriously is visible rather than asserted. $F^- = b/(a+b)$ and $F^+ = c/(c+d)$ are conditioned on *different* bases, so F^+ can exceed F^- for a strong model whose $c+d$ is small even though $b > c$ holds throughout; the denominators n_{corr} and n_{wrong} are given so the rates are not read as directly comparable. The other family is Table 13; averaging the two reproduces the Δ_{rob} column of Table 1.

Model	Class	a	b	c	d	n_{corr}	n_{wrong}	F^-	F^+	$\mathrm{OR}_{\mathrm{disc}}$
GPT-5.6 Sol (reasoning: high)	A1	1078	56	14	52	1134	66	4.9	21.2	4.0
	A2	1096	43	20	41	1139	61	3.8	32.8	2.15
	A3	1074	66	10	50	1140	60	5.8	16.7	6.6
	A4	1013	132	5	50	1145	55	11.5	9.1	26.4
	A5	1106	50	7	37	1156	44	4.3	15.9	7.14
Claude Opus 5 (extended thinking)	A1	1070	86	4	40	1156	44	7.4	9.1	21.5
	A2	1036	83	17	64	1119	81	7.4	21.0	4.88
	A3	1022	101	15	62	1123	77	9.0	19.5	6.73
	A4	962	162	8	68	1124	76	14.4	10.5	20.25
	A5	1051	71	18	60	1122	78	6.3	23.1	3.94
Gemini 3.7 Flash (thinkingLevel: high)	A1	932	150	12	106	1082	118	13.9	10.2	12.5
	A2	999	97	15	89	1096	104	8.9	14.4	6.47
	A3	931	189	12	68	1120	80	16.9	15.0	15.75
	A4	805	271	7	117	1076	124	25.2	5.6	38.71
	A5	963	144	9	84	1107	93	13.0	9.7	16.0
Gemini 3.1 Pro (thinkingLevel: high)	A1	1012	109	9	70	1121	79	9.7	11.4	12.11
	A2	1029	77	11	83	1106	94	7.0	11.7	7.0
	A3	953	155	14	78	1108	92	14.0	15.2	11.07
	A4	888	210	11	91	1098	102	19.1	10.8	19.09
	A5	977	120	15	88	1097	103	10.9	14.6	8.0
Kimi K3 (thinking)	A1	956	154	6	84	1110	90	13.9	6.7	25.67
	A2	961	122	22	95	1083	117	11.3	18.8	5.55
	A3	888	189	11	112	1077	123	17.5	8.9	17.18
	A4	808	283	9	100	1091	109	25.9	8.3	31.44
	A5	959	150	12	79	1109	91	13.5	13.2	12.5
DeepSeek-V4-Pro (thinking)	A1	917	191	8	84	1108	92	17.2	8.7	23.88
	A2	985	115	10	90	1100	100	10.5	10.0	11.5
	A3	854	222	17	107	1076	124	20.6	13.7	13.06
	A4	726	361	5	108	1087	113	33.2	4.4	72.2
	A5	926	167	12	95	1093	107	15.3	11.2	13.92
Qwen3.5-397B-A17B (thinking)	A1	854	222	11	113	1076	124	20.6	8.9	20.18
	A2	883	180	12	125	1063	137	16.9	8.8	15.0
	A3	813	259	14	114	1072	128	24.2	10.9	18.5
	A4	626	429	11	134	1055	145	40.7	7.6	39.0
	A5	905	187	9	99	1092	108	17.1	8.3	20.78

Q ALL INTERVENTIONS BY OBSERVABILITY CLASS

Table 21 gives both dimensions per configuration and Figure 13 plots them. Two patterns qualify what Section 5.5 reports. Flip rates rise as capability falls, which is consistent with stronger models depending less on the verbalized trace and equally consistent with their higher baseline accuracy leaving fewer items able to flip; this design does not separate the two, and we do not claim the first. And the novelty ordering holds on every configuration while its magnitudes do not track capability — the compositional gap is similar at both ends of the roster — which is why Dimension 4 is the least redundant with benchmark accuracy and also the hardest to attribute.

Figure 14 and Table 21 give all six interventions with the observability class attached, so the design boundary between the two manipulation kinds is visible rather than implied.

Table 16: Category-A paired outcomes on SYNTH-LOGIC, the second of the two task families averaged in Table 1, second half of the roster. Columns and definitions are those of Table 15.

Model	Class	a	b	c	d	n_{corr}	n_{wrong}	F^-	F^+	$\mathrm{OR}_{\mathrm{disc}}$
Qwen3.8-27B (thinking)	A1	816	240	14	130	1056	144	22.7	9.7	17.14
	A2	867	177	16	140	1044	156	17.0	10.3	11.06
	A3	751	310	11	128	1061	139	29.2	7.9	28.18
	A4	583	466	6	145	1049	151	44.4	4.0	77.67
	A5	848	220	13	119	1068	132	20.6	9.8	16.92
GPT-5.6 Sol (reasoning: none)	A1	647	371	13	169	1018	182	36.4	7.1	28.54
	A2	815	224	16	145	1039	161	21.6	9.9	14.0
	A3	576	461	16	147	1037	163	44.5	9.8	28.81
	A4	377	653	4	166	1030	170	63.4	2.4	163.25
	A5	697	333	15	155	1030	170	32.3	8.8	22.2
Claude Opus 5 (thinking off)	A1	602	440	17	141	1042	158	42.2	10.8	25.88
	A2	703	301	16	180	1004	196	30.0	8.2	18.81
	A3	473	546	16	165	1019	181	53.6	8.8	34.13
	A4	322	706	3	169	1028	172	68.7	1.7	235.33
	A5	583	432	16	169	1015	185	42.6	8.6	27.0
Qwen3.5-397B-A17B (thinking off)	A1	450	505	9	236	955	245	52.9	3.7	56.11
	A2	582	365	11	242	947	253	38.5	4.3	33.18
	A3	371	612	13	204	983	217	62.3	6.0	47.08
	A4	177	814	5	204	991	209	82.1	2.4	162.8
	A5	433	534	6	227	967	233	55.2	2.6	89.0
Qwen3.8-27B (thinking off)	A1	309	565	10	316	874	326	64.6	3.1	56.5
	A2	403	462	17	318	865	335	53.4	5.1	27.18
	A3	225	628	6	341	853	347	73.6	1.7	104.67
	A4	113	764	2	321	877	323	87.1	0.6	382.0
	A5	298	540	20	342	838	362	64.4	5.5	27.0
Qwen3.5-9B	A1	113	556	7	524	669	531	83.1	1.3	79.43
	A2	203	447	9	541	650	550	68.8	1.6	49.67
	A3	61	579	4	556	640	560	90.5	0.7	144.75
	A4	27	655	2	516	682	518	96.0	0.4	327.5
	A5	132	498	13	557	630	570	79.0	2.3	38.31
Qwen3.5-4B	A1	47	373	3	777	420	780	88.8	0.4	124.33
	A2	81	309	8	802	390	810	79.2	1.0	38.63
	A3	19	373	2	806	392	808	95.2	0.2	186.5
	A4	7	381	1	811	388	812	98.2	0.1	381.0
	A5	44	358	4	794	402	798	89.1	0.5	89.5

R OUTPUT-LENGTH BUDGET

Figure 15 reports the output-length diagnostics that exclude truncation as an explanation for the decline at depth.

Reasoning-enabled configurations emit roughly seven times as many output tokens per required step as reasoning-disabled ones, so the budget check matters most for the models that would otherwise be most affected. The tightest combination in the roster is Qwen3.5-397B-A17B with thinking enabled at $n = 12$, where the 95th percentile reaches 11,196 tokens against a 32,768-token limit.

S PRECISION ANALYSIS

Sample size is derived from the paired-discordance relation rather than from an assumed detectable difference. For a paired binary comparison, only the two discordant cells of the 2×2 table carry information, so the resolvable difference is governed by the discordance rate π_d. Writing p_1 for accuracy in the original condition, p_2 for accuracy in the transformed condition, $q_i = 1 - p_i$, and ρ for the within-item ϕ coefficient — the Pearson correlation between the two binary outcomes — the expected discordant proportion is

$$\pi_d \;=\; \frac{\mathbb{E}[b] + \mathbb{E}[c]}{N} \;=\; p_1 q_2 + p_2 q_1 - 2\rho\sqrt{p_1 q_1 p_2 q_2},$$

where b and c count items correct only in the original and only in the transformed condition. The covariance term $\rho\sqrt{p_1 q_1 p_2 q_2}$ identifies ρ as the ϕ coefficient specifically, not as a tetrachoric correlation: the latter is defined on latent bivariate normal variables underlying the dichotomies, and using

Table 17: Category-A gap under distractor insertion (A4) at every level of the depth ladder, on SYNTH-COMP. The headline table reports one pre-registered level, $n = 7$, shown in bold. This table exists because the methods state $N = 1{,}200$ source items *per family per level* over ten levels, which left the headline $N = 1{,}200$ ambiguous about which level it referred to; the headline row is the $n = 7$ column of this table and nothing else. It also shows directly why a single number is not enough: the gap a robustness study reports depends on where on the ladder it was measured. For the strongest configurations it grows monotonically across the ladder. For the weakest it rises and then falls, which is not a recovery but the bound $0 \leq \Delta_{\text{rob}} \leq A_{\text{orig}}$ taking effect as accuracy approaches the floor — the reason inference is drawn on the logit scale.

Model	n=3	n=4	n=5	n=6	n=7	n=8	n=9	n=10	n=11	n=12
GPT-5.6 Sol (reasoning: high)	0.8	2.3	5.1	7.3	**8.8**	12.3	10.9	13.7	18.8	19.3
Claude Opus 5 (extended thinking)	2.8	4.0	7.0	10.1	**11.3**	13.3	15.5	19.4	23.5	25.0
Gemini 3.7 Flash (thinkingLevel: high)	6.6	12.9	14.2	16.9	**22.1**	24.6	25.8	38.9	46.6	42.8
Gemini 3.1 Pro (thinkingLevel: high)	4.9	5.0	8.7	13.3	**17.4**	21.0	24.9	24.4	39.5	38.4
Kimi K3 (thinking)	10.4	16.1	17.2	21.6	**26.6**	27.1	32.9	41.3	44.7	35.5
DeepSeek-V4-Pro (thinking)	12.2	15.0	20.5	26.6	**28.6**	36.2	42.2	47.2	43.8	32.6
Qwen3.5-397B-A17B (thinking)	17.3	16.2	21.3	31.2	**33.6**	38.8	48.3	46.3	36.1	29.3
Qwen3.8-27B (thinking)	16.9	23.3	26.8	31.1	**37.4**	45.0	50.3	41.8	32.6	24.2
GPT-5.6 Sol (reasoning: none)	27.1	31.4	39.8	44.1	**55.8**	60.6	49.8	39.7	25.8	17.9
Claude Opus 5 (thinking off)	31.4	37.5	46.9	55.6	**59.9**	57.8	42.8	28.3	22.8	13.7
Qwen3.5-397B-A17B (thinking off)	38.8	47.4	56.4	60.8	**63.6**	48.8	35.6	22.8	14.2	7.6
Qwen3.8-27B (thinking off)	42.3	48.1	54.9	61.9	**60.6**	44.8	29.8	21.7	13.3	5.3
Qwen3.5-9B	47.7	57.6	61.1	64.0	**53.2**	32.2	21.3	11.8	4.6	2.3
Qwen3.5-4B	51.0	57.9	59.8	54.5	**31.4**	18.9	9.9	3.2	1.7	1.9

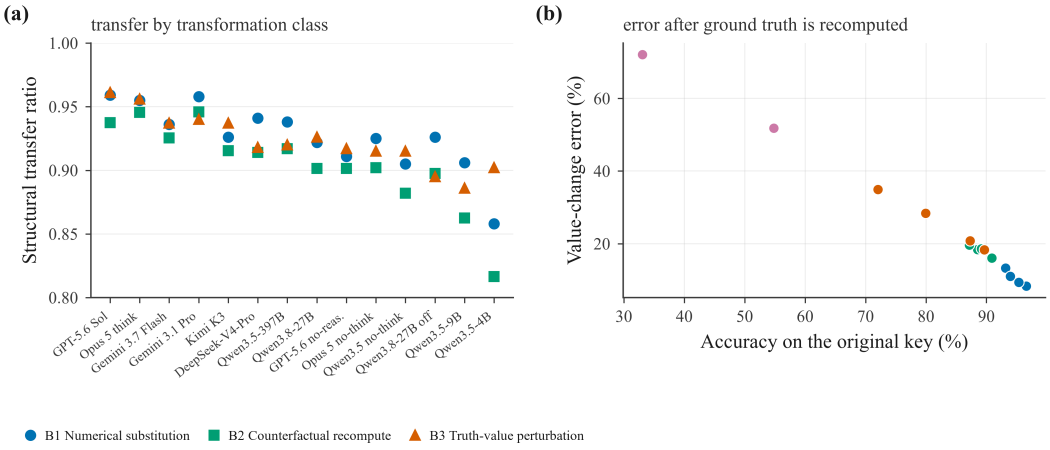

Figure 12: (a) Structural transfer ratio by transformation class and model. (b) Accuracy on the original key against the error rate once ground truth is recomputed. *Answer-consistency is not computed for this category by construction: the answer is required to change.*

it would require the joint cell probabilities to come from a bivariate normal CDF rather than from this product-moment expression. The two are not interchangeable and mixing them would misstate π_d. The minimum detectable difference at two-sided α, corrected over m classes, with power $1 - \beta$ is then $\text{MDES} = (z_{1-\alpha/2m} + z_{1-\beta})\sqrt{\pi_d / N}$.

Two properties of this expression matter for how it is used. First, it is implicit: $p_2 = p_1 - \text{MDES}$ depends on the effect being detected, so π_d cannot be evaluated before the answer is known. We solve it by fixed-point iteration from $p_2 = p_1$, which converges to within 10^{-6} in fewer than ten steps over this entire grid. Second, the familiar form $\pi_d = 2p(1-p)(1-\rho)$ is the special case $p_1 = p_2 = p$, obtained by substitution. That case is exactly the null hypothesis this design exists to reject, so using it to plan the study assumes the condition whose failure is the object of measurement. The consequence is not negligible and not in the safe direction: across this grid the equal-marginal form understates the MDES by 0.9% to 12.9%, and it understates it most where the effect is largest. We report the general form throughout and give the special case as a separate row of Table 22 for comparison.

Table 18: Run-to-run variation of the quantities the paper's claims rest on. The repeated benchmark evaluations of Table 7 could not speak to this: a score moving by under a point says nothing about a *gap* between two conditions, where decoder noise enters twice and might cancel or compound. Each quantity is therefore repeated at the scope it is estimated from, five times per configuration: the gap over the paired protocol at $n = 7$, and $\hat{\delta}$ — the common interaction estimand of Table 2 — over the whole ladder $n = 3, \ldots, 12$, since a slope through ten levels cannot be re-estimated from one of them. Columns pool both task families; per-family spreads are in Appendix W. The comparison that matters is the standard-deviation column against the gap it has to explain. Against the median gap the margin is wide, but against the smallest gap reported anywhere in the paper, 1.9 pp, the largest between-run standard deviation on any cell, 1.78 pp, leaves a factor of only 1.1. Decoder noise is therefore not a general explanation of the reported effects, and for the narrowest gaps in the roster it is not ruled out either: those cells are reported and not used to support a claim. An earlier version of this caption called the margin an order of magnitude, which held for no cell in the table.

Model	Δ_{rob} mean	s.d.	$\hat{\delta}$ mean	s.d.	s.d. as % of gap
GPT-5.6 Sol (reasoning: high)	5.5	0.40	-0.0380	0.0007	7.2%
Claude Opus 5 (extended thinking)	6.6	0.57	-0.0452	0.0012	8.6%
Gemini 3.7 Flash (thinkingLevel: high)	12.3	0.60	-0.0677	0.0019	4.9%
Gemini 3.1 Pro (thinkingLevel: high)	9.8	0.45	-0.0594	0.0016	4.6%
Kimi K3 (thinking)	14.3	0.72	-0.0769	0.0030	5.0%
DeepSeek-V4-Pro (thinking)	16.8	0.58	-0.0850	0.0031	3.4%
Qwen3.5-397B-A17B (thinking)	20.1	0.86	-0.0923	0.0021	4.3%
Qwen3.8-27B (thinking)	23.4	0.64	-0.1054	0.0030	2.7%
GPT-5.6 Sol (reasoning: none)	32.9	0.63	-0.1598	0.0025	1.9%
Claude Opus 5 (thinking off)	37.9	0.46	-0.1803	0.0048	1.2%
Qwen3.5-397B-A17B (thinking off)	45.9	0.81	-0.2097	0.0048	1.8%
Qwen3.8-27B (thinking off)	48.0	0.83	-0.2113	0.0077	1.7%
Qwen3.5-9B	43.8	0.85	-0.2555	0.0091	2.0%
Qwen3.5-4B	30.4	1.41	-0.2894	0.0088	4.6%

Table 19: Rank agreement between benchmark accuracy and stress behaviour, recomputed with one configuration per checkpoint. The roster is 14 configurations over 10 checkpoints, because four checkpoints appear twice with reasoning enabled and disabled, so statistics over all pairs include pairs that are two settings of one system. Keeping only the reasoning-enabled member removes those. The comparison is reported because it qualifies the divergence claim rather than supporting it: agreement is *higher* on the reduced set, so part of the local disagreement in Section 5.1 comes from the inference-mode pairs the design deliberately includes. The divergence among distinct checkpoints is real but smaller than the full-roster figure suggests.

Scope	Configurations	Pairs	Spearman ρ	Inversions
all configurations	14	91	0.916	8
one per checkpoint	10	45	0.976	2

The multiplicity quantile is $z_{1-\alpha/2m}$, which is the **Bonferroni** bound rather than Holm. The distinction matters in one direction only. Holm's step-down procedure rejects at least as much as Bonferroni at the same α, so the Bonferroni quantile gives a conservative MDES: a design powered against it is powered against Holm. That makes it the appropriate choice for planning, and it is why the correction used for *post hoc* inference in Appendix D — Holm within a model, Benjamini–Hochberg across models for the interaction — is deliberately different from the bound used here. Planning uses the conservative quantile; testing uses the sharper procedure.

Because a derivation is easier to get wrong than to check, each cell is verified by simulation rather than asserted. The correction family is the $m = 5$ Category-A perturbation classes tested within one model, which is the set of comparisons actually performed; the per-comparison two-sided level is therefore $\alpha/m = 0.01$, whose critical value is the quantile $z_{1-\alpha/2m}$ above. For every (N, p_1, ρ) we

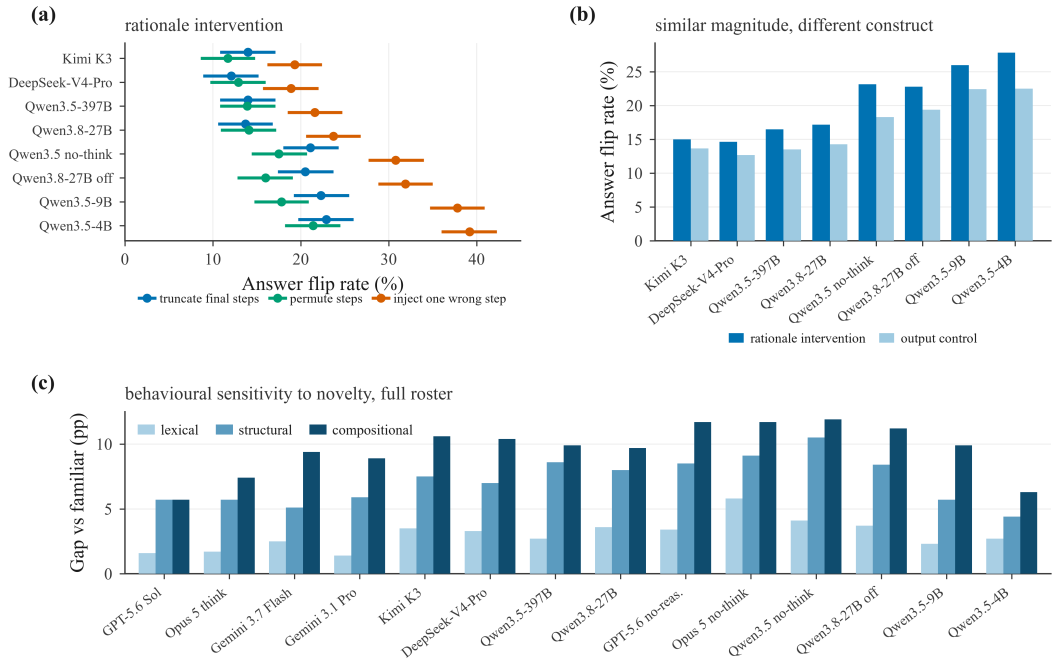

Figure 13: (a) Answer flip rate under each rationale intervention with 95% intervals, for the eight Class-A configurations that expose a manipulable trace. (b) Rationale interventions against output-control interventions on the same configurations. (c) Novelty gaps relative to familiar templates with solution depth matched, full roster. *Flip rates measure the behavioural influence of a verbalized trace on the emitted answer, not an internal computational path. Novelty gaps measure sensitivity to familiarity, not training-data exposure.*

draw 3,000 paired datasets at the solved MDES from a bivariate binary distribution with the specified marginals and correlation, apply McNemar's mid-p test at that level, and record the rejection rate. Empirical power is 0.791 to 0.829 against the nominal 0.800 across all 36 cells (Figure 16c), a spread consistent with Monte Carlo error at this replication count, so the analytic expression is accurate to within simulation error.

Items are not the unit of generalisation. Everything above treats the N item pairs as independent, which they are not: items instantiated from one seed template share its structure, so the effective information is smaller than the item count suggests. Our construction implies a template intra-class correlation of $\rho_{\text{ICC}} = 0.090$, and with 25 and 30 instances per template the design effect $1 + (\bar{m} - 1)\rho_{\text{ICC}}$ is 3.16 and 3.61 for the two families. Interval estimates already respected this, because the bootstrap resamples templates. The test and the precision analysis did not, and a template-level bootstrap cannot repair a p-value computed elsewhere under independence. Both now use the effective quantities: McNemar's mid-p is evaluated on discordant counts deflated by the design effect, which preserves the discordant split and so leaves the odds ratio unchanged while scaling only the evidence against the null; and the MDES is solved at $N_{\text{eff}} = N/\text{deff}$, reported as a separate column of Table 22. At $N = 1{,}200$ the resolvable difference is 2.4–5.5 points under independence and 5.0–10.7 once clustering is carried through, so the correction is not cosmetic. Its consequence for the reported results is nonetheless bounded, and the bound is what licenses the claims rather than the original independence assumption: widening every Category-A interval by $\sqrt{\text{deff}}$ leaves all 140 gaps distinguishable from zero, and all 192 identified interaction estimands retain intervals entirely below zero. The margin is thinnest for the two strongest models under premise reordering, where the gap is near 3 points against a clustered half-width close to 2 points; those rows are the first that a larger template ICC would remove, and we report them with that in mind rather than as securely established.

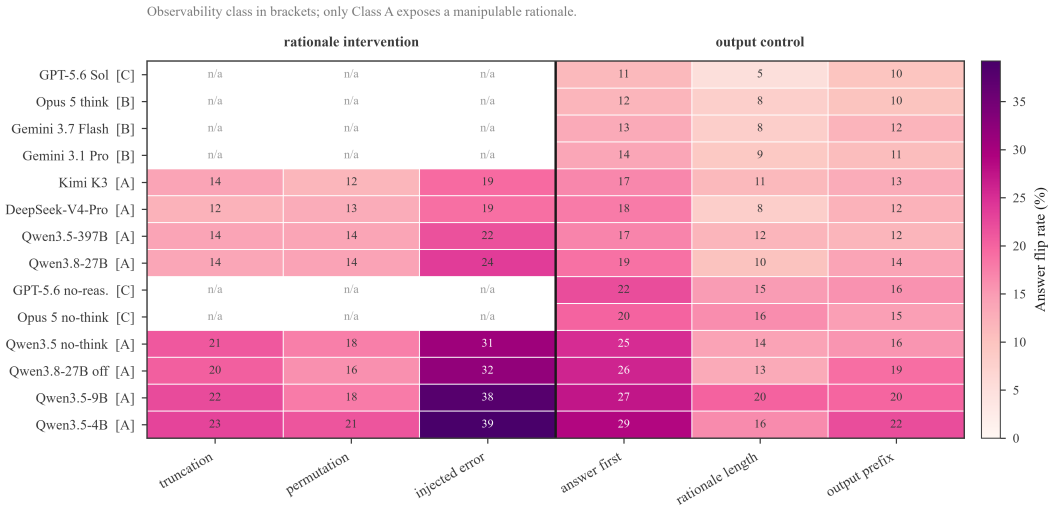

Figure 14: Answer flip rate for all six interventions, with the observability class in brackets. The rule dividing the matrix is a design boundary rather than a data gap: rationale interventions edit a trace and are available for Class A only, while output-control interventions change prompting and apply to every model. *Only the left block is faithfulness evidence. The two blocks reach comparable magnitudes, which is the reason not to pool them into a single index.*

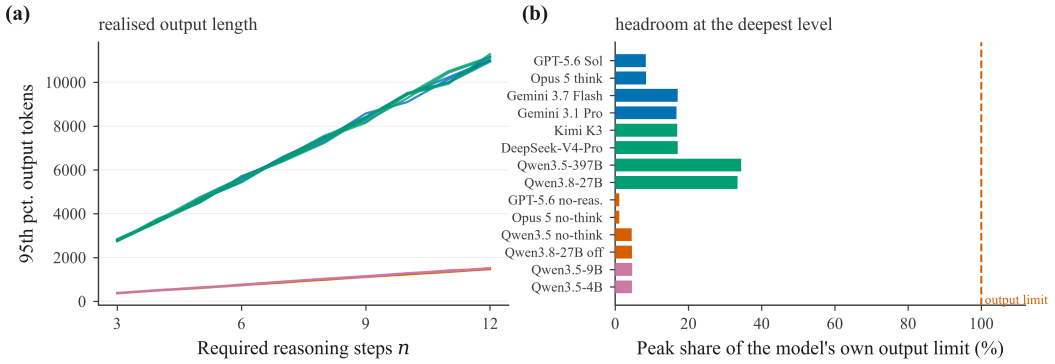

Figure 15: (a) Realised 95th-percentile output length against required steps, per model. (b) Peak share of each model's own output limit across the ladder, with the limit marked. *The highest usage anywhere in the design is 34% of the applicable limit and no cell exceeds 0.01% truncation, which excludes output truncation as an explanation for the decline at depth.*

T Transformation Examples and Validation

Table 23 instantiates every transformation class on one source item. Category A transformations are accepted only when a checker confirms the answer key is unchanged; Category B transformations are accepted with ground truth recomputed by the generator. Items failing validation are discarded rather than repaired, so a failed transform cannot silently change task identity.

Construction statistics. A single worked example shows what a transformation does but not whether the corpus behind it is broad, so the construction counts are given directly. SYNTH-COMP is built from 48 seed templates and SYNTH-LOGIC from 40, instantiated 25 and 30 times per template per level to give $N = 1{,}200$ source items per family per level over ten levels. Candidate instantiations are rejected as near-duplicates at rates of 3.1% and 2.6% — a candidate is a near-duplicate when its normalised token sequence matches an accepted item above a fixed similarity threshold after entity substitution, which catches templates that collapse to the same surface string under a particular variable draw. Of attempted transformations, 2.2% and 1.9% fail validation and are dis-

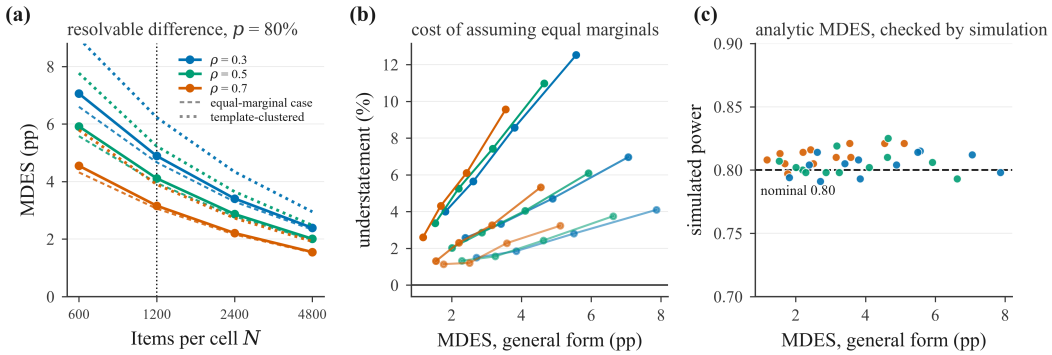

Figure 16: Precision of the paired comparison. (a) Minimum detectable effect against items per cell at 80% baseline accuracy, for three paired correlations, with the chosen $N = 1{,}200$ marked; dashed lines are the equal-marginal special case. (b) How far that special case understates the MDES, over the whole grid. (c) Empirical power from simulation at the solved MDES against the nominal 0.80. *Higher baseline accuracy and higher within-item correlation both reduce the discordance rate and therefore improve resolution, which is why the applicable figure is quoted per comparison rather than once globally.*

carded; A3, equivalent reformulation, is the class that fails most often, at 4.8% and 3.9%, since rewriting a problem while provably preserving its answer key is the hardest of the five to automate.

Two consequences follow for the analysis. First, the template count, not the item count, is the unit of independence: 1,200 items per level are 48 or 40 templates instantiated repeatedly, and items sharing a template are correlated. Every bootstrap in this paper therefore resamples templates, which is the conservative choice — resampling instances would treat 1,200 correlated observations as 1,200 independent ones and understate every interval. Second, an automatic checker that is itself wrong would propagate silently, so 200 items per family were audited by hand against the checker's decision. Agreement is 98.5% on SYNTH-COMP and 98.0% on SYNTH-LOGIC ($\kappa = 0.96$ and 0.95); the disagreements are concentrated in A3 and are cases where the checker accepted a reformulation a human read as ambiguous rather than cases where the answer key had changed.

U BUDGET ABLATION

Raising the reasoning budget from the vendor default to the maximum available shifts the estimated changepoint later by one to two levels for the frontier reasoning models and narrows the Category-A gap by 3–6 points, without removing either effect. Setting B, a vendor-matched budget where the interfaces permit it, preserves the model ordering obtained under Setting A, so the cross-vendor comparison is not driven by differences in default budgets alone. Part of what a benchmark score attributes to a model is nonetheless attributable to inference compute allocation, which is an argument for reporting the budget beside the score.

V CLOSEST PRIOR WORK

Table 24 compares this work against the closest located studies feature by feature. The difference from prior work is not the number of stress axes but that two of them are crossed, and that the interaction is estimated on a scale where the bound on the descriptive gap does not distort it. We phrase this as a difference in what is measured rather than as a priority claim.

W RUN-TO-RUN VARIANCE

Headline numbers come from one pre-registered run per item rather than from an average over repeated runs, because the two uncertainty sources are different quantities and averaging them together would leave neither interpretable. Item sampling uncertainty — which items were drawn — is ad-

dressed by the bootstrap used throughout, which resamples templates rather than instances, since items sharing a template are not independent. Generation stochasticity is a property of the decoder on a fixed item set, and is reported separately.

Where the interface exposes deterministic decoding we set it and the question does not arise. Four configurations do not: three expose no sampling parameter at all and one forces temperature to 1.0 whenever extended thinking is enabled (Table 3). For those the single run is a draw, so each was re-evaluated five times over the same items, with results in Table 7. Between-run standard deviation is 0.11 to 1.32 points on the four larger reference sets and reaches 1.83 points on AIME 2026, which is the expected consequence of 30 items: a single problem is worth 3.33 points there, so one flipped answer produces most of that spread.

The ordering of these four configurations is not stable under decoder noise alone on three of the five benchmarks, and which three is informative rather than merely inconvenient. What governs stability is the smallest gap between adjacent models measured against the between-run standard deviation. That ratio is 2.96 on FrontierMath and 5.64 on GSM-Symbolic, where the ordering is identical in all five repetitions; it is 2.29 on MATH-500, 0.49 on GPQA-Diamond and 0.00 on AIME 2026, where it is not, and on GPQA-Diamond even the top two configurations exchange places between repetitions. A saturated benchmark compresses the frontier until adjacent models sit closer together than one decoder's own run-to-run variation, at which point a reported ranking of those models is not a measurement of them. This is the same point Section 5.1 makes from the robustness gaps, arrived at from a different direction, and it is why we order the roster by MATH-500 for presentation while drawing no inference from adjacent positions within the frontier group.

What was re-run, and how often. All 14 configurations are evaluated five times, and the scope is the same for every configuration rather than reserved for those whose decoding cannot be constrained. The primary run is the first of the five and is the value reported throughout the paper; the other four establish the between-run spread and are reported as such rather than averaged into it. Keeping the primary count as the reported value means one comparison rule applies to every cell of every table, which a mean over five runs for some configurations and a single run for others would not.

The three reported quantities are re-run at the scope each needs, which differs because they are estimated from different data:

- **Benchmark accuracy**: five passes over the full item set of each of the five reference points, for all 14 configurations.
- **Category-A robustness gap**: five passes over the paired protocol at the pre-registered depth $n = 7$, for all 14 configurations and all five perturbation classes on both families.
- **The interaction estimand** $\hat{\delta}$: five passes over the *whole* ladder, $n = 3, \ldots, 12$. This quantity is a slope through ten levels, so a repetition at one depth would not produce a second estimate of it. An earlier version stated a five-fold repeat of the paired protocol at $n = 7$ and then reported a between-run spread for $\hat{\delta}$, which those runs could not have supplied.

The cost is the reason the earlier design re-ran only the configurations with no settable temperature: repeating the ladder five times for fourteen configurations is the dominant term in the evaluation budget, and Appendix R states it. The reason for paying it is that decoder variance is not confined to the configurations whose temperature is locked — eight of the fourteen run at a vendor-recommended temperature above zero, and a spread measured only on the locked four would have been generalised to them without evidence.

X Configuration and Replay

The complete configuration — model identifiers and pinned snapshots, release and access dates, prompts including system messages and the answer-extraction rule, decoding parameters with unavailable controls marked, reasoning-budget settings, generator source and seeds, transformation code, per-item outputs, analysis code and plotting code — is released with the paper. A single entry point regenerates every figure and table from the raw outputs and then runs the consistency audit,

which verifies that each number in the manuscript is locatable in a result file. No figure or table value is entered by hand.

Table 20: Category-B results. The transformation preserves the required procedure but the ground truth is recomputed by the generator, so answer-consistency is not computed: the answer is supposed to change, and scoring consistency would penalise correct behaviour. STR is the structural transfer ratio $A_{\mathrm{recomp}}/A_{\mathrm{orig}}$; because A_{recomp} is also bounded by base competence at that depth, STR is comparable across models rather than absolute.

Model	Family	Class	Transformation	A_{orig}	A_{recomp}	STR
GPT-5.6 Sol (reasoning: high)	Synth-Comp	B1	Numerical substitution	98.8	94.8	0.959
	Synth-Comp	B2	Counterfactual recompute	98.8	92.7	0.938
	Synth-Logic	B2	Counterfactual recompute	94.4	88.5	0.937
	Synth-Logic	B3	Truth-value perturbation	94.4	90.8	0.961
Claude Opus 5 (extended thinking)	Synth-Comp	B1	Numerical substitution	95.7	91.3	0.955
	Synth-Comp	B2	Counterfactual recompute	95.7	91.7	0.958
	Synth-Logic	B2	Counterfactual recompute	95.0	88.7	0.933
	Synth-Logic	B3	Truth-value perturbation	95.0	90.8	0.956
Gemini 3.7 Flash (thinkingLevel: high)	Synth-Comp	B1	Numerical substitution	94.0	88.0	0.936
	Synth-Comp	B2	Counterfactual recompute	94.0	88.3	0.939
	Synth-Logic	B2	Counterfactual recompute	92.3	84.2	0.912
	Synth-Logic	B3	Truth-value perturbation	92.3	86.5	0.937
Gemini 3.1 Pro (thinkingLevel: high)	Synth-Comp	B1	Numerical substitution	92.7	88.8	0.958
	Synth-Comp	B2	Counterfactual recompute	92.7	87.7	0.946
	Synth-Logic	B2	Counterfactual recompute	95.2	90.0	0.946
	Synth-Logic	B3	Truth-value perturbation	95.2	89.4	0.94
Kimi K3 (thinking)	Synth-Comp	B1	Numerical substitution	91.0	84.3	0.926
	Synth-Comp	B2	Counterfactual recompute	91.0	82.1	0.902
	Synth-Logic	B2	Counterfactual recompute	90.8	84.4	0.929
	Synth-Logic	B3	Truth-value perturbation	90.8	85.1	0.937
DeepSeek-V4-Pro (thinking)	Synth-Comp	B1	Numerical substitution	87.3	82.2	0.941
	Synth-Comp	B2	Counterfactual recompute	87.3	80.6	0.923
	Synth-Logic	B2	Counterfactual recompute	89.8	81.3	0.905
	Synth-Logic	B3	Truth-value perturbation	89.8	82.4	0.918
Qwen3.5-397B-A17B (thinking)	Synth-Comp	B1	Numerical substitution	84.3	79.0	0.938
	Synth-Comp	B2	Counterfactual recompute	84.3	77.1	0.915
	Synth-Logic	B2	Counterfactual recompute	90.0	82.7	0.919
	Synth-Logic	B3	Truth-value perturbation	90.0	82.8	0.92
Qwen3.8-27B (thinking)	Synth-Comp	B1	Numerical substitution	87.0	80.3	0.922
	Synth-Comp	B2	Counterfactual recompute	87.0	78.9	0.907
	Synth-Logic	B2	Counterfactual recompute	91.3	81.8	0.896
	Synth-Logic	B3	Truth-value perturbation	91.3	84.6	0.926
GPT-5.6 Sol (reasoning: none)	Synth-Comp	B1	Numerical substitution	87.5	79.8	0.911
	Synth-Comp	B2	Counterfactual recompute	87.5	78.9	0.902
	Synth-Logic	B2	Counterfactual recompute	87.1	78.5	0.901
	Synth-Logic	B3	Truth-value perturbation	87.1	79.8	0.917
Claude Opus 5 (thinking off)	Synth-Comp	B1	Numerical substitution	88.1	81.5	0.925
	Synth-Comp	B2	Counterfactual recompute	88.1	78.7	0.893
	Synth-Logic	B2	Counterfactual recompute	91.2	83.1	0.911
	Synth-Logic	B3	Truth-value perturbation	91.2	83.4	0.915
Qwen3.5-397B-A17B (thinking off)	Synth-Comp	B1	Numerical substitution	78.0	70.6	0.905
	Synth-Comp	B2	Counterfactual recompute	78.0	68.7	0.88
	Synth-Logic	B2	Counterfactual recompute	81.9	72.4	0.884
	Synth-Logic	B3	Truth-value perturbation	81.9	74.9	0.915
Qwen3.8-27B (thinking off)	Synth-Comp	B1	Numerical substitution	72.8	67.3	0.926
	Synth-Comp	B2	Counterfactual recompute	72.8	65.6	0.901
	Synth-Logic	B2	Counterfactual recompute	71.3	63.8	0.894
	Synth-Logic	B3	Truth-value perturbation	71.3	63.8	0.895
Qwen3.5-9B	Synth-Comp	B1	Numerical substitution	57.0	51.7	0.906
	Synth-Comp	B2	Counterfactual recompute	57.0	49.1	0.861
	Synth-Logic	B2	Counterfactual recompute	52.6	45.4	0.864
	Synth-Logic	B3	Truth-value perturbation	52.6	46.6	0.886
Qwen3.5-4B	Synth-Comp	B1	Numerical substitution	35.3	30.3	0.858
	Synth-Comp	B2	Counterfactual recompute	35.3	29.0	0.823
	Synth-Logic	B2	Counterfactual recompute	30.8	24.9	0.81
	Synth-Logic	B3	Truth-value perturbation	30.8	27.8	0.902

Table 21: Process-level results and novelty sensitivity, all models. Rationale interventions edit an observable trace and are therefore available for Class A only; "n/a" records the absence of a manipulable trace, not a missing measurement. Output-control interventions change prompting without touching a trace and apply to every model, so they are reported separately and are not faithfulness evidence. No composite index is computed: any weighting across interventions would be arbitrary and would hide which intervention drives the result. Novelty gaps are relative to familiar templates with solution depth matched, and measure behavioural sensitivity rather than training-data exposure.

		rationale intervention			output control			novelty gap (pp)		
Model	Obs.	trunc.	perm.	inj.	ans.-first	rat.-len	prefix	lex.	struct.	comp.
GPT-5.6 Sol (reasoning: high)	C	n/a	n/a	n/a	11.2	5.2	9.7	1.6	5.7	5.7
Claude Opus 5 (extended thinking)	B	n/a	n/a	n/a	11.9	7.9	9.8	1.7	5.7	7.4
Gemini 3.7 Flash (thinkingLevel: high)	B	n/a	n/a	n/a	13.2	8.0	12.0	2.5	5.1	9.4
Gemini 3.1 Pro (thinkingLevel: high)	B	n/a	n/a	n/a	13.7	9.1	10.8	1.4	5.9	8.9
Kimi K3 (thinking)	A	14.0	11.7	19.3	16.8	11.2	13.0	3.5	7.5	10.6
DeepSeek-V4-Pro (thinking)	A	12.1	12.9	18.9	17.9	8.0	12.2	3.3	7.0	10.4
Qwen3.5-397B-A17B (thinking)	A	14.0	13.9	21.6	17.0	11.8	11.8	2.7	8.6	9.9
Qwen3.8-27B (thinking)	A	13.7	14.1	23.7	18.7	10.1	14.0	3.6	8.0	9.7
GPT-5.6 Sol (reasoning: none)	C	n/a	n/a	n/a	22.2	15.0	15.9	3.4	8.5	11.7
Claude Opus 5 (thinking off)	C	n/a	n/a	n/a	20.0	16.3	14.6	5.8	9.1	11.7
Qwen3.5-397B-A17B (thinking off)	A	21.1	17.5	30.8	25.2	14.1	15.6	4.1	10.5	11.9
Qwen3.8-27B (thinking off)	A	20.5	16.0	31.9	25.9	13.3	19.0	3.7	8.4	11.2
Qwen3.5-9B	A	22.3	17.8	37.8	27.4	20.1	19.8	2.3	5.7	9.9
Qwen3.5-4B	A	22.9	21.4	39.2	28.9	16.5	22.1	2.7	4.4	6.3

Table 22: Precision analysis. Minimum detectable effect size (MDES, percentage points) for the paired Category-A comparison at $\alpha = 0.05$, Bonferroni-bounded over $m = 5$ classes, at 80% power. The discordance rate is taken in its general form for *unequal* marginals, $\pi_d = p_1(1-p_2) + p_2(1-p_1) - 2\rho\sqrt{p_1 q_1 p_2 q_2}$, and $\mathrm{MDES} = (z_{1-\alpha/2m} + z_{0.8})\sqrt{\pi_d/N}$. Because p_2 depends on the effect being detected, the expression is implicit and is solved by fixed-point iteration rather than evaluated once. The familiar $\pi_d = 2p(1-p)(1-\rho)$ is the special case $p_1 = p_2$, which is the null this design exists to test, and it understates the MDES by up to 13% over this grid — the larger the effect, the more it understates. Each cell is confirmed by simulation: drawing paired outcomes at the solved MDES and applying McNemar's mid-p test at the corrected level gives empirical power 0.79–0.82 against the nominal 0.80, so the derivation is checked rather than asserted. Sample size is read off this grid rather than asserted: at the chosen $N = 1{,}200$ the resolvable difference ranges from about 2.4 to 5.5 pp depending on baseline accuracy and paired correlation, and the applicable figure is quoted beside each comparison. The last two blocks carry the clustering: seed templates rather than items are the unit of generalisation, so the effective sample size is $N_{\text{eff}} = N/\text{deff}$ with $\text{deff} = 1.65$ from a measured observed-scale $\text{ICC} = 0.024$, and the clustered MDES is obtained by simulating the design that will be run — T templates, each a cluster sharing a latent offset, tested by the template-level permutation test — rather than by substituting N_{eff} into the independent expression, which assumes the clustered variance is a fixed multiple of the unclustered one.

	$p = 0.7$			$p = 0.8$			$p = 0.9$		
N	$\rho = 0.3$	$\rho = 0.5$	$\rho = 0.7$	$\rho = 0.3$	$\rho = 0.5$	$\rho = 0.7$	$\rho = 0.3$	$\rho = 0.5$	$\rho = 0.7$
600	7.87	6.63	5.11	7.06	5.92	4.55	5.57	4.65	3.55
1,200	5.5	4.63	3.58	4.89	4.11	3.16	3.8	3.18	2.43
2,400	3.85	3.25	2.51	3.41	2.87	2.21	2.62	2.2	1.69
4,800	2.71	2.29	1.77	2.39	2.01	1.55	1.82	1.53	1.18
Equal-marginal special case at $N = 1{,}200$, for comparison:									
1,200	5.35	4.52	3.5	4.67	3.95	3.06	3.5	2.96	2.29
Simulated MDES at the achieved power, same grid:									
1,200	5.58	4.68	3.61	4.91	4.12	3.19	3.83	3.25	2.48
Clustered MDES by simulation of the template design, same grid:									
1,200	7.0	5.72	4.41	6.23	5.23	3.9	4.84	3.93	3.1
Effective sample size N_{eff} behind that row:									
1,200	726.2	726.2	726.2	726.2	726.2	726.2	726.2	726.2	726.2

Table 23: One instantiation of each transformation class applied to a single source item.

Class	Transformation
Source	A shop sells notebooks at \$4 each and pens at \$2 each. Buying 3 notebooks and 5 pens costs how much?
Category A — answer preserved	
A1	A vendor offers ledgers at \$4 each and styluses at \$2 each. Buying 3 ledgers and 5 styluses costs how much?
A2	Buying 5 pens and 3 notebooks, priced at \$2 and \$4 respectively, costs how much?
A3	What is the total cost of 3 items priced at \$4 and 5 items priced at \$2?
A4	A shop sells notebooks at \$4 each and pens at \$2 each. The shop also stocks erasers. Buying 3 notebooks and 5 pens costs how much?
A5	A shop charges \$4 per notebook and \$2 per pen. What is the cost of 3 notebooks and 5 pens?
Category B — structure preserved, ground truth recomputed	
B1	. . . notebooks at \$4.30 each and pens at \$2.10 each. Buying 3 and 5 costs how much?
B2	. . . notebooks at \$4 each and pens at \$3 each. Buying 3 and 5 costs how much?
B3	SYNTH-LOGIC: one premise's truth value is flipped and the entailment is recomputed.

Table 24: Feature comparison against the closest located work. "Paired" indicates whether each transformed item is scored against its own source item; "C×P" whether perturbation sensitivity is estimated as a function of complexity; "Ladder" whether a depth ladder is used and how far it reaches. Venue is the publication of record where one was confirmed against the proceedings or the publisher, and *preprint* where no peer-reviewed version could be located; a sub-track is named as such rather than folded into its main conference.

Work	Venue	Paired	C×P	Ladder	Primary contribution
GSM-Symbolic (Mirzadeh et al., 2025)	ICLR 2025	yes	no	–	symbolic templating
MATH-Perturb (Huang et al., 2025)	ICML 2025	yes	no	–	solution-path-invalidating perturbations
PutnamGAP (Hao et al., 2025)	preprint 2025	yes	no	–	surface vs structural generalization
AR-Checker (Hou et al., 2026)	Findings of ACL 2026	yes	no	–	contamination-resistant variants
Illusion of Thinking (Shojaee et al., 2025)	NeurIPS 2025	no	no	short	complexity cliff, effort scaling
OMEGA (Sun et al., 2025a)	NeurIPS 2025	partial	no	yes	exploratory/compositional axes
DeepRD (Rameshkumar et al., 2025)	IJCNLP-AACL 2025	no	no	yes	unbounded-complexity generation
GrAlgoBench (Zhang et al., 2026)	preprint 2026	partial	no	yes	long-context and over-thinking limits
Breaking the Chain (Somov et al., 2026)	preprint 2026	partial	no	–	causal mediation over structures
Evaluatology (Zhan et al., 2024)	TBench 2024	yes	no	–	factorial decomposition of evaluation
This work	–	**yes**	**yes**	**n=3–12, 2 families**	**4 dimensions + C×P interaction**