TRUSTCONS: EXACTLY VERIFYING CONSISTENCY, COMPLETENESS, AND CONDITIONED RE-ENUMERATION IN LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reliable constraint reasoning requires more than a correct satisfiability verdict: a model should recover the target set, preserve answers under meaning-preserving changes, and — when told that a candidate set contains exactly one deterministic error — return the exact target set. We introduce TRUSTCONS, an exhaustive-oracle diagnostic with 48 boundary-selected Boolean instances, five views, and three tasks, plus a matched extension that holds visible formula dimensions fixed while raising the exact solution count from 1 to 32. An exploratory run exposed a completion-budget confound; we froze it, reran complete matrices under larger limits, treated cap-hit unparseable outputs as missing, and re-verified every solution set with Z3 5.0.0. Because v2 re-issued complete matrices rather than selected cells, 7,344 cells carry both a v1 and a v2 record under one configuration, instance, view, task, and prompt, so the correction is measured rather than asserted: the larger caps moved 3,184 censored cells into the observed population, 2,839 of them to strict correctness, a paired change of 38.4 pp (95% base-cluster interval [34.5, 42.1] pp). Unconditional worst/best-case macro ranges are 45.7–68.2%, 50.2–57.8%, and 25.7–64.2% for decision, low-cardinality exact-set recovery, and the candidate-conditioned third task; the pooled point estimates inside them are record-level averages that include the below-floor cells, and equal-weighting configurations yields different values. Among 192 fully observed decision/enumeration pairs from 4 configurations, 5.2% (95% base-cluster interval 2.1–8.9%) are decision-correct but enumeration-wrong. The third task returns the model's own enumeration set on 89.8% of strictly parsed pairs and, on those 147 parsed pairs, is correct about equally often (7 against 7, exact McNemar $p = 1.00$; unparseable answers are missing, not incorrect), so it is scored as conditioned re-enumeration, not as evidence of detection, localisation, or repair. TrustCons is a small synthetic diagnostic, not a safety certificate; its contribution is protocol-aware joint measurement and a descriptive stress test of exact-set scaling under matched visible dimensions.

1 INTRODUCTION

A binary verdict can conceal the failure that matters. If a formula is satisfiable, answering "SAT" proves neither that a model can recover every witness nor that it can reject every invalid assignment. A downstream configurator may fail when a single feasible case is omitted even though the existence verdict is correct. The same model may also change its answer when rules are reordered or variables renamed, or fail to return the exact target set after being told that a candidate contains exactly one deterministic error. These are different reliability properties and should not be collapsed into one accuracy score.

Existing resources illuminate complementary parts of this problem. RuleTaker, ProofWriter, FOLIO, and LogicBench target controlled deduction (Clark et al., 2020; Tafjord et al., 2020; Han et al., 2022; Parmar et al., 2024); SATBench, CP-Bench, and Pencil Puzzle Bench add solver-checked generation, constraint modelling, and dynamic verifiable puzzles (Wei et al., 2025; Michailidis et al., 2025;

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.

Table 1: Claim-specific comparison. Entries describe the principal evaluated object and intervention; they are not binary claims of exclusivity. The third task is scored as conditioned re-enumeration rather than a repair capability (Section 2), so it is not listed as a separating dimension; the last column names the protocol and endpoint properties TrustCons measures and its nearest neighbours do not report.

Resource	Truth access	Transformation ownership	Language diversity	Protocol censoring and endpoint separation
SATBench	solver labels/witness	benchmark-authored puzzles	broad templates	none
BeyondBench	deterministic proofs / complete sets	isomorphic generators	44 algorithmic tasks	none
LogiConBench	task-specific checks	reasoning-path variants	broad logic text	limited
BECEL / ConsisEval	task labels	paired behavior/difficulty	natural tasks	none
Consistency Checker	derived consistency	reversible tree transforms	controlled text	none
CRTBench	solver verdict	controlled reformulations	controlled text	none
Query4Regex	DFA equivalence	query/regex transforms	structured text	none
Ghosh et al. (2025)	fact-check labels	meaning-preserving queries	KG questions	none
TrustCons	complete finite truth set	five owned single-factor views	controlled CNF text	measured protocol censoring with worst/best bounds; strict versus lenient endpoint separation; matched-dimension cardinality ladder with structural-drift audit

Waugh, 2026). Most importantly, BeyondBench already combines algorithmic generation, deterministic verification, fully enumerated solutions, isomorphic transformations, broad task coverage, and repeated evaluation (Srivastava et al., 2026), so those ingredients are not individually novel here. BECEL and related resources study behavioral consistency or formally checkable transformations (Jang et al., 2022; Yang et al., 2024; Hong et al., 2025; Gu & Chen, 2026; Hahn & Han, 2026; Chen et al., 2026; Ghosh et al., 2025). Tool-augmented Logic-LM, SatLM, and LINC delegate inference to symbolic solvers (Pan et al., 2023; Ye et al., 2023; Olausson et al., 2023); TrustCons instead evaluates unaugmented API responses and does not estimate that solver-coupled ceiling. Its narrower contribution jointly audits interface compliance, semantic exactness, pure-view agreement, conditioned re-enumeration, censoring, and matched solution-cardinality interventions under one finite oracle.

This joint design matters for both measurement and audit. Exact solution sets distinguish omission from insertion without a learned judge; views are repeated measurements of the same base formula, so invariance can be separated from semantic correctness; and third-task inputs are corrupted independently of model behavior, avoiding evaluation only on errors a model happened to produce. Because the prompt discloses the single error, the endpoint is reported as a condition on enumeration rather than as a repair capability. Finally, immutable request records make protocol failures visible: in our own exploratory run low output limits interacted with hidden reasoning budgets and contaminated apparent accuracy, and rather than hide this failure we treat the correction as part of the empirical contribution and measure it as a paired transition.

We contribute: (1) a measured evaluation failure, in which a shared reasoning and output budget left 2,549 responses censored even after the caps were raised, so an observed-case accuracy is a selected quantity rather than a capability estimate; (2) an auditable remedy — protocol versioning, complete-matrix replacement, explicit missingness bounds, and a joint endpoint that separates a configuration perfect on one rendering from one that survives re-rendering; and (3) a joint diagnostic of literal interface compliance, semantic exact-set recovery, view agreement, and candidate-conditioned re-enumeration, with a matched-visible-dimension 1–32-solution stress test across three separately keyed calls. The evidence supports claims only about the tested finite domain and dated configurations.

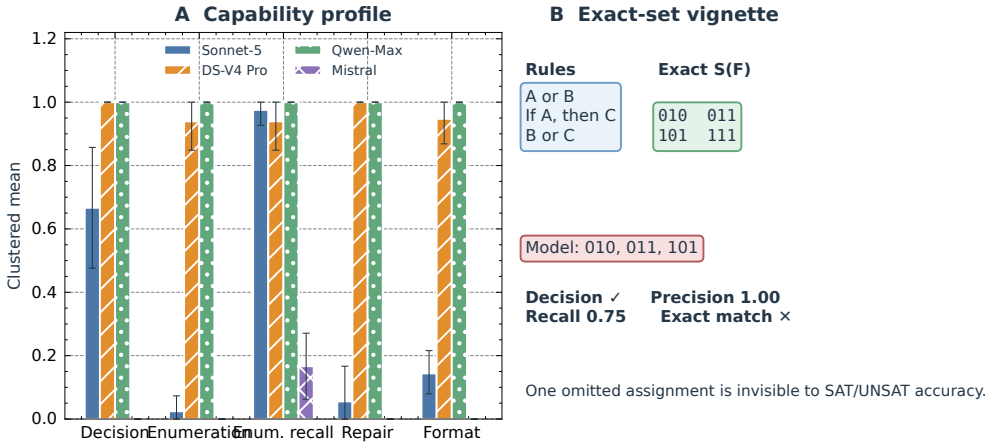

Figure 1: Strict-endpoint capability profile for four representative protocol-valid configurations (left) and an exact-set vignette (right). A correct SAT verdict can coexist with an incomplete solution set. Bars are base-instance cluster means over base-view zero-shot records with the three tasks pooled inside each bar, so the format bar weights records rather than configurations and each bar has its own base count; because the strict endpoint requires one whole-content JSON object, a configuration whose answers arrive inside code fences scores zero there by construction. Intervals and denominators appear in Table 2. Repair bars for the three low-support configurations are point estimates over 2, 8, and 8 observed bases; Table 2 withholds their observed rates under the support floor, so those bars should be read with their denominators rather than as stable configuration rates.

2 Formal Task Definition

Let $x = (x_1, \ldots, x_n) \in \{0,1\}^n$ and let F be a CNF formula:

$$F(x) = \bigwedge_{j=1}^{m} C_j(x), \qquad C_j(x) = \bigvee_{\ell \in L_j} \ell(x).$$

The exhaustive truth set is $\mathcal{T}(F) = \{x : F(x) = 1\}$. Decision predicts $\mathbf{1}[\mathcal{T}(F) \neq \varnothing]$. Enumeration predicts a set $\widehat{\mathcal{T}}$ and is exactly correct iff $\widehat{\mathcal{T}} = \mathcal{T}(F)$. Precision and recall diagnose insertions and omissions but do not replace exact match. For the third task, a deterministic corruption operator g produces $\widetilde{\mathcal{T}} = g(\mathcal{T}(F)) \neq \mathcal{T}(F)$ without querying the tested model. The prompt discloses that $\widetilde{\mathcal{T}}$ contains exactly one deterministic error, and the response must return $\mathcal{T}(F)$ together with a boolean flag $\widehat{e}$.

$$s_{\text{reenum}} = \begin{cases} 1, & \widehat{e} = 1 \ \wedge \ \widehat{\mathcal{T}} = \mathcal{T}(F), \\ 0, & \text{otherwise}, \end{cases} \tag{1}$$

Because the single-error guarantee is given to the model rather than inferred by it, and because the output schema asks for nothing further, this endpoint is a *conditioned re-enumeration* score, not a detection or repair measure: a model can ignore $\widetilde{\mathcal{T}}$, enumerate F from scratch, and report $\widehat{e} = 1$. The flag conjunct is empirically non-binding — 0 recovered flags were `false` — so on these data the score coincides with exact set equality, and Section 6.1 shows it does not separate from plain enumeration. Where the released artifacts use the internal identifier `repair` — the request namespace and the Table 4 row label — it denotes this score and nothing more.

For base, reorder, rename, and equivalent views, let π_v be the variable-name bijection and require $\pi_v^{-1}(\mathcal{T}(v(F))) = \mathcal{T}(F)$. The distractor view introduces irrelevant variables z through tautologies, so its target lies in a larger space; there we require equality only after projection ρ_x onto the original coordinates:

$$\pi_v^{-1}(\mathcal{T}(v(F))) = \mathcal{T}(F) \quad (v \neq \text{distractor}), \qquad \rho_x(\mathcal{T}(v_{\text{dist}}(F))) = \mathcal{T}(F).$$

Thus distractors may extend a valid x by arbitrary irrelevant z values but may not change the projected target; $\mathcal{V} = \{\text{base, reorder, rename, equivalent, distractor}\}$. This equality concerns the formal

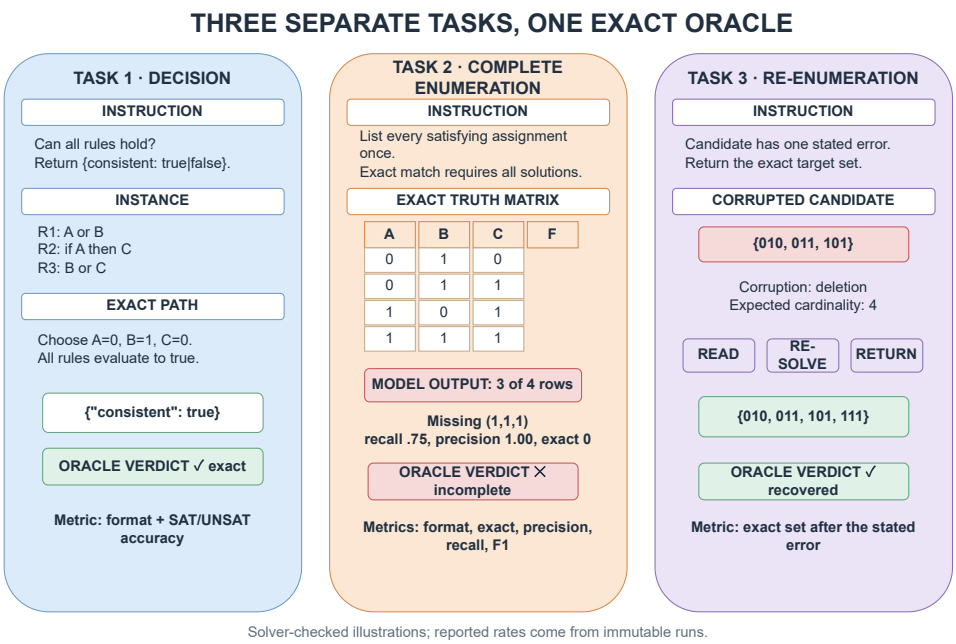

Figure 2: Three tasks share the same truth set but expose different failure modes. Decision checks existence; enumeration checks set equality; the third task supplies an independently corrupted candidate under a disclosed single-error guarantee and checks only whether the exact target set comes back, which is a conditioned re-enumeration score rather than a measurement of detection, localisation, or repair.

target, whereas output invariance concerns the model: a perfectly invariant but uniformly wrong model is not robust. We therefore report correctness by view, exact output agreement, and all-views-correct separately. Figure 3 draws the full evidence path from construction to manuscript artifact.

3 Corpus Construction and Oracle Audit

For each tier $n \in \{4, 5, 6, 7\}$, 12 independent seeds add unique random clauses until crossing the satisfiability boundary, then retain the SAT or UNSAT side by seed parity. The release is balanced (24/24) but does not contain dependent SAT/UNSAT pairs; each identifier is one base sampling unit. Exact sampling and stopping rules are in Appendix A.2.

Five renderers isolate one declared factor. Reorder permutes only rule order. Rename bijectively changes target symbols. Equivalent rendering rewrites disjunctions as implication forms without reordering. Distractor appends two irrelevant tautologies and is checked after projection onto the target coordinates. Test prompts and demonstrations use disjoint seeds and namespaces. The primary parser requires the entire trimmed response to be one JSON object with exactly the task keys; set outputs must contain unique, ascending, length-n bitstrings. The historical code-fence/prose extractor is retained only as a named lenient sensitivity endpoint.

The primary oracle evaluates all 2^n assignments in Python. An independent audit invokes the official Z3 5.0.0 binary and checks every assignment under every view. All 240 primary instance-view solution sets agree among Z3, brute force, and stored gold, and each repair candidate differs from gold; a second sha-recorded audit re-verifies the 144 matched-extension sets under the same binary, so every released truth set is covered. Neither audit validates the naturalness of the English renderings, but both remove solver disagreement as an explanation of the reported model errors.

The boundary generator is deliberately small and produces low-cardinality truth sets: 24 bases are UNSAT, 19 have one solution, and five have two or three; none has four or more. Thus 43/48 bases (89.6%) have at most one solution (Figure 4). Enumeration results measure exact-set recovery in this low-cardinality regime, not completeness under long output sets.

To hold visible formula dimensions fixed while varying solution-set size, we preregistered a matched extension after identifying the primary corpus skew. Each of 24 randomized families contains six eight-variable, sixteen-clause binary formulas sharing a three-variable core and a fixed clause-count and clause-length profile; releasing zero through five selectors changes the exact set size to 1, 2, 4, 8, 16, or 32 by changing the released selector's anchor-literal polarity as well as its gating, so clauses are not literally identical within a family. Four representative configurations receive base-view zero-shot enumeration under three separately keyed provider calls per cell (1,728 calls; USD 4.1630), at temperature zero but with no provider seed, so repetitions measure end-to-end repeatability rather than statistically independent or seeded Monte Carlo variance. Matching the visible dimensions does not match reasoning difficulty: releasing a selector from zero to five raises the number of free variables from 0 to 5 and lowers the number of variables held at a constant value from 8 to 3, while clause count, clause-length profile, mean variable occurrence (4.0) and the node count of a plain unit-propagating DPLL search (3) stay fixed (Table 26). The ladder therefore varies how long the exact answer is rather than how hard the search is, so we report it as a descriptive stress test of exact-set scaling under matched visible dimensions rather than as an isolated effect of solution count.

The third task is likewise instance-conditioned: UNSAT items receive a spurious assignment, multi-solution items lose one, and singleton items receive a replacement, so corruption kind is confounded with satisfiability and cardinality. Because Section 6.1 shows the endpoint does not separate from enumeration, these strata describe conditioned re-enumeration rather than repair proficiency.

4 Evaluation Protocol and the Token-Budget Correction

The original protocol issued 12,672 calls with task caps of 96/1,024/1,152 tokens, including reasoning. Of 7,680 high-reasoning records, 6,050 hit the cap and 5,598 were also unparseable (5,559 empty). Because the truncated answer is unknowable, these are censored rather than incorrect.

We froze v1 and versioned request keys. To avoid selective retry, v2 replaces complete broad and focal matrices, yielding 7,344 unique requests at 1,280/2,560/2,560-token caps with messages, usage, finish reason, actual model, response ID, and cost retained. The route field is null in all 21,904 responses, so no backend-routing claim is made (Table 23).

GPT-6 Astra and Claude Fable 5.1 remain exploratory v1-only configurations because complete repairs would violate the remaining budget; they are excluded from primary rankings. Mistral Large showed no cap-hit/unparseable pattern and is retained from v1. The primary breadth is therefore 12 protocol-valid configurations. Two no-thinking DeepSeek focal settings remain auxiliary v1 comparisons, with any cap-hit unparseable cells marked NA. If a v2 response still reaches its larger cap without valid JSON, it remains NAand is not called again.

Composition is outcome-independent: v1 supplies 12,672 calls, 672 exploratory cells are dropped, and v2 replaces 7,344 remaining cells. The 4,656 retained v1 records comprise two auxiliary no-thinking DeepSeek focal matrices ($2 \times 2{,}160$) and the 336-cell protocol-valid Mistral broad matrix. The result has 12,000 cells, 9,414 observed; primary statistics use only the 12 protocol-valid configurations. Table 23 registers this population, not total spend.

Whole-matrix replacement makes the correction measurable: 7,344 cells are paired on configuration, instance, view, task, and prompt. Larger caps moved 3,184 v1-censored cells into observation and 2,839 to strict correctness, while 62 moved the other way. Observability changes by 42.5 pp (95% base-cluster interval [38.5, 46.4] pp) and strict correctness by 38.4 pp ([34.5, 42.1] pp). Table 24 gives the full task-stratified transition matrix. Because the intervention crosses time, it measures the correction but does not isolate a cap effect.

Worst-case v2 cost was USD 53.2147, and adding the USD 41.8781 already spent yielded USD 95.0929, below the USD 100 hard ceiling; the catalog and reservation rules are detailed in the appendix.

5 Metrics and Statistical Inference

Literal whole-content JSON is the primary interface metric. A secondary *lenient JSON extraction* analysis reproduces the historical code-fence/prose extraction behavior; a still more conservative

recovery analysis accepts only bare unambiguous answers. Neither sensitivity endpoint replaces the declared interface contract.

All confidence intervals resample the 48 base identifiers, retaining their repeated views. Macro results average over actually tested protocol-valid configurations and describe those configurations, not a population of language models. Paired contrasts use only base identifiers observed in both conditions, and every prompt contrast is additionally bounded over the full requested population so a selected subset cannot be read as an unconditional effect. Exact two-sided McNemar tests report discordant directions and the complete-pair denominator, with Holm correction inside declared view or prompt families. In the matched extension three calls are averaged within each formula family, only families with all six $K \in \{1, 32\}$ calls observed enter the contrast, and an exact two-sided family sign test is Holm-adjusted across the four strict system contrasts. Missing pairs are counted, never filled with zero.

6 RESULTS

6.1 DECISION, ENUMERATION, AND THE CANDIDATE-CONDITIONED TASK SEPARATE

Across the 12 configurations, assigning every missing response first incorrect and then correct gives unconditional macro ranges of 45.7–68.2%, 50.2–57.8%, and 25.7–64.2% for zero-shot base decision, low-cardinality exact-set recovery, and the candidate-conditioned third task. These average each configuration's own worst and best endpoint, so they do not depend on how observed cells are pooled.

Observed-case rates are 59.0% (55.9–61.9%, $n = 446$), 54.3% (52.4–56.2%, $n = 532$), and 41.8% (38.4–45.0%, $n = 354$). Because cap-hit missingness is informative, they must be read with the unconditional ranges above. Equal-configuration means are 65.0%, 56.3%, and 57.3%; the last is inflated by three 100% cells supported by only 2, 8, and 8 bases. We therefore withhold observed cells below 12 of 48 bases and report only bounds; 3 repair configurations are excluded and the surviving mean is 43.1%.

The revision-time floor is not preregistered; its full sweep is 57.3% at a support floor of 1 (12/12 configurations), 53.4% at a support floor of 6 (11/12 configurations), 43.1% at a support floor of 12 (9/12 configurations), 42.1% at a support floor of 24 (7/12 configurations), 34.7% at a support floor of 48 (3/12 configurations)(decision: 65.0% at a support floor of 1, 65.0% at a support floor of 6, 65.0% at a support floor of 12, 61.3% at a support floor of 24, 46.4% at a support floor of 48). A quarter removes only the three lowest-support repair cells, whereas 24/48 would discard five of twelve. No defensible floor stabilises the configuration mean, so the pooled rate remains the headline.

Table 2 reports strict/lenient rates, denominators, and bounds, marking withheld cells with a dagger. Pooled literal-format rates are 60.8%, 56.6%, and 44.9%; configuration-first rates and their repair weighting gap appear there as sensitivity results.

The decision, enumeration, and interface axes are empirically nonredundant, but the size of their separation is constrained by corpus composition. Across the 4 configurations with all 48 strict decision and enumeration outcomes observed, 192 paired cells give 5.2% (95% base-cluster interval 2.1–8.9%) that are decision-correct but enumeration-wrong. Because 43/48 bases have at most one solution, enumeration is close to a decision-plus-witness task on most items; the high agreement therefore should not be generalized to long solution sets. Of 1,332 observed protocol-valid zero-shot base task cells, 23.9% (22.2–25.6%) are correct under lenient semantic extraction but fail the literal interface. Among 1,055 complete focal view settings, 8.9% (6.8–11.1%) are output-invariant yet not correct in every view; invariance is judged on the canonicalised answer, so a setting that fails the literal interface identically in all five views counts as invariant. These paired discordances, rather than benchmark assembly alone, motivate joint measurement.

The third task, by contrast, is not separately identified. Of the 348 issued pairs matched on configuration and base, 147 parse under both endpoints and agree with the model's own enumeration answer in 89.8% of those; on those parsed pairs, the two are correct on 138 and 138 instances respectively, and the discordant counts (7 third-task-only versus 7 enumeration-only) do not differ (exact McNemar $p = 1.00$; unparseable answers are missing, not incorrect, so no non-pair is counted). Of

Table 2: Primary zero-shot base results for the 12 protocol-valid configurations. The composed matrix holds two further v1 focal settings (the no-thinking DeepSeek pair); they are reported only as auxiliary comparisons and appear in neither this table nor the pooled summaries. Each cell reports observed-case strict/lenient accuracy (O), unconditional strict–lenient worst–best bounds (B), and the observed denominator n out of 48. Bounds are shown for every incomplete cell; no observation threshold licenses an unconditional point estimate. †Observed rate withheld: fewer than 12 of 48 bases are complete, so only the bounds are printed and the cell is excluded from configuration-mean summaries. Daggered repair cells nevertheless remain observed records in the pooled headline: pooled repair correctness is 148 of 354 observed cells (41.8%), or 130 of 336 (38.7%) when the 18 daggered cells are excluded.

Configuration	Decision S/L	Enumeration S/L	Repair S/L
claude-sonnet-5	O 66.7/100.0 B [29.2,85.4]/[43.8,100.0] n=21	O 2.4/97.6 B [2.1,16.7]/[83.3,97.9] n=41	O 5.6/94.4 B [2.1,64.6]/[35.4,97.9] n=18
deepseek-v4-flash-high	O 100.0/100.0 B [68.8,100.0]/[68.8,100.0] n=33	O 91.9/91.9 B [70.8,93.8]/[70.8,93.8] n=37	O NA/NA† B [16.7,100.0]/[16.7,100.0] n=8
deepseek-v4-pro-high	O 100.0/100.0 B [54.2,100.0]/[54.2,100.0] n=26	O 93.9/93.9 B [64.6,95.8]/[64.6,95.8] n=33	O NA/NA† B [4.2,100.0]/[4.2,100.0] n=2
gemini-3.1-pro-preview	O 2.1/97.9 B [2.1,2.1]/[97.9,97.9] n=48	O 0.0/85.4 B [0.0,0.0]/[85.4,85.4] n=48	O 4.2/62.5 B [4.2,4.2]/[62.5,62.5] n=48
gemini-3.8-flash	O 27.7/87.2 B [27.1,29.2]/[85.4,87.5] n=47	O 12.5/97.9 B [12.5,12.5]/[97.9,97.9] n=48	O 10.6/80.9 B [10.4,12.5]/[79.2,81.2] n=47
gpt-5.6-luna	O 83.3/83.3 B [83.3,83.3]/[83.3,83.3] n=48	O 79.2/79.2 B [79.2,79.2]/[79.2,79.2] n=48	O 80.0/80.0 B [75.0,81.2]/[75.0,81.2] n=45
grok-4.6	O 100.0/100.0 B [100.0,100.0]/[100.0,100.0] n=48	O 97.9/97.9 B [97.9,97.9]/[97.9,97.9] n=48	O 100.0/100.0 B [100.0,100.0]/[100.0,100.0] n=48
muse-spark-1.3	O 100.0/100.0 B [39.6,100.0]/[39.6,100.0] n=19	O 97.7/97.7 B [89.6,97.9]/[89.6,97.9] n=44	O 100.0/100.0 B [50.0,100.0]/[50.0,100.0] n=24
qwen3.8-a95b	O 100.0/100.0 B [66.7,100.0]/[66.7,100.0] n=32	O 100.0/100.0 B [87.5,100.0]/[87.5,100.0] n=42	O 87.5/87.5 B [29.2,95.8]/[29.2,95.8] n=16
qwen3.8-max	O 100.0/100.0 B [77.1,100.0]/[77.1,100.0] n=37	O 100.0/100.0 B [97.9,100.0]/[97.9,100.0] n=47	O NA/NA† B [16.7,100.0]/[16.7,100.0] n=8
phi-4	O 0.0/28.2 B [0.0,18.8]/[22.9,41.7] n=39	O 0.0/0.0 B [0.0,0.0]/[0.0,0.0] n=48	O 0.0/7.1 B [0.0,12.5]/[6.2,18.8] n=42
mistral-large	O 0.0/47.9 B [0.0,0.0]/[47.9,47.9] n=48	O 0.0/0.0 B [0.0,0.0]/[0.0,0.0] n=48	O 0.0/0.0 B [0.0,0.0]/[0.0,0.0] n=48

156 strict predictions, 0 echoed the corrupted candidate, so the disclosed candidate is a convenience rather than something adopted. Table 25 gives the per-configuration pairing. We therefore base no detection, localisation, or repair claim on this endpoint, though it remains a useful failure channel as the task whose censoring survives the larger caps.

6.2 Pure semantic views

The full five-view matrix exists only for four focal DeepSeek configurations. Reordering, renaming, equivalent expression, and tautological distraction preserve the mapped oracle exactly, but complete-five-view denominators vary sharply, so Table 3 withholds the conditional base and worst rates wherever complete-five-view support falls below 12 of 48 bases, prints NA in their place, and reports unconditional worst–best bounds for invariance and all-views-correct in every cell. A single-rendering score and this five-view conjunction are not the same measurement: on the 11 bases where all four focal configurations return a complete five-view response, DeepSeek V4 Pro High is at 100.0% on the base rendering but 81.8% on the conjunction, while V4 Flash High is at 90.9% and 90.9%, reversing the single-rendering order. With 11 common bases that describes two configurations rather than a population and carries no test; these results otherwise characterize the focal interfaces only.

6.3 Prompts, difficulty, and censoring sensitivity

Three-shot and certificate prompts move observed decision from 59.0% to 60.6% and 57.0%, and enumeration from 54.3% to 56.1% and 57.5%. Paired base-cluster changes are +3.7pp (+0.5 to +7.0pp), then -3.6pp (-6.5 to -0.6pp), and +2.0pp (+0.0 to +4.0pp), but censoring varies by prompt.

Table 3: Zero-shot focal DeepSeek view results. Base and worst are conditional on the complete-five-view subset shown in the denominator; invariance and all-correct are unconditional worst–best bounds over all 48 bases. †Fewer than 12 of 48 bases are complete: the conditional base and worst rates are withheld as NA and only the unconditional bounds are reported.

Configuration	Task	Complete	Base	Worst	Invariance bounds	All-correct bounds
deepseek-v4-flash-none	decision	48/48	52.1	35.4	[16.7,16.7]	[6.2,6.2]
deepseek-v4-flash-none	enumeration	45/48	20.0	8.9	[6.2,12.5]	[2.1,8.3]
deepseek-v4-flash-none	repair	47/48	8.5	4.3	[12.5,14.6]	[2.1,4.2]
deepseek-v4-flash-high	decision	15/48	100.0	100.0	[31.2,100.0]	[31.2,100.0]
deepseek-v4-flash-high	enumeration	29/48	93.1	93.1	[54.2,93.8]	[54.2,93.8]
deepseek-v4-flash-high	repair	2/48†	NA	NA	[2.1,97.9]	[2.1,97.9]
deepseek-v4-pro-none	decision	46/48	47.8	43.5	[0.0,4.2]	[0.0,4.2]
deepseek-v4-pro-none	enumeration	44/48	0.0	0.0	[56.2,64.6]	[0.0,8.3]
deepseek-v4-pro-none	repair	43/48	7.0	4.7	[8.3,18.8]	[4.2,14.6]
deepseek-v4-pro-high	decision	12/48	100.0	100.0	[25.0,100.0]	[25.0,100.0]
deepseek-v4-pro-high	enumeration	12/48	100.0	83.3	[20.8,95.8]	[20.8,95.8]
deepseek-v4-pro-high	repair	0/48†	NA	NA	[0.0,100.0]	[0.0,100.0]

Worst/best fills give -24.7 to $+26.0\%$ and -31.6 to $+26.2\%$, both spanning zero, so prompt direction is not identified over the requested population. Figure 5 and the appendix report supports and reversals.

After larger caps, 2,549 responses remain censored, unevenly by view and task; the shortest equivalent-expression view is least censored, so complete-view subsets are conditional. Lenient extraction yields 84.8%, 77.1%, and 64.4% pooled, making the strict–lenient gap an interface sensitivity. Conservative bare-answer recovery moves macro rates by at most 0.0 percentage points without changing configuration order. Tables 2–3 and 24 give all observed/missing denominators and transitions.

6.4 Matched visible-dimension solution-cardinality stress test

The extension contains 144 formulas and 1,728 responses, with 69 censored. From one to 32 solutions, family-averaged strict recovery changes from 98.6% to 97.9% (Flash), 37.5% to 8.3% (Gemini), 69.4% to 66.7% (Luna), and 0.0% to 0.0% (Phi-4). Only Gemini's paired family contrast survives Holm correction ($p = 0.02954$, 23 complete families), and it alone is significant unadjusted, so correction creates no finding. Flash's marginal rates move although all complete endpoint families tie: marginal rates use every available response, whereas the paired contrast uses only complete $K = 1/K = 32$ families. Flash and Phi-4 also sit near ceiling and floor. The result is configuration-specific, and fixed DPLL nodes do not remove the autoregressive burden of producing 32 unique sorted strings. Full differences, intervals, bounds, repetition summaries, and denominator details are in Tables 12 and 26.

7 Discussion

The methodological result is actionable: reasoning-enabled API evaluation is uninterpretable unless the completion budget leaves room for the answer and logs separate termination from semantic failure. Complete-matrix replacement avoids selective retry and measures the correction: 3,184 censored cells became observable, 2,839 strict-correct, while 62 moved the other way.

TrustCons separates semantic equivalence, answer invariance, and all-views correctness; completeness is set equality, not one witness. The third task is conditioned re-enumeration, not repair: the prompt discloses one error, the flag never returns false, and the endpoint nearly always matches the model's enumeration. Decision/enumeration, strict/lenient, and invariant/wrong discordances are empirically nonredundant but reveal no internal mechanism.

Limitations. The primary corpus has only 48 controlled formulas, at most seven variables, and 43/48 have at most one solution. The matched extension uses synthetic eight-variable 2-CNF (polynomial-time classically) and four selected configurations; structural complexity and cardinality are not fully separated, and three temperature-zero calls measure end-to-end repeatability, not seeded sampling. The third task cannot separate candidate use from re-enumeration; moreover, its

frozen prompt requests diagnosis without a diagnosis field and omits the parser's ascending-order requirement. Detection, localisation, diagnosis, and repair are therefore unidentified, and corruption-type contrasts are confounded. The study tests unaugmented calls without a solver-coupled ceiling. Its API sample is dated and budget-limited, full views cover only DeepSeek, censoring remains informative, and backend routing is unavailable. The paired cap correction crosses time and cannot exclude provider drift. These scores certify neither deployment safety nor general reasoning competence.

8 CONCLUSION

TrustCons connects satisfiability decision, exact-set recovery, focal pure-view analysis, and deterministic candidate-conditioned re-enumeration under one finite oracle. Relative to broader algorithmic resources such as BeyondBench the new knowledge is narrower: literal-interface and semantic endpoints can diverge, shared reasoning/output budgets create informative missingness that can be measured as a paired transition rather than asserted, and exact-set recovery can be stressed under a matched-visible-dimension 1–32-solution ladder whose structural drift we report. Freezing the flawed run, replacing full matrices, bounding missingness, and independently checking view oracles yields a defensible protocol study rather than a universal leaderboard or trust certificate.

Ethics Statement

TrustCons contains synthetic Boolean constraints and no personal data. Its principal risk is interpretive: readers could mistake a narrow logic diagnostic for a certification of safety or trustworthiness. We explicitly prohibit that inference. Commercial API names, costs, and recorded model slugs are historical observations; the serving stack returned no backend provider identifier for any request, so no routing claim and no provider endorsement is implied.

Reproducibility Statement

The artifact includes versioned datasets, development/test hashes, job and response ledgers, prompts, exact gold sets, the Z3 audit, catalog/price snapshots, cost guards, strict and sensitivity parsers, base- and family-clustered statistics, source data for every table and figure, editable draw.io schematics, and deterministic builds. v1, v2, and both cardinality extensions use distinct request namespaces. Every response carries its requested, canonical, and actual model slug, response identifier, usage, and cost; backend provider identifiers were not returned by the serving stack and are recorded as unavailable rather than inferred. Total execution cost is USD 91.3338.

AI Use Statement

Generative AI assisted literature discovery, code and test drafting, language editing, visualization prototyping, document generation, and three disclosed simulated reviews. Final method diagrams were recreated as editable vectors; model outputs and gold labels were not altered. Primary sources, numerical claims, code execution, and interpretations require human verification. The simulated reviews do not impersonate real reviewers or predict an ICLR decision.

References

Zheng Chen, Chuan Zhou, Fengxiang Cheng, Tin Po Yip, Fenrong Liu, Yisen Wang, Jiajun Chai, Xiaohan Wang, Guojun Yin, Wei Lin, Bo Li, Haoxuan Li, and Zhouchen Lin. Logiconbench: Benchmarking logical consistencies of llms. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/01db36a646c07c64dd39a92b4eceb417-Abstract-Conference.html.

Peter Clark, Oyvind Tafjord, and Kyle Richardson. Transformers as soft reasoners over language. *arXiv preprint arXiv:2002.05867*, 2020. URL https://arxiv.org/abs/2002.05867.

Bishwamittra Ghosh, Sarah Hasan, Naheed Anjum Arafat, and Arijit Khan. Logical consistency of large language models in fact-checking. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/3209cf3312b2cbb68e33644362ddc2bd-Abstract-Conference.html.

Alexander Gu and Alan Chen. Controlled reformulation testing for logical consistency in large language models. *arXiv preprint arXiv:2607.14528*, 2026. URL https://arxiv.org/abs/2607.14528.

Joonghyuk Hahn and Yo-Sub Han. Query4Regex: Verifiable regex transformation through formal operations from NL and DSL queries. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 6297–6305, 2026. doi: 10.18653/v1/2026.findings-eacl.331. URL https://aclanthology.org/2026.findings-eacl.331/.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Wenfei Zhou, James Coady, David Peng, Yujie Qiao, Luke Benson, Lucy Sun, Alex Wardle-Solano, Hannah Szabo, Ekaterina Zubova, Matthew Burtell, Jonathan Fan, Yixin Liu, Brian Wong, Malcolm Sailor, Ansong Ni, Linyong Nan, Jungo Kasai, Tao Yu, Rui Zhang, Alexander R. Fabbri, Wojciech Kryscinski, Semih Yavuz, Ye Liu, Xi Victoria Lin, Shafiq Joty, Yingbo Zhou, Caiming Xiong, Rex Ying, Arman Cohan, and Dragomir Radev. Folio: Natural language reasoning with first-order logic. *arXiv preprint arXiv:2209.00840*, 2022. URL https://arxiv.org/abs/2209.00840.

Zhaochen Hong, Haofei Yu, and Jiaxuan You. ConsistencyChecker: Tree-based evaluation of LLM generalization capabilities. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 33039–33075, 2025. doi: 10.18653/v1/2025.acl-long.1585. URL https://aclanthology.org/2025.acl-long.1585/.

Myeongjun Jang, Deuk Sin Kwon, and Thomas Lukasiewicz. BECEL: Benchmark for consistency evaluation of language models. In *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 3680–3696, 2022. URL https://aclanthology.org/2022.coling-1.324/.

Kostis Michailidis, Dimos Tsouros, and Tias Guns. CP-Bench: Evaluating large language models for constraint modelling. In *European Conference on Artificial Intelligence*, pp. 861–868, 2025. doi: 10.3233/FAIA250890. URL https://arxiv.org/abs/2506.06052v2.

Theo X. Olausson, Alex Gu, Benjamin Lipkin, Cedegao E. Zhang, Armando Solar-Lezama, Joshua B. Tenenbaum, and Roger Levy. Linc: A neurosymbolic approach for logical reasoning by combining language models with first-order logic provers. *arXiv preprint arXiv:2310.15164*, 2023. URL https://arxiv.org/abs/2310.15164.

Liangming Pan, Alon Albalak, Xinyi Wang, and William Yang Wang. Logic-LM: Empowering large language models with symbolic solvers for faithful logical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023. doi: 10.48550/arXiv.2305.12295. URL https://arxiv.org/abs/2305.12295.

Mihir Parmar, Nisarg Patel, Neeraj Varshney, Mutsumi Nakamura, Man Luo, Santosh Mashetty, Arindam Mitra, and Chitta Baral. Logicbench: Towards systematic evaluation of logical reasoning ability of large language models. *arXiv preprint arXiv:2404.15522*, 2024. URL https://arxiv.org/abs/2404.15522.

Gaurav Srivastava, Aafiya Hussain, Zhenyu Bi, Swastik Roy, Priya Pitre, Meng Lu, Morteza Ziyadi, and Xuan Wang. BeyondBench: Contamination-resistant evaluation of reasoning in language models. In *International Conference on Learning Representations*, pp. 20501–20636, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/22b4e30a7a660b17f1fb58ab49671e77-Abstract-Conference.html.

Oyvind Tafjord, Bhavana Dalvi Mishra, and Peter Clark. Proofwriter: Generating implications, proofs, and abductive statements over natural language. *arXiv preprint arXiv:2012.13048*, 2020. URL https://arxiv.org/abs/2012.13048.

Justin Waugh. Pencil puzzle bench: A benchmark for multi-step verifiable reasoning. *arXiv preprint arXiv:2603.02119*, 2026. doi: 10.48550/arXiv.2603.02119. URL https://arxiv.org/abs/2603.02119.

Anjiang Wei, Yuheng Wu, Yingjia Wan, Tarun Suresh, Huanmi Tan, Zhanke Zhou, Sanmi Koyejo, Ke Wang, and Alex Aiken. SATBench: Benchmarking LLMs' logical reasoning via automated puzzle generation from SAT formulas. *arXiv preprint arXiv:2505.14615*, 2025. doi: 10.48550/arXiv.2505.14615. URL https://arxiv.org/abs/2505.14615.

Zhe Yang, Yichang Zhang, Tianyu Liu, Jian Yang, Junyang Lin, Chang Zhou, and Zhifang Sui. Can large language models always solve easy problems if they can solve harder ones? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1531–1555, 2024. doi: 10.18653/v1/2024.emnlp-main.92. URL https://aclanthology.org/2024.emnlp-main.92/.

Xi Ye, Qiaochu Chen, Isil Dillig, and Greg Durrett. SatLM: Satisfiability-aided language models using declarative prompting. In *Advances in Neural Information Processing Systems*, 2023. doi: 10.48550/arXiv.2305.09656. URL https://arxiv.org/abs/2305.09656.

Appendix notation. D, E, and R denote decision, enumeration, and candidate-conditioned re-enumeration; 0S, 3S, and 3S+C denote zero-shot, three-shot, and three-shot plus certificate. A, P, R, F1, Fmt, and Flag denote accuracy, precision, recall, F1, format validity, and the share returning `error_found=true`; CI is a base-cluster interval, pH is a Holm-adjusted p value, nFam is the number of complete formula families, and NA is not applicable or not defined. B, O, N, Q, and X denote base, reorder-only, rename-only, equivalent-expression-only, and distractor-only views. SC, SW, FI, and NA in Table 18 denote strict-correct, strict-wrong with valid format, format-invalid, and completion-censored outcomes.

A Related Work and Operational Comparison

The compact comparison uses claim-specific dimensions rather than binary novelty flags. BeyondBench is the decisive nearest neighbor because it already provides algorithmic generation, deterministic verification, fully enumerated solutions, isomorphic transformations, large task breadth, and repeated evaluation. SATBench is closest on SAT-grounded generation; Query4Regex on formal equivalence; Ghosh et al. on meaning-preserving propositional queries. TrustCons does not claim priority for these ingredients. It differs by jointly treating literal-interface compliance, semantic exact-set recovery, owned single-factor views, candidate-conditioned re-enumeration, protocol censoring, and a matched visible-dimension solution-count intervention as separately auditable endpoints.

What a single-rendering evaluation would have missed. A resource that verifies an exact solution set on one rendering reports the base-view rate; TrustCons additionally requires the same formula to come back right after it is reordered, renamed, rewritten, and distracted. On the 11 bases where all four focal configurations supply a complete five-view response, DeepSeek V4 Pro High scores 100.0% on the base rendering but 81.8% on the five-view conjunction, while V4 Flash High scores 90.9% and 90.9%; the single-rendering endpoint orders the two the other way round. The magnitude is the stable part of that comparison and the reversal rests on 11 common bases, so it describes two configurations rather than a population and carries no test. This is the kind of statement that owned single-factor renderings of one formula produce and that an evaluation spread over many differently generated tasks cannot, and it is offered as such.

B Prompts, Parsing, and Censoring

The primary parser requires the whole response to be one JSON object, exact task keys, correct types, and unique ascending bitstrings. The frozen third-task prompt says "Diagnose and repair it" although its schema has no diagnosis field; unlike the enumeration prompt it also omits an explicit sorting instruction. These mismatches can depress strict-interface scores and reinforce why the endpoint is not interpreted as diagnosis or repair. The lenient extractor and conservative bare-answer recovery are labeled sensitivity endpoints. The prompts, the three disjoint development demonstrations, the deterministic repair corruption rule, and the parser contract are printed in full below. Tables 4–5 report third-task and cross-task summaries under the `repair` identifier defined in Section 2, including observed and censored counts.

The prose above describes the interface; this appendix supplies it. Every prompt below is regenerated from the frozen dataset by the same function the harness used, and each one is checked against the SHA-256 recorded in the request ledger before this document is built: all 7,344 registered prompts in `expanded-v2-tokenfix` reproduce byte-for-byte, so the strings printed here are the strings that were sent.

Worked example. All nine cells below are the same instance (`n7-002`, 7 variables, 1 satisfying assignment, difficulty `n7`), so the reader can diff exactly what each task instruction and each prompt mode adds and removes.

The prompt text is a deterministic function of four released objects: the task, the prompt mode, the instance's rendered rules, and the three development instances. The nine cells below therefore pin down the form of every prompt, and the frozen datasets pin down the content, so the printed subset

plus the released data determine the remaining registered prompts; the hash check at the end of this section confirms it for all of them rather than for the nine shown.

TASK INSTRUCTION: DECISION

```
Return JSON only:  {"consistent":  true} or {"consistent":
false}.
```

Zero-shot. sha256=a20ab8da3d554bb7... (hash matches ledger).

```
Solve the Boolean constraint system.  Every rule must hold;
variables are Boolean.

Target variable order:  A, B, C, D, E, F, G.

Only target variables belong in the answer; X and Z are
irrelevant when present.

R1.  At least one of the following is true:  C OR NOT G.
R2.  At least one of the following is true:  NOT B OR D OR
G.
R3.  At least one of the following is true:  B OR NOT D.
R4.  At least one of the following is true:  C OR F.
R5.  At least one of the following is true:  B OR C.
R6.  At least one of the following is true:  A OR B OR F.
R7.  At least one of the following is true:  NOT A OR G.
R8.  At least one of the following is true:  C OR D.
R9.  At least one of the following is true:  NOT D OR E.
R10.  At least one of the following is true:  NOT B OR NOT C
OR D.
R11.  At least one of the following is true:  E OR G.
R12.  At least one of the following is true:  NOT A OR NOT
E.
R13.  At least one of the following is true:  NOT B OR NOT
F.
R14.  At least one of the following is true:  NOT B OR C OR
D.
R15.  At least one of the following is true:  B OR NOT E.
R16.  At least one of the following is true:  E OR NOT F.
R17.  At least one of the following is true:  NOT E OR NOT
G.
R18.  At least one of the following is true:  NOT C OR NOT
F.
R19.  At least one of the following is true:  A OR NOT C.
R20.  At least one of the following is true:  C OR D OR F.

Return JSON only:  {"consistent":  true} or {"consistent":
false}.
```

Three-shot. sha256=df89923fbddc6318... (hash matches ledger).

```
Solve the Boolean constraint system.  Every rule must hold;
variables are Boolean.

Target variable order:  A, B, C, D, E, F, G.

Only target variables belong in the answer; X and Z are
irrelevant when present.

Three independent development examples follow:

Variables:  A, B, C, D
R1.  At least one of the following is true:  A OR NOT D.
R2.  At least one of the following is true:  NOT C OR D.
```

```
R3.  At least one of the following is true:  NOT A OR NOT B
OR NOT D.
R4.  At least one of the following is true:  NOT A OR C.
R5.  At least one of the following is true:  C OR NOT D.
R6.  At least one of the following is true:  NOT A OR B OR
C.
R7.  At least one of the following is true:  B OR NOT C OR
D.
R8.  At least one of the following is true:  A OR B.
R9.  At least one of the following is true:  A OR C.
R10.  At least one of the following is true:  NOT A OR NOT
B.
R11.  At least one of the following is true:  NOT B OR NOT
D.
R12.  At least one of the following is true:  NOT A OR NOT C
OR D.
Answer:  {"consistent":true}

--

Variables:  A, B, C, D
R1.  At least one of the following is true:  B OR C.
R2.  At least one of the following is true:  C OR NOT D.
R3.  At least one of the following is true:  A OR D.
R4.  At least one of the following is true:  NOT A OR B OR
D.
R5.  At least one of the following is true:  A OR B.
R6.  At least one of the following is true:  NOT B OR C.
R7.  At least one of the following is true:  A OR NOT D.
R8.  At least one of the following is true:  A OR B OR NOT
C.
R9.  At least one of the following is true:  NOT B OR NOT D.
R10.  At least one of the following is true:  NOT B OR NOT
C.
R11.  At least one of the following is true:  B OR D.
R12.  At least one of the following is true:  NOT A OR B OR
NOT C.
Answer:  {"consistent":false}

--

Variables:  A, B, C, D
R1.  At least one of the following is true:  NOT B OR D.
R2.  At least one of the following is true:  NOT A OR NOT D.
R3.  At least one of the following is true:  NOT C OR D.
R4.  At least one of the following is true:  A OR C.
R5.  At least one of the following is true:  A OR NOT B OR
D.
R6.  At least one of the following is true:  NOT A OR NOT B
OR NOT D.
R7.  At least one of the following is true:  NOT A OR C OR
NOT D.
R8.  At least one of the following is true:  C OR D.
R9.  At least one of the following is true:  NOT A OR NOT C
OR D.
R10.  At least one of the following is true:  NOT A OR D.
R11.  At least one of the following is true:  NOT A OR B OR
C.
Answer:  {"consistent":true}

R1.  At least one of the following is true:  C OR NOT G.
R2.  At least one of the following is true:  NOT B OR D OR
G.
R3.  At least one of the following is true:  B OR NOT D.
R4.  At least one of the following is true:  C OR F.
```

```
R5.  At least one of the following is true:  B OR C.
R6.  At least one of the following is true:  A OR B OR F.
R7.  At least one of the following is true:  NOT A OR G.
R8.  At least one of the following is true:  C OR D.
R9.  At least one of the following is true:  NOT D OR E.
R10.  At least one of the following is true:  NOT B OR NOT C
OR D.
R11.  At least one of the following is true:  E OR G.
R12.  At least one of the following is true:  NOT A OR NOT
E.
R13.  At least one of the following is true:  NOT B OR NOT
F.
R14.  At least one of the following is true:  NOT B OR C OR
D.
R15.  At least one of the following is true:  B OR NOT E.
R16.  At least one of the following is true:  E OR NOT F.
R17.  At least one of the following is true:  NOT E OR NOT
G.
R18.  At least one of the following is true:  NOT C OR NOT
F.
R19.  At least one of the following is true:  A OR NOT C.
R20.  At least one of the following is true:  C OR D OR F.

Return JSON only:  {"consistent":  true} or {"consistent":
false}.
```

Three-shot with verifier certificate. `sha256=17a5bd13abd9cad3...` (hash matches ledger).

```
Solve the Boolean constraint system.  Every rule must hold;
variables are Boolean.

Target variable order:  A, B, C, D, E, F, G.

Only target variables belong in the answer; X and Z are
irrelevant when present.

Three independent development examples follow:

Variables:  A, B, C, D
R1.  At least one of the following is true:  A OR NOT D.
R2.  At least one of the following is true:  NOT C OR D.
R3.  At least one of the following is true:  NOT A OR NOT B
OR NOT D.
R4.  At least one of the following is true:  NOT A OR C.
R5.  At least one of the following is true:  C OR NOT D.
R6.  At least one of the following is true:  NOT A OR B OR
C.
R7.  At least one of the following is true:  B OR NOT C OR
D.
R8.  At least one of the following is true:  A OR B.
R9.  At least one of the following is true:  A OR C.
R10.  At least one of the following is true:  NOT A OR NOT
B.
R11.  At least one of the following is true:  NOT B OR NOT
D.
R12.  At least one of the following is true:  NOT A OR NOT C
OR D.
Answer:  {"consistent":true}
Verifier certificate:  checked 16 assignments; exact
satisfying count=1.

--
```

```
Variables:  A, B, C, D
R1.  At least one of the following is true:  B OR C.
R2.  At least one of the following is true:  C OR NOT D.
R3.  At least one of the following is true:  A OR D.
R4.  At least one of the following is true:  NOT A OR B OR
D.
R5.  At least one of the following is true:  A OR B.
R6.  At least one of the following is true:  NOT B OR C.
R7.  At least one of the following is true:  A OR NOT D.
R8.  At least one of the following is true:  A OR B OR NOT
C.
R9.  At least one of the following is true:  NOT B OR NOT D.
R10.  At least one of the following is true:  NOT B OR NOT
C.
R11.  At least one of the following is true:  B OR D.
R12.  At least one of the following is true:  NOT A OR B OR
NOT C.
Answer:  {"consistent":false}
Verifier certificate:  checked 16 assignments; exact
satisfying count=0.

--

Variables:  A, B, C, D
R1.  At least one of the following is true:  NOT B OR D.
R2.  At least one of the following is true:  NOT A OR NOT D.
R3.  At least one of the following is true:  NOT C OR D.
R4.  At least one of the following is true:  A OR C.
R5.  At least one of the following is true:  A OR NOT B OR
D.
R6.  At least one of the following is true:  NOT A OR NOT B
OR NOT D.
R7.  At least one of the following is true:  NOT A OR C OR
NOT D.
R8.  At least one of the following is true:  C OR D.
R9.  At least one of the following is true:  NOT A OR NOT C
OR D.
R10.  At least one of the following is true:  NOT A OR D.
R11.  At least one of the following is true:  NOT A OR B OR
C.
Answer:  {"consistent":true}
Verifier certificate:  checked 16 assignments; exact
satisfying count=2.

R1.  At least one of the following is true:  C OR NOT G.
R2.  At least one of the following is true:  NOT B OR D OR
G.
R3.  At least one of the following is true:  B OR NOT D.
R4.  At least one of the following is true:  C OR F.
R5.  At least one of the following is true:  B OR C.
R6.  At least one of the following is true:  A OR B OR F.
R7.  At least one of the following is true:  NOT A OR G.
R8.  At least one of the following is true:  C OR D.
R9.  At least one of the following is true:  NOT D OR E.
R10.  At least one of the following is true:  NOT B OR NOT C
OR D.
R11.  At least one of the following is true:  E OR G.
R12.  At least one of the following is true:  NOT A OR NOT
E.
R13.  At least one of the following is true:  NOT B OR NOT
F.
R14.  At least one of the following is true:  NOT B OR C OR
D.
R15.  At least one of the following is true:  B OR NOT E.
R16.  At least one of the following is true:  E OR NOT F.
```

```
R17.  At least one of the following is true:  NOT E OR NOT
G.
R18.  At least one of the following is true:  NOT C OR NOT
F.
R19.  At least one of the following is true:  A OR NOT C.
R20.  At least one of the following is true:  C OR D OR F.

Return JSON only:  {"consistent":  true} or {"consistent":
false}.
```

TASK INSTRUCTION: ENUMERATION

```
Return JSON only as {"solutions":  [...]}.  Enumerate every
satisfying bitstring in ascending order; use an empty list
when none exists.
```

Zero-shot. `sha256=abc35e689aec5cdf...` (hash matches ledger).

```
Solve the Boolean constraint system.  Every rule must hold;
variables are Boolean.

Target variable order:  A, B, C, D, E, F, G.

Only target variables belong in the answer; X and Z are
irrelevant when present.

R1.  At least one of the following is true:  C OR NOT G.
R2.  At least one of the following is true:  NOT B OR D OR
G.
R3.  At least one of the following is true:  B OR NOT D.
R4.  At least one of the following is true:  C OR F.
R5.  At least one of the following is true:  B OR C.
R6.  At least one of the following is true:  A OR B OR F.
R7.  At least one of the following is true:  NOT A OR G.
R8.  At least one of the following is true:  C OR D.
R9.  At least one of the following is true:  NOT D OR E.
R10.  At least one of the following is true:  NOT B OR NOT C
OR D.
R11.  At least one of the following is true:  E OR G.
R12.  At least one of the following is true:  NOT A OR NOT
E.
R13.  At least one of the following is true:  NOT B OR NOT
F.
R14.  At least one of the following is true:  NOT B OR C OR
D.
R15.  At least one of the following is true:  B OR NOT E.
R16.  At least one of the following is true:  E OR NOT F.
R17.  At least one of the following is true:  NOT E OR NOT
G.
R18.  At least one of the following is true:  NOT C OR NOT
F.
R19.  At least one of the following is true:  A OR NOT C.
R20.  At least one of the following is true:  C OR D OR F.

Return JSON only as {"solutions":  [...]}.  Enumerate every
satisfying bitstring in ascending order; use an empty list
when none exists.
```

Three-shot. `sha256=fc1be263cc5f1188...` (hash matches ledger).

```
Solve the Boolean constraint system.  Every rule must hold;
variables are Boolean.

Target variable order:  A, B, C, D, E, F, G.
```

```
Only target variables belong in the answer; X and Z are
irrelevant when present.

Three independent development examples follow:

Variables:  A, B, C, D
R1.  At least one of the following is true:  A OR NOT D.
R2.  At least one of the following is true:  NOT C OR D.
R3.  At least one of the following is true:  NOT A OR NOT B
OR NOT D.
R4.  At least one of the following is true:  NOT A OR C.
R5.  At least one of the following is true:  C OR NOT D.
R6.  At least one of the following is true:  NOT A OR B OR
C.
R7.  At least one of the following is true:  B OR NOT C OR
D.
R8.  At least one of the following is true:  A OR B.
R9.  At least one of the following is true:  A OR C.
R10.  At least one of the following is true:  NOT A OR NOT
B.
R11.  At least one of the following is true:  NOT B OR NOT
D.
R12.  At least one of the following is true:  NOT A OR NOT C
OR D.
Answer:  {"solutions":["1011"]}

--

Variables:  A, B, C, D
R1.  At least one of the following is true:  B OR C.
R2.  At least one of the following is true:  C OR NOT D.
R3.  At least one of the following is true:  A OR D.
R4.  At least one of the following is true:  NOT A OR B OR
D.
R5.  At least one of the following is true:  A OR B.
R6.  At least one of the following is true:  NOT B OR C.
R7.  At least one of the following is true:  A OR NOT D.
R8.  At least one of the following is true:  A OR B OR NOT
C.
R9.  At least one of the following is true:  NOT B OR NOT D.
R10.  At least one of the following is true:  NOT B OR NOT
C.
R11.  At least one of the following is true:  B OR D.
R12.  At least one of the following is true:  NOT A OR B OR
NOT C.
Answer:  {"solutions":[]}

--

Variables:  A, B, C, D
R1.  At least one of the following is true:  NOT B OR D.
R2.  At least one of the following is true:  NOT A OR NOT D.
R3.  At least one of the following is true:  NOT C OR D.
R4.  At least one of the following is true:  A OR C.
R5.  At least one of the following is true:  A OR NOT B OR
D.
R6.  At least one of the following is true:  NOT A OR NOT B
OR NOT D.
R7.  At least one of the following is true:  NOT A OR C OR
NOT D.
R8.  At least one of the following is true:  C OR D.
R9.  At least one of the following is true:  NOT A OR NOT C
OR D.
R10.  At least one of the following is true:  NOT A OR D.
```

```
R11.  At least one of the following is true:  NOT A OR B OR
C.
Answer:  {"solutions":["0011","0111"]}

R1.  At least one of the following is true:  C OR NOT G.
R2.  At least one of the following is true:  NOT B OR D OR
G.
R3.  At least one of the following is true:  B OR NOT D.
R4.  At least one of the following is true:  C OR F.
R5.  At least one of the following is true:  B OR C.
R6.  At least one of the following is true:  A OR B OR F.
R7.  At least one of the following is true:  NOT A OR G.
R8.  At least one of the following is true:  C OR D.
R9.  At least one of the following is true:  NOT D OR E.
R10.  At least one of the following is true:  NOT B OR NOT C
OR D.
R11.  At least one of the following is true:  E OR G.
R12.  At least one of the following is true:  NOT A OR NOT
E.
R13.  At least one of the following is true:  NOT B OR NOT
F.
R14.  At least one of the following is true:  NOT B OR C OR
D.
R15.  At least one of the following is true:  B OR NOT E.
R16.  At least one of the following is true:  E OR NOT F.
R17.  At least one of the following is true:  NOT E OR NOT
G.
R18.  At least one of the following is true:  NOT C OR NOT
F.
R19.  At least one of the following is true:  A OR NOT C.
R20.  At least one of the following is true:  C OR D OR F.

Return JSON only as {"solutions":  [...]}.  Enumerate every
satisfying bitstring in ascending order; use an empty list
when none exists.
```

Three-shot with verifier certificate. `sha256=0b49c3c0e894bd51...` (hash matches ledger).

```
Solve the Boolean constraint system.  Every rule must hold;
variables are Boolean.

Target variable order:  A, B, C, D, E, F, G.

Only target variables belong in the answer; X and Z are
irrelevant when present.

Three independent development examples follow:

Variables:  A, B, C, D
R1.  At least one of the following is true:  A OR NOT D.
R2.  At least one of the following is true:  NOT C OR D.
R3.  At least one of the following is true:  NOT A OR NOT B
OR NOT D.
R4.  At least one of the following is true:  NOT A OR C.
R5.  At least one of the following is true:  C OR NOT D.
R6.  At least one of the following is true:  NOT A OR B OR
C.
R7.  At least one of the following is true:  B OR NOT C OR
D.
R8.  At least one of the following is true:  A OR B.
R9.  At least one of the following is true:  A OR C.
R10.  At least one of the following is true:  NOT A OR NOT
B.
```

```
R11.  At least one of the following is true:  NOT B OR NOT
D.
R12.  At least one of the following is true:  NOT A OR NOT C
OR D.
Answer:  {"solutions":["1011"]}
Verifier certificate:  checked 16 assignments; exact
satisfying count=1.

--

Variables:  A, B, C, D
R1.  At least one of the following is true:  B OR C.
R2.  At least one of the following is true:  C OR NOT D.
R3.  At least one of the following is true:  A OR D.
R4.  At least one of the following is true:  NOT A OR B OR
D.
R5.  At least one of the following is true:  A OR B.
R6.  At least one of the following is true:  NOT B OR C.
R7.  At least one of the following is true:  A OR NOT D.
R8.  At least one of the following is true:  A OR B OR NOT
C.
R9.  At least one of the following is true:  NOT B OR NOT D.
R10.  At least one of the following is true:  NOT B OR NOT
C.
R11.  At least one of the following is true:  B OR D.
R12.  At least one of the following is true:  NOT A OR B OR
NOT C.
Answer:  {"solutions":[]}
Verifier certificate:  checked 16 assignments; exact
satisfying count=0.

--

Variables:  A, B, C, D
R1.  At least one of the following is true:  NOT B OR D.
R2.  At least one of the following is true:  NOT A OR NOT D.
R3.  At least one of the following is true:  NOT C OR D.
R4.  At least one of the following is true:  A OR C.
R5.  At least one of the following is true:  A OR NOT B OR
D.
R6.  At least one of the following is true:  NOT A OR NOT B
OR NOT D.
R7.  At least one of the following is true:  NOT A OR C OR
NOT D.
R8.  At least one of the following is true:  C OR D.
R9.  At least one of the following is true:  NOT A OR NOT C
OR D.
R10.  At least one of the following is true:  NOT A OR D.
R11.  At least one of the following is true:  NOT A OR B OR
C.
Answer:  {"solutions":["0011","0111"]}
Verifier certificate:  checked 16 assignments; exact
satisfying count=2.

R1.  At least one of the following is true:  C OR NOT G.
R2.  At least one of the following is true:  NOT B OR D OR
G.
R3.  At least one of the following is true:  B OR NOT D.
R4.  At least one of the following is true:  C OR F.
R5.  At least one of the following is true:  B OR C.
R6.  At least one of the following is true:  A OR B OR F.
R7.  At least one of the following is true:  NOT A OR G.
R8.  At least one of the following is true:  C OR D.
R9.  At least one of the following is true:  NOT D OR E.
```

```
R10.  At least one of the following is true:  NOT B OR NOT C
OR D.
R11.  At least one of the following is true:  E OR G.
R12.  At least one of the following is true:  NOT A OR NOT
E.
R13.  At least one of the following is true:  NOT B OR NOT
F.
R14.  At least one of the following is true:  NOT B OR C OR
D.
R15.  At least one of the following is true:  B OR NOT E.
R16.  At least one of the following is true:  E OR NOT F.
R17.  At least one of the following is true:  NOT E OR NOT
G.
R18.  At least one of the following is true:  NOT C OR NOT
F.
R19.  At least one of the following is true:  A OR NOT C.
R20.  At least one of the following is true:  C OR D OR F.

Return JSON only as {"solutions":  [...]}.  Enumerate every
satisfying bitstring in ascending order; use an empty list
when none exists.
```

TASK INSTRUCTION: REPAIR

```
The candidate solution set below is known to contain exactly
one deterministic error.  Candidate:  ["0000000"].  Diagnose
and repair it.  Return JSON only as {"error_found":  true,
"solutions":  [...]}.
```

Zero-shot. `sha256=31f000726596615d...` (hash matches ledger).

```
Solve the Boolean constraint system.  Every rule must hold;
variables are Boolean.

Target variable order:  A, B, C, D, E, F, G.

Only target variables belong in the answer; X and Z are
irrelevant when present.

R1.  At least one of the following is true:  C OR NOT G.
R2.  At least one of the following is true:  NOT B OR D OR
G.
R3.  At least one of the following is true:  B OR NOT D.
R4.  At least one of the following is true:  C OR F.
R5.  At least one of the following is true:  B OR C.
R6.  At least one of the following is true:  A OR B OR F.
R7.  At least one of the following is true:  NOT A OR G.
R8.  At least one of the following is true:  C OR D.
R9.  At least one of the following is true:  NOT D OR E.
R10.  At least one of the following is true:  NOT B OR NOT C
OR D.
R11.  At least one of the following is true:  E OR G.
R12.  At least one of the following is true:  NOT A OR NOT
E.
R13.  At least one of the following is true:  NOT B OR NOT
F.
R14.  At least one of the following is true:  NOT B OR C OR
D.
R15.  At least one of the following is true:  B OR NOT E.
R16.  At least one of the following is true:  E OR NOT F.
R17.  At least one of the following is true:  NOT E OR NOT
G.
R18.  At least one of the following is true:  NOT C OR NOT
F.
```

```
R19.  At least one of the following is true:  A OR NOT C.
R20.  At least one of the following is true:  C OR D OR F.

The candidate solution set below is known to contain exactly
one deterministic error.  Candidate:  ["0000000"].  Diagnose
and repair it.  Return JSON only as {"error_found":  true,
"solutions":  [...]}.
```

Three-shot. `sha256=3c1c142f86293feb...` (hash matches ledger).

```
Solve the Boolean constraint system.  Every rule must hold;
variables are Boolean.

Target variable order:  A, B, C, D, E, F, G.

Only target variables belong in the answer; X and Z are
irrelevant when present.

Three independent development examples follow:

Variables:  A, B, C, D
R1.  At least one of the following is true:  A OR NOT D.
R2.  At least one of the following is true:  NOT C OR D.
R3.  At least one of the following is true:  NOT A OR NOT B
OR NOT D.
R4.  At least one of the following is true:  NOT A OR C.
R5.  At least one of the following is true:  C OR NOT D.
R6.  At least one of the following is true:  NOT A OR B OR
C.
R7.  At least one of the following is true:  B OR NOT C OR
D.
R8.  At least one of the following is true:  A OR B.
R9.  At least one of the following is true:  A OR C.
R10.  At least one of the following is true:  NOT A OR NOT
B.
R11.  At least one of the following is true:  NOT B OR NOT
D.
R12.  At least one of the following is true:  NOT A OR NOT C
OR D.
Candidate:  ["0000"]
Answer:  {"error_found":true,"solutions":["1011"]}

--

Variables:  A, B, C, D
R1.  At least one of the following is true:  B OR C.
R2.  At least one of the following is true:  C OR NOT D.
R3.  At least one of the following is true:  A OR D.
R4.  At least one of the following is true:  NOT A OR B OR
D.
R5.  At least one of the following is true:  A OR B.
R6.  At least one of the following is true:  NOT B OR C.
R7.  At least one of the following is true:  A OR NOT D.
R8.  At least one of the following is true:  A OR B OR NOT
C.
R9.  At least one of the following is true:  NOT B OR NOT D.
R10.  At least one of the following is true:  NOT B OR NOT
C.
R11.  At least one of the following is true:  B OR D.
R12.  At least one of the following is true:  NOT A OR B OR
NOT C.
Candidate:  ["0000"]
Answer:  {"error_found":true,"solutions":[]}
```

```
--

Variables:  A, B, C, D
R1.  At least one of the following is true:  NOT B OR D.
R2.  At least one of the following is true:  NOT A OR NOT D.
R3.  At least one of the following is true:  NOT C OR D.
R4.  At least one of the following is true:  A OR C.
R5.  At least one of the following is true:  A OR NOT B OR
D.
R6.  At least one of the following is true:  NOT A OR NOT B
OR NOT D.
R7.  At least one of the following is true:  NOT A OR C OR
NOT D.
R8.  At least one of the following is true:  C OR D.
R9.  At least one of the following is true:  NOT A OR NOT C
OR D.
R10.  At least one of the following is true:  NOT A OR D.
R11.  At least one of the following is true:  NOT A OR B OR
C.
Candidate:  ["0111"]
Answer:  {"error_found":true,"solutions":["0011","0111"]}

R1.  At least one of the following is true:  C OR NOT G.
R2.  At least one of the following is true:  NOT B OR D OR
G.
R3.  At least one of the following is true:  B OR NOT D.
R4.  At least one of the following is true:  C OR F.
R5.  At least one of the following is true:  B OR C.
R6.  At least one of the following is true:  A OR B OR F.
R7.  At least one of the following is true:  NOT A OR G.
R8.  At least one of the following is true:  C OR D.
R9.  At least one of the following is true:  NOT D OR E.
R10.  At least one of the following is true:  NOT B OR NOT C
OR D.
R11.  At least one of the following is true:  E OR G.
R12.  At least one of the following is true:  NOT A OR NOT
E.
R13.  At least one of the following is true:  NOT B OR NOT
F.
R14.  At least one of the following is true:  NOT B OR C OR
D.
R15.  At least one of the following is true:  B OR NOT E.
R16.  At least one of the following is true:  E OR NOT F.
R17.  At least one of the following is true:  NOT E OR NOT
G.
R18.  At least one of the following is true:  NOT C OR NOT
F.
R19.  At least one of the following is true:  A OR NOT C.
R20.  At least one of the following is true:  C OR D OR F.

The candidate solution set below is known to contain exactly
one deterministic error.  Candidate:  ["0000000"].  Diagnose
and repair it.  Return JSON only as {"error_found":  true,
"solutions":  [...]}.
```

Three-shot with verifier certificate. `sha256=732af03deaa4342d...` (hash matches ledger).

```
Solve the Boolean constraint system.  Every rule must hold;
variables are Boolean.

Target variable order:  A, B, C, D, E, F, G.
```

```
Only target variables belong in the answer; X and Z are
irrelevant when present.

Three independent development examples follow:

Variables:  A, B, C, D
R1.  At least one of the following is true:  A OR NOT D.
R2.  At least one of the following is true:  NOT C OR D.
R3.  At least one of the following is true:  NOT A OR NOT B
OR NOT D.
R4.  At least one of the following is true:  NOT A OR C.
R5.  At least one of the following is true:  C OR NOT D.
R6.  At least one of the following is true:  NOT A OR B OR
C.
R7.  At least one of the following is true:  B OR NOT C OR
D.
R8.  At least one of the following is true:  A OR B.
R9.  At least one of the following is true:  A OR C.
R10.  At least one of the following is true:  NOT A OR NOT
B.
R11.  At least one of the following is true:  NOT B OR NOT
D.
R12.  At least one of the following is true:  NOT A OR NOT C
OR D.
Candidate:  ["0000"]
Answer:  {"error_found":true,"solutions":["1011"]}
Verifier certificate:  checked 16 assignments; exact
satisfying count=1.

--

Variables:  A, B, C, D
R1.  At least one of the following is true:  B OR C.
R2.  At least one of the following is true:  C OR NOT D.
R3.  At least one of the following is true:  A OR D.
R4.  At least one of the following is true:  NOT A OR B OR
D.
R5.  At least one of the following is true:  A OR B.
R6.  At least one of the following is true:  NOT B OR C.
R7.  At least one of the following is true:  A OR NOT D.
R8.  At least one of the following is true:  A OR B OR NOT
C.
R9.  At least one of the following is true:  NOT B OR NOT D.
R10.  At least one of the following is true:  NOT B OR NOT
C.
R11.  At least one of the following is true:  B OR D.
R12.  At least one of the following is true:  NOT A OR B OR
NOT C.
Candidate:  ["0000"]
Answer:  {"error_found":true,"solutions":[]}
Verifier certificate:  checked 16 assignments; exact
satisfying count=0.

--

Variables:  A, B, C, D
R1.  At least one of the following is true:  NOT B OR D.
R2.  At least one of the following is true:  NOT A OR NOT D.
R3.  At least one of the following is true:  NOT C OR D.
R4.  At least one of the following is true:  A OR C.
R5.  At least one of the following is true:  A OR NOT B OR
D.
R6.  At least one of the following is true:  NOT A OR NOT B
OR NOT D.
```

```
R7.  At least one of the following is true:  NOT A OR C OR
NOT D.
R8.  At least one of the following is true:  C OR D.
R9.  At least one of the following is true:  NOT A OR NOT C
OR D.
R10.  At least one of the following is true:  NOT A OR D.
R11.  At least one of the following is true:  NOT A OR B OR
C.
Candidate:  ["0111"]
Answer:  {"error_found":true,"solutions":["0011","0111"]}
Verifier certificate:  checked 16 assignments; exact
satisfying count=2.

R1.  At least one of the following is true:  C OR NOT G.
R2.  At least one of the following is true:  NOT B OR D OR
G.
R3.  At least one of the following is true:  B OR NOT D.
R4.  At least one of the following is true:  C OR F.
R5.  At least one of the following is true:  B OR C.
R6.  At least one of the following is true:  A OR B OR F.
R7.  At least one of the following is true:  NOT A OR G.
R8.  At least one of the following is true:  C OR D.
R9.  At least one of the following is true:  NOT D OR E.
R10.  At least one of the following is true:  NOT B OR NOT C
OR D.
R11.  At least one of the following is true:  E OR G.
R12.  At least one of the following is true:  NOT A OR NOT
E.
R13.  At least one of the following is true:  NOT B OR NOT
F.
R14.  At least one of the following is true:  NOT B OR C OR
D.
R15.  At least one of the following is true:  B OR NOT E.
R16.  At least one of the following is true:  E OR NOT F.
R17.  At least one of the following is true:  NOT E OR NOT
G.
R18.  At least one of the following is true:  NOT C OR NOT
F.
R19.  At least one of the following is true:  A OR NOT C.
R20.  At least one of the following is true:  C OR D OR F.

The candidate solution set below is known to contain exactly
one deterministic error.  Candidate:  ["0000000"].  Diagnose
and repair it.  Return JSON only as {"error_found":  true,
"solutions":  [...]}.
```

DEVELOPMENT DEMONSTRATIONS

The three-shot and certificate prompts prepend the same three disjoint development instances, drawn from a separate four-instance development set whose identifiers never intersect the 48 evaluated bases. The certificate suffix states only how many assignments were checked and how many satisfied the rules; it never reveals a test answer.

Demonstration 1 (decision task).

```
Variables:  A, B, C, D
R1.  At least one of the following is true:  A OR NOT D.
R2.  At least one of the following is true:  NOT C OR D.
R3.  At least one of the following is true:  NOT A OR NOT B
OR NOT D.
R4.  At least one of the following is true:  NOT A OR C.
R5.  At least one of the following is true:  C OR NOT D.
R6.  At least one of the following is true:  NOT A OR B OR
C.
```

```
R7.  At least one of the following is true:  B OR NOT C OR
D.
R8.  At least one of the following is true:  A OR B.
R9.  At least one of the following is true:  A OR C.
R10.  At least one of the following is true:  NOT A OR NOT
B.
R11.  At least one of the following is true:  NOT B OR NOT
D.
R12.  At least one of the following is true:  NOT A OR NOT C
OR D.
Answer:  {"consistent":true}
```

Demonstration 2 (decision task).

```
Variables:  A, B, C, D
R1.  At least one of the following is true:  B OR C.
R2.  At least one of the following is true:  C OR NOT D.
R3.  At least one of the following is true:  A OR D.
R4.  At least one of the following is true:  NOT A OR B OR
D.
R5.  At least one of the following is true:  A OR B.
R6.  At least one of the following is true:  NOT B OR C.
R7.  At least one of the following is true:  A OR NOT D.
R8.  At least one of the following is true:  A OR B OR NOT
C.
R9.  At least one of the following is true:  NOT B OR NOT D.
R10.  At least one of the following is true:  NOT B OR NOT
C.
R11.  At least one of the following is true:  B OR D.
R12.  At least one of the following is true:  NOT A OR B OR
NOT C.
Answer:  {"consistent":false}
```

Demonstration 3 (decision task).

```
Variables:  A, B, C, D
R1.  At least one of the following is true:  NOT B OR D.
R2.  At least one of the following is true:  NOT A OR NOT D.
R3.  At least one of the following is true:  NOT C OR D.
R4.  At least one of the following is true:  A OR C.
R5.  At least one of the following is true:  A OR NOT B OR
D.
R6.  At least one of the following is true:  NOT A OR NOT B
OR NOT D.
R7.  At least one of the following is true:  NOT A OR C OR
NOT D.
R8.  At least one of the following is true:  C OR D.
R9.  At least one of the following is true:  NOT A OR NOT C
OR D.
R10.  At least one of the following is true:  NOT A OR D.
R11.  At least one of the following is true:  NOT A OR B OR
C.
Answer:  {"consistent":true}
```

Certificate suffix. In `three_shot_certificate` each demonstration above gains one final line, and no other text changes:

```
Verifier certificate:  checked 16 assignments; exact
satisfying count=1.
Verifier certificate:  checked 16 assignments; exact
satisfying count=0.
Verifier certificate:  checked 16 assignments; exact
satisfying count=2.
```

DETERMINISTIC REPAIR CORRUPTION

Repair inputs are derived from the gold set by a rule that depends only on satisfiability and cardinality, never on model behaviour:

- no satisfying assignment: the candidate set is `["0000000"]` (`spurious_solution`);
- more than one satisfying assignment: the candidate set omits `corrupted[0]` (`missing_solution`);
- exactly one satisfying assignment: the candidate set replaces it with the lexicographically first unsatisfying assignment (`replaced_solution`).

The prompt states only that exactly one deterministic error is present. Because the corruption type is a function of the instance, corruption-type differences are descriptive and are not given a causal reading.

PARSER CONTRACT AND SENSITIVITY ENDPOINTS

```
PRIMARY (literal interface, reported as strict):
whole response must be one JSON object
exact task keys; boolean for decision; list of bitstrings
for enumeration
bitstrings unique, ascending, correct length, over target
variables only

SENSITIVITY 1 (lenient JSON extraction, historical
endpoint):
strip a surrounding code fence, then take the outermost
{...}  span

SENSITIVITY 2 (conservative bare-answer recovery):
accept only an unambiguous bare answer with no surrounding
prose
```

PROMPT REGENERATION CHECK

Task	Prompt mode	Registered calls	Hashes reproduced
decision	zero_shot	912	912
decision	three_shot	912	912
decision	three_shot_certificate	912	912
enumeration	zero_shot	912	912
enumeration	three_shot	912	912
enumeration	three_shot_certificate	912	912
repair	zero_shot	912	912
repair	three_shot	480	480
repair	three_shot_certificate	480	480
Total		7,344	7,344

Table 4: Task 3 candidate-conditioned re-enumeration by model and difficulty. The abbreviations are Flag=error flag returned true (the prompt's single-error guarantee makes this near-constant, so it is reported for completeness and is not evidence of detection), E=exact set, P=precision, R=recall, F1=F1, Fmt=contract-valid interface. Full machine-readable values are supplied as CSV.

system_id	n4 Flag/E/P/R/F/Fmt	n5 Flag/E/P/R/F/Fmt	n6 Flag/E/P/R/F/Fmt	n7 Flag/E/P/R/F/Fmt
claude-sonnet-5	1.00/.17/1.00/1.00/1.00/.17	1.00/.00/1.00/1.00/1.00/.00	1.00/.00/1.00/1.00/1.00/.00	.67/.00/.67/.67/.67/.00
deepseek-v4-flash-high	1.00/1.00/1.00/1.00/1.00/1.00	1.00/1.00/1.00/1.00/1.00/1.00	1.00/1.00/1.00/1.00/1.00/1.00	1.00/1.00/1.00/1.00/1.00/1.00
deepseek-v4-pro-high	1.00/1.00/1.00/1.00/1.00/1.00	NA/NA/NA/NA/NA/NA	1.00/1.00/1.00/1.00/1.00/1.00	NA/NA/NA/NA/NA/NA
gemini-3.1-pro-preview	.92/.08/.75/.75/.75/.17	.83/.00/.58/.58/.58/.00	.58/.00/.50/.50/.50/.00	.92/.08/.67/.67/.67/.08
gemini-3.8-flash	.92/.00/.92/.92/.92/.00	.92/.17/.92/.92/.92/.17	.83/.08/.75/.75/.75/.08	.64/.18/.64/.64/.64/.18
gpt-5.6-luna	1.00/.67/.71/.75/.72/1.00	1.00/.91/.95/1.00/.97/1.00	1.00/.67/.71/.75/.72/.92	1.00/1.00/1.00/1.00/1.00/1.00
grok-4.6	1.00/1.00/1.00/1.00/1.00/1.00	1.00/1.00/1.00/1.00/1.00/1.00	1.00/1.00/1.00/1.00/1.00/1.00	1.00/1.00/1.00/1.00/1.00/1.00
mistral-large	.58/.00/.00/.00/.00/.00	.17/.00/.08/.03/.04/.00	.33/.00/.08/.06/.07/.00	.92/.00/.00/.00/.00/.00
muse-spark-1.3	1.00/1.00/1.00/1.00/1.00/1.00	1.00/1.00/1.00/1.00/1.00/1.00	1.00/1.00/1.00/1.00/1.00/1.00	1.00/1.00/1.00/1.00/1.00/1.00
phi-4	1.00/.00/.01/.17/.03/.00	1.00/.00/.14/.28/.14/.00	1.00/.00/.28/.15/.19/.00	1.00/.00/.22/.22/.22/.00
qwen3.8-a95b	1.00/.86/.86/.86/.86/1.00	1.00/1.00/1.00/1.00/1.00/1.00	1.00/.50/.50/.50/.50/1.00	1.00/1.00/1.00/1.00/1.00/1.00
qwen3.8-max	1.00/1.00/1.00/1.00/1.00/1.00	1.00/1.00/1.00/1.00/1.00/1.00	1.00/1.00/1.00/1.00/1.00/1.00	NA/NA/NA/NA/NA/NA

Table 5: Cross-task and stability associations. Task labels are D=decision, E=enumeration, and R=candidate-conditioned re-enumeration; prompt labels are 0S=zero-shot, 3S=three-shot, and 3S+C=three-shot plus certificate. Each cell is Exact/Format/CI-low/CI-high. Full machine-readable values are supplied as CSV.

system_id	D-0S E/F/CIlo/ CIhi	D-3S E/F/CIlo/ CIhi	D-3S+C E/F/CIlo/ CIhi	E-0S E/F/CIlo/ CIhi	E-3S E/F/CIlo/ CIhi	E-3S+C E/F/CIlo/ CIhi	R-0S E/F/CIlo/ CIhi	R-3S E/F/CIlo/ CIhi	R-3S+C E/F/CIlo/ CIhi
claude-sonnet-5	.67/.67/.43/.86	.90/.90/.76/1.00	.64/.64/.36/.86	.02/.02/.00/.07	.19/.19/.08/.33	.09/.09/.00/.20	.06/.06/.00/.17	--	--
deepseek-v4-flash-high	1.00/1.00/1.00/ 1.00	.96/1.00/.88/1.00	1.00/1.00/1.00/ 1.00	.92/.95/.84/1.00	.97/1.00/.92/1.00	1.00/1.00/1.00/ 1.00	1.00/1.00/1.00/ 1.00	.76/1.00/.53/.94	.92/1.00/.79/1.00
deepseek-v4-pro-high	1.00/1.00/1.00/ 1.00	1.00/1.00/1.00/ 1.00	.91/.91/.77/1.00	.94/.94/.85/1.00	1.00/1.00/1.00/ 1.00	1.00/1.00/1.00/ 1.00	1.00/1.00/1.00/ 1.00	1.00/1.00/1.00/ 1.00	1.00/1.00/1.00/ 1.00
gemini-3.1-pro-preview	.02/.02/.00/.06	.10/.10/.02/.19	.06/.06/.00/.15	.00/.00/.00/.00	.00/.00/.00/.00	.00/.00/.00/.00	.04/.06/.00/.10	--	--
gemini-3.8-flash	.28/.28/.15/.40	.34/.36/.21/.47	.19/.21/.08/.29	.12/.12/.04/.23	.04/.04/.00/.10	.08/.08/.02/.17	.11/.11/.02/.19	--	--
gpt-5.6-luna	.83/1.00/.73/.94	.96/1.00/.90/1.00	.92/1.00/.83/.98	.79/1.00/.67/.90	.77/1.00/.65/.88	.88/1.00/.77/.96	.80/.98/.67/.91	--	--
grok-4.6	1.00/1.00/1.00/ 1.00	1.00/1.00/1.00/ 1.00	1.00/1.00/1.00/ 1.00	.98/1.00/.94/1.00	1.00/1.00/1.00/ 1.00	1.00/1.00/1.00/ 1.00	1.00/1.00/1.00/ 1.00	--	--
mistral-large	.00/.00/.00/.00	.00/.00/.00/.00	.00/.00/.00/.00	.00/.00/.00/.00	.00/.00/.00/.00	.00/.00/.00/.00	.00/.00/.00/.00	--	--
muse-spark-1.3	1.00/1.00/1.00/ 1.00	1.00/1.00/1.00/ 1.00	1.00/1.00/1.00/ 1.00	.98/.98/.93/1.00	1.00/1.00/1.00/ 1.00	1.00/1.00/1.00/ 1.00	1.00/1.00/1.00/ 1.00	--	--
phi-4	.00/.00/.00/.00	.00/.00/.00/.00	.00/.00/.00/.00	.00/.00/.00/.00	.00/.00/.00/.00	.00/.00/.00/.00	.00/.00/.00/.00	--	--
qwen3.8-a95b	1.00/1.00/1.00/ 1.00	1.00/1.00/1.00/ 1.00	1.00/1.00/1.00/ 1.00	1.00/1.00/1.00/ 1.00	1.00/1.00/1.00/ 1.00	1.00/1.00/1.00/ 1.00	.88/1.00/.69/1.00	--	--
qwen3.8-max	1.00/1.00/1.00/ 1.00	1.00/1.00/1.00/ 1.00	1.00/1.00/1.00/ 1.00	1.00/1.00/1.00/ 1.00	.98/1.00/.93/1.00	1.00/1.00/1.00/ 1.00	1.00/1.00/1.00/ 1.00	--	--

C Corpus and Transformation Audit

Figure 3 draws the evidence path these audits sit on, and Figure 4 profiles the released corpus by variable tier, clause count, solution cardinality, and view length.

Exact generator. At each draw, clause length is two with probability 0.65 and three otherwise; variables are sampled uniformly without replacement and literal polarities are independent fair draws. Duplicate clauses are rejected. The next unique clause defines the satisfiability boundary, with at most $12n$ clauses per attempt and 200 attempts per seed. Construction keeps the SAT or UNSAT side by seed parity.

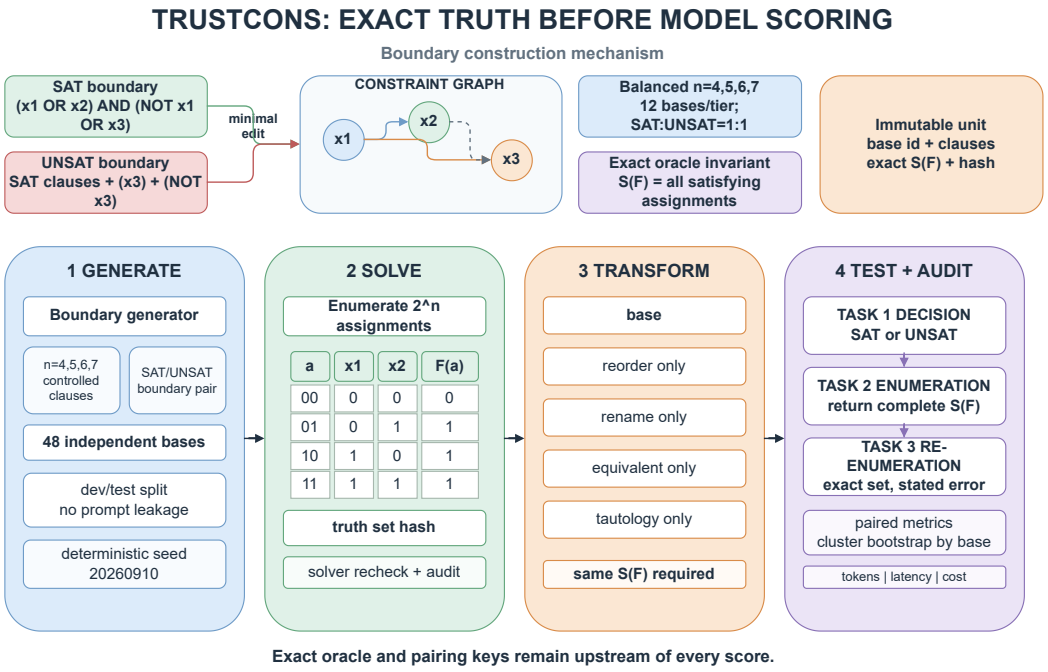

Figure 3: TrustCons evidence path. Boundary-selected formulas are solved before rendering; five pure views feed three exact tasks. The immutable response ledger, censoring audit, paired statistics, and manuscript artifacts remain downstream of the oracle.

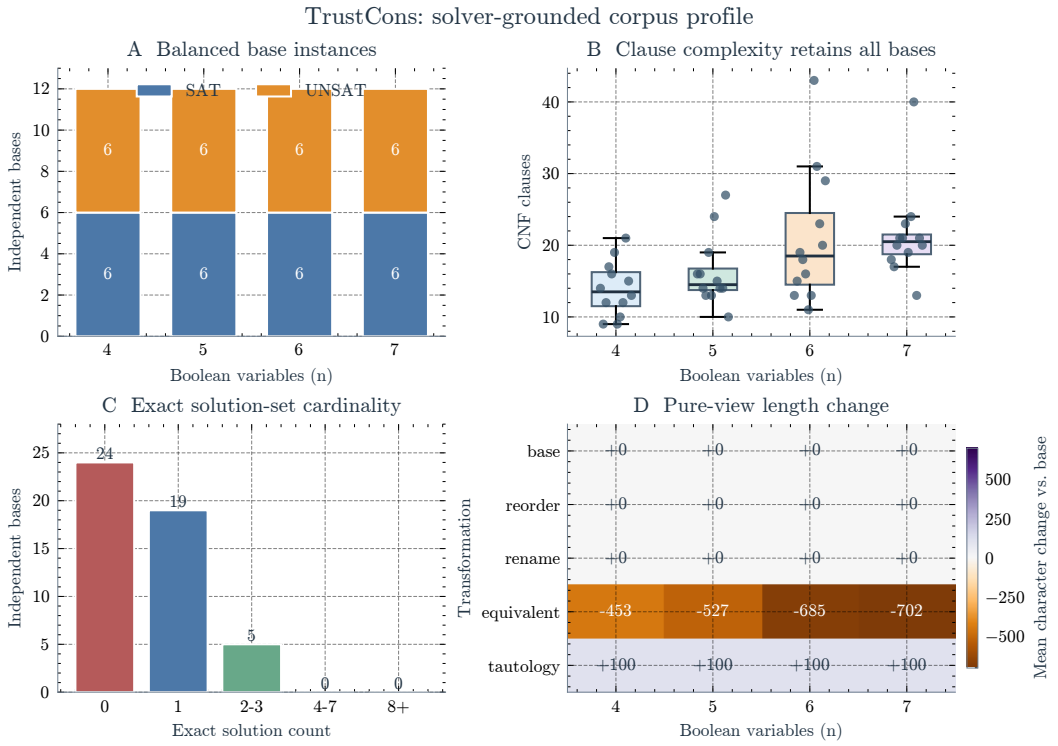

Figure 4: Corpus distribution by variable tier, clause count, solution cardinality, and view length. Each count refers to base instances unless labeled as a rendering measurement.

Table 6: CNF construction and validation rules. Full machine-readable values are supplied as CSV.

rule_id	rule	implemented	deterministic	checked_in_generate	checked_in_test	audit_owner
1	literal domain	True	True	True	True	pipeline
2	clause width	True	True	True	True	pipeline
3	conjunction semantics	True	True	True	True	pipeline
4	boundary SAT seed	True	True	True	True	pipeline
5	minimal UNSAT edit	True	True	True	True	pipeline
6	SAT/UNSAT balance	True	True	True	True	pipeline
7	difficulty tier	True	True	True	True	pipeline
8	unique base hash	True	True	True	True	pipeline
9	exhaustive oracle	True	True	True	True	pipeline
10	solver recheck	True	True	True	True	pipeline
11	dev/test isolation	True	True	True	True	pipeline
12	stable request key	True	True	True	True	pipeline
13	transport-only retry	True	True	True	True	pipeline
14	budget guard	True	True	True	True	pipeline
15	immutable append	True	True	True	True	pipeline
16	pure reorder	True	True	True	True	pipeline
17	pure rename	True	True	True	True	pipeline
18	pure equivalence	True	True	True	True	pipeline
19	pure tautology	True	True	True	True	pipeline
20	paired base cluster	True	True	True	True	pipeline

Table 7: Dataset distribution cube. Full machine-readable values are supplied as CSV.

instance	tier	n	label	clauses	solutions	B chars	O chars	N chars	Q chars	X chars
n4-000	n4	4	SAT	14	1	728	728	728	270	828
n4-001	n4	4	UNSAT	9	0	460	460	460	169	560
n4-002	n4	4	SAT	12	1	629	629	629	242	729
n4-003	n4	4	UNSAT	17	0	906	906	906	341	1006
n4-004	n4	4	SAT	16	1	837	837	837	318	937
n4-005	n4	4	UNSAT	19	0	1014	1014	1014	411	1114
n4-006	n4	4	SAT	9	1	491	491	491	191	591
n4-007	n4	4	UNSAT	10	0	503	503	503	196	603
n4-008	n4	4	SAT	12	1	642	642	642	244	742
n4-009	n4	4	UNSAT	21	0	1105	1105	1105	455	1205
n4-010	n4	4	SAT	15	1	789	789	789	273	889
n4-011	n4	4	UNSAT	13	0	681	681	681	239	781
n5-000	n5	5	SAT	16	1	860	860	860	332	960
n5-001	n5	5	UNSAT	16	0	836	836	836	296	936
n5-002	n5	5	SAT	14	3	746	746	746	290	846
n5-003	n5	5	UNSAT	13	0	658	658	658	253	758
n5-004	n5	5	SAT	19	1	997	997	997	393	1097
n5-005	n5	5	UNSAT	13	0	662	662	662	273	762
n5-006	n5	5	SAT	24	1	1286	1286	1286	496	1386
n5-007	n5	5	UNSAT	15	0	800	800	800	283	900
n5-008	n5	5	SAT	14	1	734	734	734	270	834
n5-009	n5	5	UNSAT	14	0	738	738	738	298	838
n5-010	n5	5	SAT	27	1	1391	1391	1391	541	1491
n5-011	n5	5	UNSAT	10	0	514	514	514	178	614
n6-000	n6	6	SAT	13	1	682	682	682	237	782
n6-001	n6	6	UNSAT	15	0	794	794	794	283	894
n6-002	n6	6	SAT	19	1	1026	1026	1026	403	1126
n6-003	n6	6	UNSAT	18	0	968	968	968	374	1068
n6-004	n6	6	SAT	16	2	847	847	847	314	947
n6-005	n6	6	UNSAT	11	0	568	568	568	195	668
n6-006	n6	6	SAT	13	2	705	705	705	275	805
n6-007	n6	6	UNSAT	43	0	2280	2280	2280	883	2380
n6-008	n6	6	SAT	31	1	1618	1618	1618	659	1718
n6-009	n6	6	UNSAT	23	0	1230	1230	1230	495	1330
n6-010	n6	6	SAT	20	3	1086	1086	1086	404	1186
n6-011	n6	6	UNSAT	29	0	1576	1576	1576	641	1676
n7-000	n7	7	SAT	18	2	930	930	930	350	1030
n7-001	n7	7	UNSAT	17	0	917	917	917	347	1017
n7-002	n7	7	SAT	20	1	1021	1021	1021	386	1121
n7-003	n7	7	UNSAT	21	0	1128	1128	1128	433	1228
n7-004	n7	7	SAT	21	1	1118	1118	1118	437	1218
n7-005	n7	7	UNSAT	23	0	1255	1255	1255	493	1355
n7-006	n7	7	SAT	19	1	989	989	989	353	1089
n7-007	n7	7	UNSAT	24	0	1268	1268	1268	512	1368
n7-008	n7	7	SAT	40	1	2084	2084	2084	764	2184

instance	tier	n	label	clauses	solutions	B chars	O chars	N chars	Q chars	X chars
n7-009	n7	7	UNSAT	13	0	672	672	672	233	772
n7-010	n7	7	SAT	21	1	1093	1093	1093	399	1193
n7-011	n7	7	UNSAT	20	0	1070	1070	1070	412	1170

Table 8: Five pure semantic transformations. Columns are B=base, O=reorder-only, N=rename-only, Q=equivalent-expression-only, and X=distractor-only; each cell names the operation, clause/rendered-rule counts, variable-name signature, and whether the rendering differs from base. Full machine-readable values are supplied as CSV.

instance	B	O	N	Q	X
n4-000	op=identity; clauses/rendered=14/14; names=ABCD; changed=N	op=order; clauses/rendered=14/14; names=ABCD; changed=Y	op=rename; clauses/rendered=14/14; names=KMRT; changed=Y	op=implication; clauses/rendered=14/14; names=ABCD; changed=Y	op=+2 tautologies; clauses/rendered=14/16; names=ABCD; changed=Y
n4-001	op=identity; clauses/rendered=9/9; names=ABCD; changed=N	op=order; clauses/rendered=9/9; names=ABCD; changed=Y	op=rename; clauses/rendered=9/9; names=KMRT; changed=Y	op=implication; clauses/rendered=9/9; names=ABCD; changed=Y	op=+2 tautologies; clauses/rendered=9/11; names=ABCD; changed=Y
n4-002	op=identity; clauses/rendered=12/12; names=ABCD; changed=N	op=order; clauses/rendered=12/12; names=ABCD; changed=Y	op=rename; clauses/rendered=12/12; names=KMRT; changed=Y	op=implication; clauses/rendered=12/12; names=ABCD; changed=Y	op=+2 tautologies; clauses/rendered=12/14; names=ABCD; changed=Y
n4-003	op=identity; clauses/rendered=17/17; names=ABCD; changed=N	op=order; clauses/rendered=17/17; names=ABCD; changed=Y	op=rename; clauses/rendered=17/17; names=KMRT; changed=Y	op=implication; clauses/rendered=17/17; names=ABCD; changed=Y	op=+2 tautologies; clauses/rendered=17/19; names=ABCD; changed=Y
n4-004	op=identity; clauses/rendered=16/16; names=ABCD; changed=N	op=order; clauses/rendered=16/16; names=ABCD; changed=Y	op=rename; clauses/rendered=16/16; names=KMRT; changed=Y	op=implication; clauses/rendered=16/16; names=ABCD; changed=Y	op=+2 tautologies; clauses/rendered=16/18; names=ABCD; changed=Y

D Complete Decision and Enumeration Results

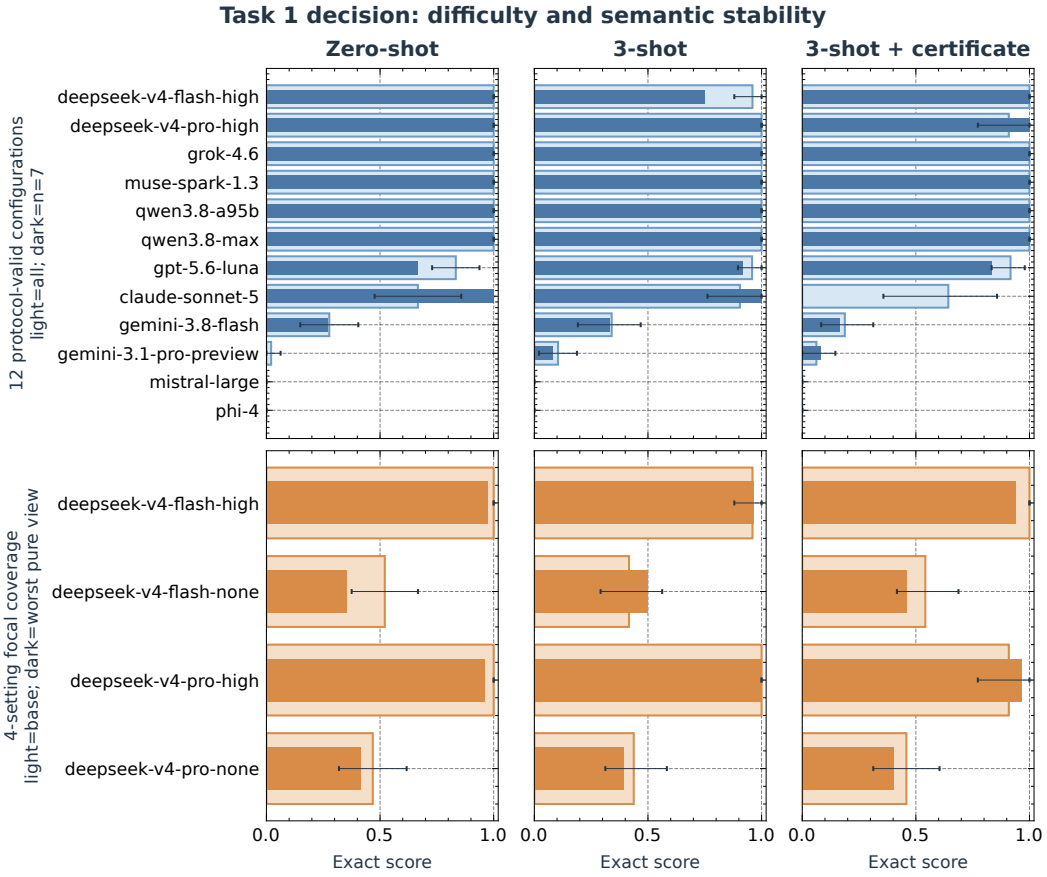

Figure 5: Decision accuracy across prompts and structural tiers. Light bars show all 48 bases and dark bars the $n = 7$ tier; the lower row contrasts base with the worst pure view for the four focal settings. Censored cells are absent, not zero.

Table 9: Decision decomposition across prompts and views. Full machine-readable values are supplied as CSV.

system_id | prompt_mode | difficulty	B A/F	X A/F	Q A/F	N A/F	O A/F
deepseek-v4-flash-high | 3S | n4	1.00/1.00	.92/1.00	1.00/1.00	.90/1.00	1.00/1.00
deepseek-v4-flash-high | 3S | n5	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00
deepseek-v4-flash-high | 3S | n6	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00
deepseek-v4-flash-high | 3S | n7	.75/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00
deepseek-v4-flash-high | 3S+C | n4	1.00/1.00	1.00/1.00	1.00/1.00	.91/1.00	1.00/1.00
deepseek-v4-flash-high | 3S+C | n5	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00
deepseek-v4-flash-high | 3S+C | n6	1.00/1.00	1.00/1.00	1.00/1.00	.88/1.00	1.00/1.00
deepseek-v4-flash-high | 3S+C | n7	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00
deepseek-v4-flash-high | 0S | n4	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00
deepseek-v4-flash-high | 0S | n5	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00
deepseek-v4-flash-high | 0S | n6	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00
deepseek-v4-flash-high | 0S | n7	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	.89/.89
deepseek-v4-flash-none | 3S | n4	.33/1.00	.33/1.00	.50/1.00	.58/1.00	.50/1.00
deepseek-v4-flash-none | 3S | n5	.42/1.00	.50/1.00	.50/1.00	.50/1.00	.50/1.00
deepseek-v4-flash-none | 3S | n6	.50/1.00	.58/1.00	.50/1.00	.67/1.00	.58/1.00
deepseek-v4-flash-none | 3S | n7	.42/1.00	.58/1.00	.58/1.00	.50/1.00	.50/1.00
deepseek-v4-flash-none | 3S+C | n4	.50/1.00	.42/1.00	.50/1.00	.42/1.00	.50/1.00
deepseek-v4-flash-none | 3S+C | n5	.50/1.00	.42/1.00	.50/1.00	.50/1.00	.33/1.00
deepseek-v4-flash-none | 3S+C | n6	.67/1.00	.50/1.00	.58/1.00	.42/1.00	.50/1.00
deepseek-v4-flash-none | 3S+C | n7	.50/1.00	.67/1.00	.42/1.00	.50/1.00	.50/1.00
deepseek-v4-flash-none | 0S | n4	.50/1.00	.42/1.00	.50/1.00	.50/1.00	.42/1.00
deepseek-v4-flash-none | 0S | n5	.42/1.00	.50/1.00	.50/1.00	.42/1.00	.25/1.00
deepseek-v4-flash-none | 0S | n6	.58/1.00	.50/1.00	.58/1.00	.67/1.00	.17/1.00
deepseek-v4-flash-none | 0S | n7	.58/1.00	.33/1.00	.50/1.00	.50/1.00	.58/1.00
deepseek-v4-pro-high | 3S | n4	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00
deepseek-v4-pro-high | 3S | n5	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00
deepseek-v4-pro-high | 3S | n6	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00
deepseek-v4-pro-high | 3S | n7	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00
deepseek-v4-pro-high | 3S+C | n4	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00
deepseek-v4-pro-high | 3S+C | n5	.83/.83	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00
deepseek-v4-pro-high | 3S+C | n6	.75/.75	1.00/1.00	1.00/1.00	.83/.83	1.00/1.00
deepseek-v4-pro-high | 3S+C | n7	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00

system_id | prompt_mode | difficulty	B A/F	X A/F	Q A/F	N A/F	O A/F
deepseek-v4-pro-high | 0S | n4	1.00/1.00	1.00/1.00	.90/.90	1.00/1.00	1.00/1.00
deepseek-v4-pro-high | 0S | n5	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00
deepseek-v4-pro-high | 0S | n6	1.00/1.00	.80/.80	1.00/1.00	.80/.80	1.00/1.00
deepseek-v4-pro-high | 0S | n7	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00
deepseek-v4-pro-none | 3S | n4	.42/1.00	.42/1.00	.45/1.00	.33/1.00	.58/1.00
deepseek-v4-pro-none | 3S | n5	.33/1.00	.50/1.00	.50/1.00	.33/1.00	.42/1.00
deepseek-v4-pro-none | 3S | n6	.50/1.00	.50/1.00	.50/1.00	.33/1.00	.25/1.00
deepseek-v4-pro-none | 3S | n7	.50/1.00	.50/1.00	.42/1.00	.58/1.00	.42/1.00
deepseek-v4-pro-none | 3S+C | n4	.50/1.00	.50/1.00	.50/1.00	.45/1.00	.58/1.00
deepseek-v4-pro-none | 3S+C | n5	.25/1.00	.33/1.00	.50/1.00	.25/1.00	.33/1.00
deepseek-v4-pro-none | 3S+C | n6	.58/1.00	.42/1.00	.50/1.00	.50/1.00	.58/1.00
deepseek-v4-pro-none | 3S+C | n7	.50/1.00	.50/1.00	.50/1.00	.42/1.00	.50/1.00
deepseek-v4-pro-none | 0S | n4	.42/.92	.50/1.00	.50/1.00	.67/1.00	.50/1.00
deepseek-v4-pro-none | 0S | n5	.42/1.00	.50/1.00	.50/1.00	.42/1.00	.42/1.00
deepseek-v4-pro-none | 0S | n6	.58/1.00	.50/1.00	.45/1.00	.42/1.00	.25/1.00
deepseek-v4-pro-none | 0S | n7	.45/1.00	.50/1.00	.42/1.00	.58/1.00	.50/1.00

Table 10: Decision label-conditioned results. Full machine-readable values are supplied as CSV.

system_id	0S: L,tier=A/F	3S: L,tier=A/F	3S+C: L,tier=A/F
claude-sonnet-5	L=SAT,D=n4:.67/.67; L=SAT,D=n5:.50/.50; L=SAT,D=n6:1.00/1.00; L=SAT,D=n7:1.00/1.00; L=UNSAT,D=n4:.67/.67; L=UNSAT,D=n5:.50/.50; L=UNSAT,D=n6:.50/.50; L=UNSAT,D=n7:NA/NA	L=SAT,D=n4:.75/.75; L=SAT,D=n5:.67/.67; L=SAT,D=n6:1.00/1.00; L=SAT,D=n7:1.00/1.00; L=UNSAT,D=n4:1.00/1.00; L=UNSAT,D=n5:1.00/1.00; L=UNSAT,D=n6:1.00/1.00; L=UNSAT,D=n7:NA/NA	L=SAT,D=n4:.67/.67; L=SAT,D=n5:NA/NA; L=SAT,D=n6:1.00/1.00; L=SAT,D=n7:NA/NA; L=UNSAT,D=n4:.75/.75; L=UNSAT,D=n5:1.00/1.00; L=UNSAT,D=n6:.33/.33; L=UNSAT,D=n7:.00/.00
deepseek-v4-flash-high	L=SAT,D=n4:1.00/1.00; L=SAT,D=n5:1.00/1.00; L=SAT,D=n6:1.00/1.00; L=SAT,D=n7:1.00/1.00; L=UNSAT,D=n4:1.00/1.00; L=UNSAT,D=n5:1.00/1.00; L=UNSAT,D=n6:1.00/1.00; L=UNSAT,D=n7:1.00/1.00	L=SAT,D=n4:1.00/1.00; L=SAT,D=n5:1.00/1.00; L=SAT,D=n6:1.00/1.00; L=SAT,D=n7:.50/1.00; L=UNSAT,D=n4:1.00/1.00; L=UNSAT,D=n5:1.00/1.00; L=UNSAT,D=n6:1.00/1.00; L=UNSAT,D=n7:1.00/1.00	L=SAT,D=n4:1.00/1.00; L=SAT,D=n5:1.00/1.00; L=SAT,D=n6:1.00/1.00; L=SAT,D=n7:1.00/1.00; L=UNSAT,D=n4:1.00/1.00; L=UNSAT,D=n5:1.00/1.00; L=UNSAT,D=n6:1.00/1.00; L=UNSAT,D=n7:1.00/1.00
deepseek-v4-pro-high	L=SAT,D=n4:1.00/1.00; L=SAT,D=n5:1.00/1.00; L=SAT,D=n6:1.00/1.00; L=SAT,D=n7:1.00/1.00; L=UNSAT,D=n4:1.00/1.00; L=UNSAT,D=n5:1.00/1.00; L=UNSAT,D=n6:1.00/1.00; L=UNSAT,D=n7:1.00/1.00	L=SAT,D=n4:1.00/1.00; L=SAT,D=n5:1.00/1.00; L=SAT,D=n6:1.00/1.00; L=SAT,D=n7:NA/NA; L=UNSAT,D=n4:1.00/1.00; L=UNSAT,D=n5:1.00/1.00; L=UNSAT,D=n6:1.00/1.00; L=UNSAT,D=n7:1.00/1.00	L=SAT,D=n4:1.00/1.00; L=SAT,D=n5:.75/.75; L=SAT,D=n6:1.00/1.00; L=SAT,D=n7:1.00/1.00; L=UNSAT,D=n4:1.00/1.00; L=UNSAT,D=n5:1.00/1.00; L=UNSAT,D=n6:.50/.50; L=UNSAT,D=n7:1.00/1.00
gemini-3.1-pro-preview	L=SAT,D=n4:.17/.17; L=SAT,D=n5:.00/.00; L=SAT,D=n6:.00/.00; L=SAT,D=n7:.00/.00; L=UNSAT,D=n4:.00/.00; L=UNSAT,D=n5:.00/.00; L=UNSAT,D=n6:.00/.00; L=UNSAT,D=n7:.00/.00	L=SAT,D=n4:.00/.00; L=SAT,D=n5:.00/.00; L=SAT,D=n6:.17/.17; L=SAT,D=n7:.00/.00; L=UNSAT,D=n4:.17/.17; L=UNSAT,D=n5:.33/.33; L=UNSAT,D=n6:.00/.00; L=UNSAT,D=n7:.17/.17	L=SAT,D=n4:.00/.00; L=SAT,D=n5:.00/.00; L=SAT,D=n6:.33/.33; L=SAT,D=n7:.00/.00; L=UNSAT,D=n4:.00/.00; L=UNSAT,D=n5:.00/.00; L=UNSAT,D=n6:.00/.00; L=UNSAT,D=n7:.17/.17
gemini-3.8-flash	L=SAT,D=n4:.17/.17; L=SAT,D=n5:.50/.50; L=SAT,D=n6:.17/.17; L=SAT,D=n7:.40/.40; L=UNSAT,D=n4:.33/.33; L=UNSAT,D=n5:.50/.50; L=UNSAT,D=n6:.00/.00; L=UNSAT,D=n7:.17/.17	L=SAT,D=n4:.50/.50; L=SAT,D=n5:.33/.33; L=SAT,D=n6:.33/.33; L=SAT,D=n7:.50/.50; L=UNSAT,D=n4:.17/.17; L=UNSAT,D=n5:.33/.33; L=UNSAT,D=n6:.40/.40; L=UNSAT,D=n7:.17/.33	L=SAT,D=n4:.00/.00; L=SAT,D=n5:.17/.17; L=SAT,D=n6:.17/.17; L=SAT,D=n7:.33/.33; L=UNSAT,D=n4:.33/.33; L=UNSAT,D=n5:.50/.50; L=UNSAT,D=n6:.00/.17; L=UNSAT,D=n7:.00/.00
gpt-5.6-luna	L=SAT,D=n4:1.00/1.00; L=SAT,D=n5:1.00/1.00; L=SAT,D=n6:1.00/1.00; L=SAT,D=n7:1.00/1.00; L=UNSAT,D=n4:.50/1.00; L=UNSAT,D=n5:1.00/1.00; L=UNSAT,D=n6:.83/1.00; L=UNSAT,D=n7:.33/1.00	L=SAT,D=n4:1.00/1.00; L=SAT,D=n5:1.00/1.00; L=SAT,D=n6:1.00/1.00; L=SAT,D=n7:.83/1.00; L=UNSAT,D=n4:1.00/1.00; L=UNSAT,D=n5:1.00/1.00; L=UNSAT,D=n6:.83/1.00; L=UNSAT,D=n7:1.00/1.00	L=SAT,D=n4:1.00/1.00; L=SAT,D=n5:1.00/1.00; L=SAT,D=n6:1.00/1.00; L=SAT,D=n7:1.00/1.00; L=UNSAT,D=n4:.83/1.00; L=UNSAT,D=n5:1.00/1.00; L=UNSAT,D=n6:.83/1.00; L=UNSAT,D=n7:.67/1.00
grok-4.6	L=SAT,D=n4:1.00/1.00; L=SAT,D=n5:1.00/1.00; L=SAT,D=n6:1.00/1.00; L=SAT,D=n7:1.00/1.00; L=UNSAT,D=n4:1.00/1.00; L=UNSAT,D=n5:1.00/1.00; L=UNSAT,D=n6:1.00/1.00; L=UNSAT,D=n7:1.00/1.00	L=SAT,D=n4:1.00/1.00; L=SAT,D=n5:1.00/1.00; L=SAT,D=n6:1.00/1.00; L=SAT,D=n7:1.00/1.00; L=UNSAT,D=n4:1.00/1.00; L=UNSAT,D=n5:1.00/1.00; L=UNSAT,D=n6:1.00/1.00; L=UNSAT,D=n7:1.00/1.00	L=SAT,D=n4:1.00/1.00; L=SAT,D=n5:1.00/1.00; L=SAT,D=n6:1.00/1.00; L=SAT,D=n7:1.00/1.00; L=UNSAT,D=n4:1.00/1.00; L=UNSAT,D=n5:1.00/1.00; L=UNSAT,D=n6:1.00/1.00; L=UNSAT,D=n7:1.00/1.00
mistral-large	L=SAT,D=n4:.00/.00; L=SAT,D=n5:.00/.00; L=SAT,D=n6:.00/.00; L=SAT,D=n7:.00/.00; L=UNSAT,D=n4:.00/.00; L=UNSAT,D=n5:.00/.00; L=UNSAT,D=n6:.00/.00; L=UNSAT,D=n7:.00/.00	L=SAT,D=n4:.00/.00; L=SAT,D=n5:.00/.00; L=SAT,D=n6:.00/.00; L=SAT,D=n7:.00/.00; L=UNSAT,D=n4:.00/.00; L=UNSAT,D=n5:.00/.00; L=UNSAT,D=n6:.00/.00; L=UNSAT,D=n7:.00/.00	L=SAT,D=n4:.00/.00; L=SAT,D=n5:.00/.00; L=SAT,D=n6:.00/.00; L=SAT,D=n7:.00/.00; L=UNSAT,D=n4:.00/.00; L=UNSAT,D=n5:.00/.00; L=UNSAT,D=n6:.00/.00; L=UNSAT,D=n7:.00/.00
muse-spark-1.3	L=SAT,D=n4:1.00/1.00; L=SAT,D=n5:1.00/1.00; L=SAT,D=n6:NA/NA; L=SAT,D=n7:NA/NA; L=UNSAT,D=n4:1.00/1.00; L=UNSAT,D=n5:1.00/1.00; L=UNSAT,D=n6:1.00/1.00; L=UNSAT,D=n7:1.00/1.00	L=SAT,D=n4:1.00/1.00; L=SAT,D=n5:1.00/1.00; L=SAT,D=n6:1.00/1.00; L=SAT,D=n7:NA/NA; L=UNSAT,D=n4:1.00/1.00; L=UNSAT,D=n5:1.00/1.00; L=UNSAT,D=n6:NA/NA; L=UNSAT,D=n7:1.00/1.00	L=SAT,D=n4:1.00/1.00; L=SAT,D=n5:1.00/1.00; L=SAT,D=n6:NA/NA; L=SAT,D=n7:1.00/1.00; L=UNSAT,D=n4:1.00/1.00; L=UNSAT,D=n5:1.00/1.00; L=UNSAT,D=n6:NA/NA; L=UNSAT,D=n7:NA/NA
phi-4	L=SAT,D=n4:.00/.00; L=SAT,D=n5:.00/.00; L=SAT,D=n6:.00/.00; L=SAT,D=n7:.00/.00; L=UNSAT,D=n4:.00/.00; L=UNSAT,D=n5:.00/.00; L=UNSAT,D=n6:.00/.00; L=UNSAT,D=n7:.00/.00	L=SAT,D=n4:.00/.00; L=SAT,D=n5:.00/.00; L=SAT,D=n6:.00/.00; L=SAT,D=n7:.00/.00; L=UNSAT,D=n4:.00/.00; L=UNSAT,D=n5:.00/.00; L=UNSAT,D=n6:.00/.00; L=UNSAT,D=n7:.00/.00	L=SAT,D=n4:.00/.00; L=SAT,D=n5:.00/.00; L=SAT,D=n6:.00/.00; L=SAT,D=n7:.00/.00; L=UNSAT,D=n4:.00/.00; L=UNSAT,D=n5:.00/.00; L=UNSAT,D=n6:.00/.00; L=UNSAT,D=n7:.00/.00

system_id	0S: L,tier=A/F	3S: L,tier=A/F	3S+C: L,tier=A/F
qwen3.8-a95b	L=SAT,D=n4:1.00/1.00; L=SAT,D=n5:1.00/1.00; L=SAT,D=n6:1.00/1.00; L=SAT,D=n7:1.00/1.00; L=UNSAT,D=n4:1.00/1.00; L=UNSAT,D=n5:1.00/1.00; L=UNSAT,D=n6:1.00/1.00; L=UNSAT,D=n7:1.00/1.00	L=SAT,D=n4:1.00/1.00; L=SAT,D=n5:1.00/1.00; L=SAT,D=n6:1.00/1.00; L=SAT,D=n7:1.00/1.00; L=UNSAT,D=n4:1.00/1.00; L=UNSAT,D=n5:1.00/1.00; L=UNSAT,D=n6:1.00/1.00; L=UNSAT,D=n7:1.00/1.00	L=SAT,D=n4:1.00/1.00; L=SAT,D=n5:1.00/1.00; L=SAT,D=n6:1.00/1.00; L=SAT,D=n7:1.00/1.00; L=UNSAT,D=n4:1.00/1.00; L=UNSAT,D=n5:1.00/1.00; L=UNSAT,D=n6:1.00/1.00; L=UNSAT,D=n7:1.00/1.00
qwen3.8-max	L=SAT,D=n4:1.00/1.00; L=SAT,D=n5:1.00/1.00; L=SAT,D=n6:1.00/1.00; L=SAT,D=n7:1.00/1.00; L=UNSAT,D=n4:1.00/1.00; L=UNSAT,D=n5:1.00/1.00; L=UNSAT,D=n6:1.00/1.00; L=UNSAT,D=n7:1.00/1.00	L=SAT,D=n4:1.00/1.00; L=SAT,D=n5:1.00/1.00; L=SAT,D=n6:1.00/1.00; L=SAT,D=n7:1.00/1.00; L=UNSAT,D=n4:1.00/1.00; L=UNSAT,D=n5:1.00/1.00; L=UNSAT,D=n6:1.00/1.00; L=UNSAT,D=n7:1.00/1.00	L=SAT,D=n4:1.00/1.00; L=SAT,D=n5:1.00/1.00; L=SAT,D=n6:1.00/1.00; L=SAT,D=n7:1.00/1.00; L=UNSAT,D=n4:1.00/1.00; L=UNSAT,D=n5:1.00/1.00; L=UNSAT,D=n6:1.00/1.00; L=UNSAT,D=n7:1.00/1.00

Table 11: Enumeration precision and recall. Full machine-readable values are supplied as CSV.

system_id | prompt_mode	n4 E/P/R/F/Fmt	n5 E/P/R/F/Fmt	n6 E/P/R/F/Fmt	n7 E/P/R/F/Fmt
claude-sonnet-5 | 3S	.55/1.00/1.00/1.00/.55	.09/1.00/1.00/1.00/.09	.00/.71/.71/.71/.00	.00/1.00/1.00/1.00/.00
claude-sonnet-5 | 3S+C	.17/1.00/1.00/1.00/.17	.10/1.00/1.00/1.00/.10	.00/1.00/1.00/1.00/.00	.00/1.00/1.00/1.00/.00
claude-sonnet-5 | 0S	.08/1.00/1.00/1.00/.08	.00/1.00/1.00/1.00/.00	.00/1.00/1.00/1.00/.00	.00/.88/.88/.88/.00
deepseek-v4-flash-high | 3S	1.00/1.00/1.00/1.00/1.00	1.00/1.00/1.00/1.00/1.00	1.00/1.00/1.00/1.00/1.00	.89/.89/.89/.89/1.00
deepseek-v4-flash-high | 3S+C	1.00/1.00/1.00/1.00/1.00	1.00/1.00/1.00/1.00/1.00	1.00/1.00/1.00/1.00/1.00	1.00/1.00/1.00/1.00/1.00
deepseek-v4-flash-high | 0S	.91/.91/.91/.91/.91	.90/.90/.90/.90/.90	.88/1.00/.92/.94/1.00	1.00/1.00/1.00/1.00/1.00
deepseek-v4-pro-high | 3S	1.00/1.00/1.00/1.00/1.00	1.00/1.00/1.00/1.00/1.00	1.00/1.00/1.00/1.00/1.00	1.00/1.00/1.00/1.00/1.00
deepseek-v4-pro-high | 3S+C	1.00/1.00/1.00/1.00/1.00	1.00/1.00/1.00/1.00/1.00	1.00/1.00/1.00/1.00/1.00	1.00/1.00/1.00/1.00/1.00
deepseek-v4-pro-high | 0S	.90/.90/.90/.90/.90	1.00/1.00/1.00/1.00/1.00	1.00/1.00/1.00/1.00/1.00	.88/.88/.88/.88/.88
gemini-3.1-pro-preview | 3S	.00/1.00/1.00/1.00/.00	.00/1.00/1.00/1.00/.00	.00/1.00/1.00/1.00/.00	.00/1.00/1.00/1.00/.00
gemini-3.1-pro-preview | 3S+C	.00/1.00/1.00/1.00/.00	.00/1.00/1.00/1.00/.00	.00/1.00/1.00/1.00/.00	.00/1.00/1.00/1.00/.00
gemini-3.1-pro-preview | 0S	.00/1.00/1.00/1.00/.00	.00/1.00/1.00/1.00/.00	.00/.83/.83/.83/.00	.00/.58/.58/.58/.00
gemini-3.8-flash | 3S	.08/1.00/1.00/1.00/.08	.00/.92/.92/.92/.00	.08/.92/.92/.92/.08	.00/.92/.92/.92/.00
gemini-3.8-flash | 3S+C	.00/.92/.92/.92/.00	.25/.92/.92/.92/.25	.08/.92/.92/.92/.08	.00/.75/.75/.75/.00
gemini-3.8-flash | 0S	.25/1.00/1.00/1.00/.25	.17/1.00/1.00/1.00/.17	.08/.92/.92/.92/.08	.00/1.00/1.00/1.00/.00
gpt-5.6-luna | 3S	.75/.82/.92/.85/1.00	.92/.92/.92/.92/1.00	.67/.72/.75/.73/1.00	.75/.78/.83/.79/1.00
gpt-5.6-luna | 3S+C	.92/.92/.92/.92/1.00	1.00/1.00/1.00/1.00/1.00	.83/.89/.92/.90/1.00	.75/.75/.75/.75/1.00
gpt-5.6-luna | 0S	.92/.96/1.00/.97/1.00	.83/.89/.89/.89/1.00	.83/.83/.83/.83/1.00	.58/.71/.83/.75/1.00
grok-4.6 | 3S	1.00/1.00/1.00/1.00/1.00	1.00/1.00/1.00/1.00/1.00	1.00/1.00/1.00/1.00/1.00	1.00/1.00/1.00/1.00/1.00
grok-4.6 | 3S+C	1.00/1.00/1.00/1.00/1.00	1.00/1.00/1.00/1.00/1.00	1.00/1.00/1.00/1.00/1.00	1.00/1.00/1.00/1.00/1.00
grok-4.6 | 0S	1.00/1.00/1.00/1.00/1.00	1.00/1.00/1.00/1.00/1.00	1.00/1.00/1.00/1.00/1.00	.92/.92/.92/.92/1.00
mistral-large | 3S	.00/.06/.17/.08/.00	.00/.05/.17/.07/.00	.00/.12/.19/.13/.00	.00/.01/.08/.02/.00
mistral-large | 3S+C	.00/.04/.25/.07/.00	.00/.04/.08/.05/.00	.00/.04/.17/.06/.00	.00/.00/.00/.00/.00
mistral-large | 0S	.00/.01/.08/.02/.00	.00/.02/.17/.03/.00	.00/.01/.17/.02/.00	.00/.02/.25/.03/.00
muse-spark-1.3 | 3S	1.00/1.00/1.00/1.00/1.00	1.00/1.00/1.00/1.00/1.00	1.00/1.00/1.00/1.00/1.00	1.00/1.00/1.00/1.00/1.00
muse-spark-1.3 | 3S+C	1.00/1.00/1.00/1.00/1.00	1.00/1.00/1.00/1.00/1.00	1.00/1.00/1.00/1.00/1.00	1.00/1.00/1.00/1.00/1.00
muse-spark-1.3 | 0S	.91/.91/.91/.91/.91	1.00/1.00/1.00/1.00/1.00	1.00/1.00/1.00/1.00/1.00	1.00/1.00/1.00/1.00/1.00
phi-4 | 3S	.00/.02/.08/.03/.00	.00/.02/.08/.03/.00	.00/.03/.09/.05/.00	.00/.00/.00/.00/.00
phi-4 | 3S+C	.00/.02/.08/.03/.00	.00/.00/.00/.00/.00	.00/.02/.12/.03/.00	.00/.00/.25/.01/.00
phi-4 | 0S	.00/.04/.33/.07/.00	.00/.04/.25/.06/.00	.00/.01/.25/.02/.00	.00/.00/.17/.00/.00
qwen3.8-a95b | 3S	1.00/1.00/1.00/1.00/1.00	1.00/1.00/1.00/1.00/1.00	1.00/1.00/1.00/1.00/1.00	1.00/1.00/1.00/1.00/1.00
qwen3.8-a95b | 3S+C	1.00/1.00/1.00/1.00/1.00	1.00/1.00/1.00/1.00/1.00	1.00/1.00/1.00/1.00/1.00	1.00/1.00/1.00/1.00/1.00
qwen3.8-a95b | 0S	1.00/1.00/1.00/1.00/1.00	1.00/1.00/1.00/1.00/1.00	1.00/1.00/1.00/1.00/1.00	1.00/1.00/1.00/1.00/1.00
qwen3.8-max | 3S	1.00/1.00/1.00/1.00/1.00	.91/.91/.91/.91/1.00	1.00/1.00/1.00/1.00/1.00	1.00/1.00/1.00/1.00/1.00
qwen3.8-max | 3S+C	1.00/1.00/1.00/1.00/1.00	1.00/1.00/1.00/1.00/1.00	1.00/1.00/1.00/1.00/1.00	1.00/1.00/1.00/1.00/1.00
qwen3.8-max | 0S	1.00/1.00/1.00/1.00/1.00	1.00/1.00/1.00/1.00/1.00	1.00/1.00/1.00/1.00/1.00	1.00/1.00/1.00/1.00/1.00

Matched-cardinality contrasts. The family-level $K{=}32 - K{=}1$ effects are 0.0 pp (Flash), -30.4 pp (Gemini), -2.8 pp (Luna), and 0.0 pp (Phi-4). Their Holm-adjusted sign-test values are 1, 0.02954, 1, and 1, respectively; strict-versus-lenient aggregate rank correlation is 1.00. Table 12 gives the full intervals, observed denominators, bounds, and repetition summaries.

Table 12: Enumeration by solution-set cardinality. The eight slash-separated fields are E/CI-low/CI-high/R/Fmt/p/pH/nFam, where pH is Holm-adjusted and nFam is the number of complete formula families. Full machine-readable values are supplied as CSV.

system_id	0S: band,tier=E/CIlo/CIhi/R/Fmt/p/pH/nFam	3S: band,tier=E/CIlo/CIhi/R/Fmt/p/pH/nFam	3S+C: band,tier=E/CIlo/CIhi/R/Fmt/p/pH/nFam	family_contrast: band,tier=E/CIlo/CIhi/R/Fmt/p/pH/nFam
claude-sonnet-5	S=0,D=n4:.00/.00/.00/1.00/.00/--/--/--; S=0,D=n5:.00/.00/.00/1.00/.00/--/--/--; S=0,D=n6:.00/.00/.00/1.00/.00/--/--/--; S=0,D=n7:.00/.00/.00/1.00/.00/--/--/--; S=1,D=n4:.17/.00/.50/1.00/.17/--/--/--; S=1,D=n5:.00/.00/.00/1.00/.00/--/--/--; S=1,D=n6:.00/.00/.00/1.00/.00/--/--/--; S=1,D=n7:.00/.00/.00/1.00/.00/--/--/--; S=2-3,D=n5:.00/.00/.00/1.00/.00/--/--/--; S=2-3,D=n6:.00/.00/.00/1.00/.00/--/--/--; S=2-3,D=n7:.00/.00/.00/.00/.00/--/--/--	S=0,D=n4:.20/.00/.60/1.00/.20/--/--/--; S=0,D=n5:.17/.00/.50/1.00/.17/--/--/--; S=0,D=n6:.00/.00/.00/1.00/.00/--/--/--; S=0,D=n7:.00/.00/.00/1.00/.00/--/--/--; S=1,D=n4:.83/.50/1.00/1.00/.83/--/--/--; S=1,D=n5:.00/.00/.00/1.00/.00/--/--/--; S=1,D=n6:.00/.00/.00/1.00/.00/--/--/--; S=1,D=n7:.00/.00/.00/1.00/.00/--/--/--; S=2-3,D=n5:.00/.00/.00/1.00/.00/--/--/--; S=2-3,D=n6:.00/.00/.00/.00/.00/--/--/--; S=2-3,D=n7:.00/.00/.00/1.00/.00/--/--/--	S=0,D=n4:.17/.00/.50/1.00/.17/--/--/--; S=0,D=n5:.00/.00/.00/1.00/.00/--/--/--; S=0,D=n6:.00/.00/.00/1.00/.00/--/--/--; S=0,D=n7:.00/.00/.00/1.00/.00/--/--/--; S=1,D=n4:.17/.00/.50/1.00/.17/--/--/--; S=1,D=n5:.00/.00/.00/1.00/.00/--/--/--; S=1,D=n6:.00/.00/.00/1.00/.00/--/--/--; S=1,D=n7:.00/.00/.00/1.00/.00/--/--/--; S=2-3,D=n5:1.00/1.00/1.00/1.00/1.00/--/--/--; S=2-3,D=n6:.00/.00/.00/1.00/.00/--/--/--; S=2-3,D=n7:.00/.00/.00/1.00/.00/--/--/--	--
deepseek-v4-flash-high	S=0,D=n4:1.00/1.00/1.00/1.00/1.00/--/--/--; S=0,D=n5:1.00/1.00/1.00/1.00/1.00/--/--/--; S=0,D=n6:1.00/1.00/1.00/1.00/1.00/--/--/--; S=0,D=n7:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=matched-extension:.99/.96/1.00/.99/1.00/--/--/--; S=1,D=n4:.83/.50/1.00/.83/.83/--/--/--; S=1,D=n5:.75/.25/1.00/.75/.75/--/--/--; S=1,D=n6:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n7:1.00/1.00/1.00/1.00/1.00/--/--/--; S=16,D=matched-extension:.84/.69/.96/.84/1.00/--/--/--; S=2,D=matched-extension:.92/.86/.97/.92/.93/--/--/--; S=2-3,D=n5:1.00/1.00/1.00/1.00/1.00/--/--/--; S=2-3,D=n6:.50/.00/1.00/.67/1.00/--/--/--; S=2-3,D=n7:1.00/1.00/1.00/1.00/1.00/--/--/--; S=32,D=matched-extension:.98/.94/1.00/.98/1.00/--/--/--; S=4,D=matched-extension:.93/.86/.99/.93/1.00/--/--/--; S=8,D=matched-extension:.93/.85/1.00/.93/1.00/--/--/--	S=0,D=n4:1.00/1.00/1.00/1.00/1.00/--/--/--; S=0,D=n5:1.00/1.00/1.00/1.00/1.00/--/--/--; S=0,D=n6:1.00/1.00/1.00/1.00/1.00/--/--/--; S=0,D=n7:.83/.50/1.00/.83/1.00/--/--/--; S=1,D=n4:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n5:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n6:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n7:1.00/1.00/1.00/1.00/1.00/--/--/--; S=2-3,D=n5:1.00/1.00/1.00/1.00/1.00/--/--/--; S=2-3,D=n6:1.00/1.00/1.00/1.00/1.00/--/--/--; S=2-3,D=n7:1.00/1.00/1.00/1.00/1.00/--/--/--	S=0,D=n4:1.00/1.00/1.00/1.00/1.00/--/--/--; S=0,D=n5:1.00/1.00/1.00/1.00/1.00/--/--/--; S=0,D=n6:1.00/1.00/1.00/1.00/1.00/--/--/--; S=0,D=n7:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n4:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n5:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n6:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n7:1.00/1.00/1.00/1.00/1.00/--/--/--; S=2-3,D=n5:1.00/1.00/1.00/1.00/1.00/--/--/--; S=2-3,D=n6:1.00/1.00/1.00/1.00/1.00/--/--/--; S=2-3,D=n7:1.00/1.00/1.00/1.00/1.00/--/--/--	S=K32-K1,D=matched-extension:.00/.00/.00/--/--/1.00/1.00/17

system_id	0S: band,tier=E/CIlo/CIhi/R/Fmt/p/pH/nFam	3S: band,tier=E/CIlo/CIhi/R/Fmt/p/pH/nFam	3S+C: band,tier=E/CIlo/CIhi/R/Fmt/p/pH/nFam	family_contrast: band,tier=E/CIlo/CIhi/R/Fmt/p/pH/nFam
deepseek-v4-pro-high	S=0,D=n4:1.00/1.00/1.00/1.00/1.00/--/--/--; S=0,D=n5:1.00/1.00/1.00/1.00/1.00/--/--/--; S=0,D=n6:1.00/1.00/1.00/1.00/1.00/--/--/--; S=0,D=n7:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n4:.80/.40/1.00/.80/.80/--/--/--; S=1,D=n5:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n6:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n7:.50/.00/1.00/.50/.50/--/--/--; S=2-3,D=n5:NA/NA/NA/NA/NA/--/--/--; S=2-3,D=n6:1.00/1.00/1.00/1.00/1.00/--/--/--; S=2-3,D=n7:1.00/1.00/1.00/1.00/1.00/--/--/--	S=0,D=n4:1.00/1.00/1.00/1.00/1.00/--/--/--; S=0,D=n5:1.00/1.00/1.00/1.00/1.00/--/--/--; S=0,D=n6:1.00/1.00/1.00/1.00/1.00/--/--/--; S=0,D=n7:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n4:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n5:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n6:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n7:1.00/1.00/1.00/1.00/1.00/--/--/--; S=2-3,D=n5:1.00/1.00/1.00/1.00/1.00/--/--/--; S=2-3,D=n6:1.00/1.00/1.00/1.00/1.00/--/--/--; S=2-3,D=n7:1.00/1.00/1.00/1.00/1.00/--/--/--	S=0,D=n4:1.00/1.00/1.00/1.00/1.00/--/--/--; S=0,D=n5:1.00/1.00/1.00/1.00/1.00/--/--/--; S=0,D=n6:1.00/1.00/1.00/1.00/1.00/--/--/--; S=0,D=n7:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n4:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n5:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n6:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n7:1.00/1.00/1.00/1.00/1.00/--/--/--; S=2-3,D=n5:1.00/1.00/1.00/1.00/1.00/--/--/--; S=2-3,D=n6:1.00/1.00/1.00/1.00/1.00/--/--/--; S=2-3,D=n7:1.00/1.00/1.00/1.00/1.00/--/--/--	--
gemini-3.1-pro-preview	S=0,D=n4:.00/.00/.00/1.00/.00/--/--/--; S=0,D=n5:.00/.00/.00/1.00/.00/--/--/--; S=0,D=n6:.00/.00/.00/1.00/.00/--/--/--; S=0,D=n7:.00/.00/.00/1.00/.00/--/--/--; S=1,D=n4:.00/.00/.00/1.00/.00/--/--/--; S=1,D=n5:.00/.00/.00/1.00/.00/--/--/--; S=1,D=n6:.00/.00/.00/.67/.00/--/--/--; S=1,D=n7:.00/.00/.00/.20/.00/--/--/--; S=2-3,D=n5:.00/.00/.00/1.00/.00/--/--/--; S=2-3,D=n6:.00/.00/.00/.67/.00/--/--/--; S=2-3,D=n7:.00/.00/.00/.00/.00/--/--/--	S=0,D=n4:.00/.00/.00/1.00/.00/--/--/--; S=0,D=n5:.00/.00/.00/1.00/.00/--/--/--; S=0,D=n6:.00/.00/.00/1.00/.00/--/--/--; S=0,D=n7:.00/.00/.00/1.00/.00/--/--/--; S=1,D=n4:.00/.00/.00/1.00/.00/--/--/--; S=1,D=n5:.00/.00/.00/1.00/.00/--/--/--; S=1,D=n6:.00/.00/.00/1.00/.00/--/--/--; S=1,D=n7:.00/.00/.00/1.00/.00/--/--/--; S=2-3,D=n5:.00/.00/.00/1.00/.00/--/--/--; S=2-3,D=n6:.00/.00/.00/1.00/.00/--/--/--; S=2-3,D=n7:.00/.00/.00/1.00/.00/--/--/--	S=0,D=n4:.00/.00/.00/1.00/.00/--/--/--; S=0,D=n5:.00/.00/.00/1.00/.00/--/--/--; S=0,D=n6:.00/.00/.00/1.00/.00/--/--/--; S=0,D=n7:.00/.00/.00/1.00/.00/--/--/--; S=1,D=n4:.00/.00/.00/1.00/.00/--/--/--; S=1,D=n5:.00/.00/.00/1.00/.00/--/--/--; S=1,D=n6:.00/.00/.00/1.00/.00/--/--/--; S=1,D=n7:.00/.00/.00/1.00/.00/--/--/--; S=2-3,D=n5:.00/.00/.00/1.00/.00/--/--/--; S=2-3,D=n6:.00/.00/.00/1.00/.00/--/--/--; S=2-3,D=n7:.00/.00/.00/1.00/.00/--/--/--	--
gemini-3.8-flash	S=0,D=n4:.33/.00/.67/1.00/.33/--/--/--; S=0,D=n5:.00/.00/.00/1.00/.00/--/--/--; S=0,D=n6:.00/.00/.00/.83/.00/--/--/--; S=0,D=n7:.00/.00/.00/1.00/.00/--/--/--; S=1,D=matched-extension:.38/.24/.53/.97/.38/--/--/--; S=1,D=n4:.17/.00/.50/1.00/.17/--/--/--; S=1,D=n5:.40/.00/.80/1.00/.40/--/--/--; S=1,D=n6:.33/.00/1.00/1.00/.33/--/--/--; S=1,D=n7:.00/.00/.00/1.00/.00/--/--/--; S=16,D=matched-extension:.11/.04/.19/.38/.11/--/--/--; S=2,D=matched-extension:.25/.15/.36/.99/.25/--/--/--; S=2-3,D=n5:.00/.00/.00/1.00/.00/--/--/--; S=2-3,D=n6:.00/.00/.00/1.00/.00/--/--/--; S=2-3,D=n7:.00/.00/.00/1.00/.00/--/--/--; S=32,D=matched-extension:.08/.01/.17/.27/.08/--/--/--; S=4,D=matched-extension:.21/.10/.33/.99/.21/--/--/--; S=8,D=matched-extension:.31/.21/.40/.51/.31/--/--/--	S=0,D=n4:.00/.00/.00/1.00/.00/--/--/--; S=0,D=n5:.00/.00/.00/.83/.00/--/--/--; S=0,D=n6:.17/.00/.50/.83/.17/--/--/--; S=0,D=n7:.00/.00/.00/.83/.00/--/--/--; S=1,D=n4:.17/.00/.50/1.00/.17/--/--/--; S=1,D=n5:.00/.00/.00/1.00/.00/--/--/--; S=1,D=n6:.00/.00/.00/1.00/.00/--/--/--; S=1,D=n7:.00/.00/.00/1.00/.00/--/--/--; S=2-3,D=n5:.00/.00/.00/1.00/.00/--/--/--; S=2-3,D=n6:.00/.00/.00/1.00/.00/--/--/--; S=2-3,D=n7:.00/.00/.00/1.00/.00/--/--/--	S=0,D=n4:.00/.00/.00/.83/.00/--/--/--; S=0,D=n5:.33/.00/.67/.83/.33/--/--/--; S=0,D=n6:.00/.00/.00/.83/.00/--/--/--; S=0,D=n7:.00/.00/.00/.83/.00/--/--/--; S=1,D=n4:.00/.00/.00/1.00/.00/--/--/--; S=1,D=n5:.20/.00/.60/1.00/.20/--/--/--; S=1,D=n6:.00/.00/.00/1.00/.00/--/--/--; S=1,D=n7:.00/.00/.00/.80/.00/--/--/--; S=2-3,D=n5:.00/.00/.00/1.00/.00/--/--/--; S=2-3,D=n6:.33/.00/1.00/1.00/.33/--/--/--; S=2-3,D=n7:.00/.00/.00/.00/.00/--/--/--	S=K32-K1,D=matched-extension:-0.30/-0.46/-0.14/--/--/.01/.03/23

system_id	0S: band,tier=E/CIlo/CIhi/R/Fmt/p/pH/nFam	3S: band,tier=E/CIlo/CIhi/R/Fmt/p/pH/nFam	3S+C: band,tier=E/CIlo/CIhi/R/Fmt/p/pH/nFam	family_contrast: band,tier=E/CIlo/CIhi/R/Fmt/p/pH/nFam
gpt-5.6-luna	S=0,D=n4:1.00/1.00/1.00/1.00/1.00/--/--/--; S=0,D=n5:.83/.50/1.00/.83/1.00/--/--/--; S=0,D=n6:.67/.33/1.00/.67/1.00/--/--/--; S=0,D=n7:.67/.33/1.00/.67/1.00/--/--/--; S=1,D=matched-extension:.69/.56/.82/.69/1.00/--/--/--; S=1,D=n4:.83/.50/1.00/1.00/1.00/--/--/--; S=1,D=n5:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n6:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n7:.60/.20/1.00/1.00/1.00/--/--/--; S=16,D=matched-extension:.76/.64/.88/.76/1.00/--/--/--; S=2,D=matched-extension:.65/.51/.79/.65/1.00/--/--/--; S=2-3,D=n5:.00/.00/.00/.67/1.00/--/--/--; S=2-3,D=n6:1.00/1.00/1.00/1.00/1.00/--/--/--; S=2-3,D=n7:.00/.00/.00/1.00/1.00/--/--/--; S=32,D=matched-extension:.67/.51/.81/.67/1.00/--/--/--; S=4,D=matched-extension:.75/.62/.86/.75/1.00/--/--/--; S=8,D=matched-extension:.68/.56/.81/.68/1.00/--/--/--	S=0,D=n4:.83/.50/1.00/.83/1.00/--/--/--; S=0,D=n5:.83/.50/1.00/.83/1.00/--/--/--; S=0,D=n6:.50/.17/.83/.50/1.00/--/--/--; S=0,D=n7:.67/.33/1.00/.67/1.00/--/--/--; S=1,D=n4:.67/.33/1.00/1.00/1.00/--/--/--; S=1,D=n5:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n6:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n7:.80/.40/1.00/1.00/1.00/--/--/--; S=2-3,D=n5:1.00/1.00/1.00/1.00/1.00/--/--/--; S=2-3,D=n6:.67/.00/1.00/1.00/1.00/--/--/--; S=2-3,D=n7:1.00/1.00/1.00/1.00/1.00/--/--/--	S=0,D=n4:.83/.50/1.00/.83/1.00/--/--/--; S=0,D=n5:1.00/1.00/1.00/1.00/1.00/--/--/--; --; S=0,D=n6:.83/.50/1.00/.83/1.00/--/--/--; S=0,D=n7:.50/.17/.83/.50/1.00/--/--/--; S=1,D=n4:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n5:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n6:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n7:1.00/1.00/1.00/1.00/1.00/--/--/--; S=2-3,D=n5:1.00/1.00/1.00/1.00/1.00/--/--/--; S=2-3,D=n6:.67/.00/1.00/1.00/1.00/--/--/--; S=2-3,D=n7:1.00/1.00/1.00/1.00/1.00/--/--/--	S=K32-K1,D=matched-extension:-0.03/-0.14/.08/--/--/.75/1.00/24
grok-4.6	S=0,D=n4:1.00/1.00/1.00/1.00/1.00/--/--/--; S=0,D=n5:1.00/1.00/1.00/1.00/1.00/--/--/--; S=0,D=n6:1.00/1.00/1.00/1.00/1.00/--/--/--; S=0,D=n7:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n4:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n5:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n6:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n7:.80/.40/1.00/.80/1.00/--/--/--; S=2-3,D=n5:1.00/1.00/1.00/1.00/1.00/--/--/--; S=2-3,D=n6:1.00/1.00/1.00/1.00/1.00/--/--/--; S=2-3,D=n7:1.00/1.00/1.00/1.00/1.00/--/--/--	S=0,D=n4:1.00/1.00/1.00/1.00/1.00/--/--/--; S=0,D=n5:1.00/1.00/1.00/1.00/1.00/--/--/--; S=0,D=n6:1.00/1.00/1.00/1.00/1.00/--/--/--; S=0,D=n7:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n4:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n5:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n6:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n7:1.00/1.00/1.00/1.00/1.00/--/--/--; S=2-3,D=n5:1.00/1.00/1.00/1.00/1.00/--/--/--; S=2-3,D=n6:1.00/1.00/1.00/1.00/1.00/--/--/--; S=2-3,D=n7:1.00/1.00/1.00/1.00/1.00/--/--/--	S=0,D=n4:1.00/1.00/1.00/1.00/1.00/--/--/--; S=0,D=n5:1.00/1.00/1.00/1.00/1.00/--/--/--; S=0,D=n6:1.00/1.00/1.00/1.00/1.00/--/--/--; S=0,D=n7:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n4:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n5:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n6:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n7:1.00/1.00/1.00/1.00/1.00/--/--/--; S=2-3,D=n5:1.00/1.00/1.00/1.00/1.00/--/--/--; S=2-3,D=n6:1.00/1.00/1.00/1.00/1.00/--/--/--; S=2-3,D=n7:1.00/1.00/1.00/1.00/1.00/--/--/--	--
mistral-large	S=0,D=n4:.00/.00/.00/.00/.00/--/--/--; S=0,D=n5:.00/.00/.00/.00/.00/--/--/--; S=0,D=n6:.00/.00/.00/.00/.00/--/--/--; S=0,D=n7:.00/.00/.00/.00/.00/--/--/--; S=1,D=n4:.00/.00/.00/.17/.00/--/--/--; S=1,D=n5:.00/.00/.00/.40/.00/--/--/--; S=1,D=n6:.00/.00/.00/.33/.00/--/--/--; S=1,D=n7:.00/.00/.00/.40/.00/--/--/--; S=2-3,D=n5:.00/.00/.00/.00/.00/--/--/--; S=2-3,D=n6:.00/.00/.00/.33/.00/--/--/--; S=2-3,D=n7:.00/.00/.00/1.00/.00/--/--/--	S=0,D=n4:.00/.00/.00/.00/.00/--/--/--; S=0,D=n5:.00/.00/.00/.00/.00/--/--/--; S=0,D=n6:.00/.00/.00/.00/.00/--/--/--; S=0,D=n7:.00/.00/.00/.00/.00/--/--/--; S=1,D=n4:.00/.00/.00/.33/.00/--/--/--; S=1,D=n5:.00/.00/.00/.20/.00/--/--/--; S=1,D=n6:.00/.00/.00/.33/.00/--/--/--; S=1,D=n7:.00/.00/.00/.20/.00/--/--/--; S=2-3,D=n5:.00/.00/.00/1.00/.00/--/--/--; S=2-3,D=n6:.00/.00/.00/.44/.00/--/--/--; S=2-3,D=n7:.00/.00/.00/.00/.00/--/--/--	S=0,D=n4:.00/.00/.00/.00/.00/--/--/--; S=0,D=n5:.00/.00/.00/.00/.00/--/--/--; S=0,D=n6:.00/.00/.00/.00/.00/--/--/--; S=0,D=n7:.00/.00/.00/.00/.00/--/--/--; S=1,D=n4:.00/.00/.00/.50/.00/--/--/--; S=1,D=n5:.00/.00/.00/.00/.00/--/--/--; S=1,D=n6:.00/.00/.00/.33/.00/--/--/--; S=1,D=n7:.00/.00/.00/.00/.00/--/--/--; S=2-3,D=n5:.00/.00/.00/1.00/.00/--/--/--; S=2-3,D=n6:.00/.00/.00/.33/.00/--/--/--; S=2-3,D=n7:.00/.00/.00/.00/.00/--/--/--	--

system_id	0S: band,tier=E/CIlo/CIhi/R/Fmt/p/pH/nFam	3S: band,tier=E/CIlo/CIhi/R/Fmt/p/pH/nFam	3S+C: band,tier=E/CIlo/CIhi/R/Fmt/p/pH/nFam	family_contrast: band,tier=E/CIlo/CIhi/R/Fmt/p/pH/nFam
muse-spark-1.3	S=0,D=n4:1.00/1.00/1.00/1.00/1.00/--/--/--; S=0,D=n5:1.00/1.00/1.00/1.00/1.00/--/--/--; S=0,D=n6:1.00/1.00/1.00/1.00/1.00/--/--/--; S=0,D=n7:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n4:.83/.50/1.00/.83/.83/--/--/--; S=1,D=n5:1.00/1.00/1.00/1.00/--/--/--; S=1,D=n6:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n7:1.00/1.00/1.00/1.00/1.00/--/--/--; S=2-3,D=n5:1.00/1.00/1.00/1.00/1.00/--/--/--; S=2-3,D=n6:1.00/1.00/1.00/1.00/1.00/--/--/--; S=2-3,D=n7:1.00/1.00/1.00/1.00/1.00/--/--/--	S=0,D=n4:1.00/1.00/1.00/1.00/1.00/--/--/--; S=0,D=n5:1.00/1.00/1.00/1.00/1.00/--/--/--; S=0,D=n6:1.00/1.00/1.00/1.00/1.00/--/--/--; S=0,D=n7:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n4:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n5:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n6:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n7:1.00/1.00/1.00/1.00/1.00/--/--/--; S=2-3,D=n5:1.00/1.00/1.00/1.00/1.00/--/--/--; S=2-3,D=n6:1.00/1.00/1.00/1.00/1.00/--/--/--; S=2-3,D=n7:1.00/1.00/1.00/1.00/1.00/--/--/--	S=0,D=n4:1.00/1.00/1.00/1.00/1.00/--/--/--; S=0,D=n5:1.00/1.00/1.00/1.00/1.00/--/--/--; S=0,D=n6:1.00/1.00/1.00/1.00/1.00/--/--/--; S=0,D=n7:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n4:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n5:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n6:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n7:1.00/1.00/1.00/1.00/1.00/--/--/--; S=2-3,D=n5:1.00/1.00/1.00/1.00/1.00/--/--/--; S=2-3,D=n6:1.00/1.00/1.00/1.00/1.00/--/--/--; S=2-3,D=n7:1.00/1.00/1.00/1.00/1.00/--/--/--	--
phi-4	S=0,D=n4:.00/.00/.00/.00/.00/--/--/--; S=0,D=n5:.00/.00/.00/.00/.00/--/--/--; S=0,D=n6:.00/.00/.00/.00/.00/--/--/--; S=0,D=n7:.00/.00/.00/.00/.00/--/--/--; S=1,D=matched-extension:.00/.00/.00/.00/.00/--/--/--; S=1,D=n4:.00/.00/.00/.67/.00/--/--/--; S=1,D=n5:.00/.00/.00/.40/.00/--/--/--; S=1,D=n6:.00/.00/.00/.67/.00/--/--/--; S=1,D=n7:.00/.00/.00/.40/.00/--/--/--; S=16,D=matched-extension:.00/.00/.00/.01/.00/--/--/--; S=2,D=matched-extension:.00/.00/.00/.00/.00/--/--/--; S=2-3,D=n5:.00/.00/.00/1.00/.00/--/--/--; S=2-3,D=n6:.00/.00/.00/.33/.00/--/--/--; S=2-3,D=n7:.00/.00/.00/.00/.00/--/--/--; S=32,D=matched-extension:.00/.00/.00/.03/.00/--/--/--; S=4,D=matched-extension:.00/.00/.00/.00/.00/--/--/--; S=8,D=matched-extension:.00/.00/.00/.00/.00/--/--/--	S=0,D=n4:.00/.00/.00/.00/.00/--/--/--; S=0,D=n5:.00/.00/.00/.00/.00/--/--/--; S=0,D=n6:.00/.00/.00/.00/.00/--/--/--; S=0,D=n7:.00/.00/.00/.00/.00/--/--/--; S=1,D=n4:.00/.00/.00/.17/.00/--/--/--; S=1,D=n5:.00/.00/.00/.20/.00/--/--/--; S=1,D=n6:.00/.00/.00/.33/.00/--/--/--; S=1,D=n7:.00/.00/.00/.00/.00/--/--/--; S=2-3,D=n5:.00/.00/.00/.00/.00/--/--/--; S=2-3,D=n6:.00/.00/.00/.00/.00/--/--/--; S=2-3,D=n7:.00/.00/.00/.00/.00/--/--/--	S=0,D=n4:.00/.00/.00/.00/.00/--/--/--; S=0,D=n5:.00/.00/.00/.00/.00/--/--/--; S=0,D=n6:.00/.00/.00/.00/.00/--/--/--; S=0,D=n7:.00/.00/.00/.00/.00/--/--/--; S=1,D=n4:.00/.00/.00/.17/.00/--/--/--; S=1,D=n5:.00/.00/.00/.00/.00/--/--/--; S=1,D=n6:.00/.00/.00/.00/.00/--/--/--; S=1,D=n7:.00/.00/.00/.60/.00/--/--/--; S=2-3,D=n5:.00/.00/.00/.00/.00/--/--/--; S=2-3,D=n6:.00/.00/.00/.50/.00/--/--/--; S=2-3,D=n7:.00/.00/.00/.00/.00/--/--/--	S=K32-K1,D=matched-extension:.00/.00/.00/--/--/1.00/1.00/14
qwen3.8-a95b	S=0,D=n4:1.00/1.00/1.00/1.00/1.00/--/--/--; S=0,D=n5:1.00/1.00/1.00/1.00/1.00/--/--/--; S=0,D=n6:1.00/1.00/1.00/1.00/1.00/--/--/--; S=0,D=n7:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n4:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n5:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n6:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n7:1.00/1.00/1.00/1.00/1.00/--/--/--; S=2-3,D=n5:1.00/1.00/1.00/1.00/1.00/--/--/--; S=2-3,D=n6:1.00/1.00/1.00/1.00/1.00/--/--/--; S=2-3,D=n7:1.00/1.00/1.00/1.00/1.00/--/--/--	S=0,D=n4:1.00/1.00/1.00/1.00/1.00/--/--/--; S=0,D=n5:1.00/1.00/1.00/1.00/1.00/--/--/--; S=0,D=n6:1.00/1.00/1.00/1.00/1.00/--/--/--; S=0,D=n7:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n4:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n5:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n6:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n7:1.00/1.00/1.00/1.00/1.00/--/--/--; S=2-3,D=n5:1.00/1.00/1.00/1.00/1.00/--/--/--; S=2-3,D=n6:1.00/1.00/1.00/1.00/1.00/--/--/--; S=2-3,D=n7:1.00/1.00/1.00/1.00/1.00/--/--/--	S=0,D=n4:1.00/1.00/1.00/1.00/1.00/--/--/--; S=0,D=n5:1.00/1.00/1.00/1.00/1.00/--/--/--; S=0,D=n6:1.00/1.00/1.00/1.00/1.00/--/--/--; S=0,D=n7:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n4:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n5:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n6:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n7:1.00/1.00/1.00/1.00/1.00/--/--/--; S=2-3,D=n5:1.00/1.00/1.00/1.00/1.00/--/--/--; S=2-3,D=n6:1.00/1.00/1.00/1.00/1.00/--/--/--; S=2-3,D=n7:1.00/1.00/1.00/1.00/1.00/--/--/--	--

system_id	0S: band,tier=E/CIlo/CIhi/R/Fmt/p/pH/nFam	3S: band,tier=E/CIlo/CIhi/R/Fmt/p/pH/nFam	3S+C: band,tier=E/CIlo/CIhi/R/Fmt/p/pH/nFam	family_contrast: band,tier=E/CIlo/CIhi/R/Fmt/p/pH/nFam
qwen3.8-max	S=0,D=n4:1.00/1.00/1.00/1.00/1.00/--/--/--; S=0,D=n5:1.00/1.00/1.00/1.00/1.00/--/--/--; S=0,D=n6:1.00/1.00/1.00/1.00/1.00/--/--/--; S=0,D=n7:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n4:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n5:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n6:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n7:1.00/1.00/1.00/1.00/1.00/--/--/--; S=2-3,D=n5:1.00/1.00/1.00/1.00/1.00/--/--/--; S=2-3,D=n6:1.00/1.00/1.00/1.00/1.00/--/--/--; S=2-3,D=n7:1.00/1.00/1.00/1.00/1.00/--/--/--	S=0,D=n4:1.00/1.00/1.00/1.00/1.00/--/--/--; S=0,D=n5:.83/.50/1.00/.83/1.00/--/--/--; S=0,D=n6:1.00/1.00/1.00/1.00/1.00/--/--/--; S=0,D=n7:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n4:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n5:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n6:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n7:1.00/1.00/1.00/1.00/1.00/--/--/--; S=2-3,D=n5:1.00/1.00/1.00/1.00/1.00/--/--/--; S=2-3,D=n6:1.00/1.00/1.00/1.00/1.00/--/--/--; S=2-3,D=n7:1.00/1.00/1.00/1.00/1.00/--/--/--	S=0,D=n4:1.00/1.00/1.00/1.00/1.00/--/--/--; S=0,D=n5:1.00/1.00/1.00/1.00/1.00/--/--/--; S=0,D=n6:1.00/1.00/1.00/1.00/1.00/--/--/--; S=0,D=n7:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n4:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n5:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n6:1.00/1.00/1.00/1.00/1.00/--/--/--; S=1,D=n7:1.00/1.00/1.00/1.00/1.00/--/--/--; S=2-3,D=n5:1.00/1.00/1.00/1.00/1.00/--/--/--; S=2-3,D=n6:1.00/1.00/1.00/1.00/1.00/--/--/--; S=2-3,D=n7:1.00/1.00/1.00/1.00/1.00/--/--/--	--

Table 13: Enumeration by clause complexity. Cells report E/P/R; F1 is omitted because it is not defined in this aggregate table. Full machine-readable values are supplied as CSV.

system_id	0S: band,tier=E/P/R	3S: band,tier=E/P/R	3S+C: band,tier=E/P/R
claude-sonnet-5	C=high,D=n4:.00/1.00/1.00; C=high,D=n5:.00/1.00/1.00; C=high,D=n6:.00/1.00/1.00; C=high,D=n7:.00/1.00/1.00; C=low,D=n4:.14/1.00/1.00; C=low,D=n5:.00/1.00/1.00; C=low,D=n6:.00/1.00/1.00; C=low,D=n7:.00/1.00/1.00; C=mid,D=n4:.00/1.00/1.00; C=mid,D=n5:.00/1.00/1.00; C=mid,D=n6:.00/1.00/1.00; C=mid,D=n7:.00/.67/.67	C=high,D=n4:.00/1.00/1.00; C=high,D=n5:.00/1.00/1.00; C=high,D=n6:.00/1.00/1.00; C=high,D=n7:.00/1.00/1.00; C=low,D=n4:.71/1.00/1.00; C=low,D=n5:.17/1.00/1.00; C=low,D=n6:.00/.67/.67; C=low,D=n7:.00/1.00/1.00; C=mid,D=n4:.33/1.00/1.00; C=mid,D=n5:.00/1.00/1.00; C=mid,D=n6:.00/.67/.67; C=mid,D=n7:.00/1.00/1.00	C=high,D=n4:.00/1.00/1.00; C=high,D=n5:NA/NA/NA; C=high,D=n6:.00/1.00/1.00; C=high,D=n7:.00/1.00/1.00; C=low,D=n4:.14/1.00/1.00; C=low,D=n5:.17/1.00/1.00; C=low,D=n6:.00/1.00/1.00; C=low,D=n7:.00/1.00/1.00; C=mid,D=n4:.25/1.00/1.00; C=mid,D=n5:.00/1.00/1.00; C=mid,D=n6:.00/1.00/1.00; C=mid,D=n7:.00/1.00/1.00
deepseek-v4-flash-high	C=high,D=n4:1.00/1.00/1.00; C=high,D=n5:1.00/1.00/1.00; C=high,D=n6:1.00/1.00/1.00; C=high,D=n7:1.00/1.00/1.00; C=low,D=n4:.86/.86/.86; C=low,D=n5:1.00/1.00/1.00; C=low,D=n6:1.00/1.00/1.00; C=low,D=n7:1.00/1.00/1.00; C=mid,D=n4:1.00/1.00/1.00; C=mid,D=n5:.67/.67/.67; C=mid,D=n6:.67/1.00/.78; C=mid,D=n7:1.00/1.00/1.00	C=high,D=n4:1.00/1.00/1.00; C=high,D=n5:1.00/1.00/1.00; C=high,D=n6:1.00/1.00/1.00; C=high,D=n7:.80/.80/.80; C=low,D=n4:1.00/1.00/1.00; C=low,D=n5:1.00/1.00/1.00; C=low,D=n6:1.00/1.00/1.00; C=low,D=n7:1.00/1.00/1.00; C=mid,D=n4:1.00/1.00/1.00; C=mid,D=n5:1.00/1.00/1.00; C=mid,D=n6:1.00/1.00/1.00; C=mid,D=n7:1.00/1.00/1.00	C=high,D=n4:1.00/1.00/1.00; C=high,D=n5:1.00/1.00/1.00; C=high,D=n6:1.00/1.00/1.00; C=high,D=n7:1.00/1.00/1.00; C=low,D=n4:1.00/1.00/1.00; C=low,D=n5:1.00/1.00/1.00; C=low,D=n6:1.00/1.00/1.00; C=low,D=n7:1.00/1.00/1.00; C=mid,D=n4:1.00/1.00/1.00; C=mid,D=n5:1.00/1.00/1.00; C=mid,D=n6:1.00/1.00/1.00; C=mid,D=n7:1.00/1.00/1.00
deepseek-v4-pro-high	C=high,D=n4:1.00/1.00/1.00; C=high,D=n5:NA/NA/NA; C=high,D=n6:1.00/1.00/1.00; C=high,D=n7:.80/.80/.80; C=low,D=n4:.80/.80/.80; C=low,D=n5:1.00/1.00/1.00; C=low,D=n6:1.00/1.00/1.00; C=low,D=n7:1.00/1.00/1.00; C=mid,D=n4:1.00/1.00/1.00; C=mid,D=n5:1.00/1.00/1.00; C=mid,D=n6:1.00/1.00/1.00; C=mid,D=n7:1.00/1.00/1.00	C=high,D=n4:1.00/1.00/1.00; C=high,D=n5:1.00/1.00/1.00; C=high,D=n6:1.00/1.00/1.00; C=high,D=n7:1.00/1.00/1.00; C=low,D=n4:1.00/1.00/1.00; C=low,D=n5:1.00/1.00/1.00; C=low,D=n6:1.00/1.00/1.00; C=low,D=n7:1.00/1.00/1.00; C=mid,D=n4:1.00/1.00/1.00; C=mid,D=n5:1.00/1.00/1.00; C=mid,D=n6:1.00/1.00/1.00; C=mid,D=n7:1.00/1.00/1.00	C=high,D=n4:1.00/1.00/1.00; C=high,D=n5:1.00/1.00/1.00; C=high,D=n6:1.00/1.00/1.00; C=high,D=n7:1.00/1.00/1.00; C=low,D=n4:1.00/1.00/1.00; C=low,D=n5:1.00/1.00/1.00; C=low,D=n6:1.00/1.00/1.00; C=low,D=n7:1.00/1.00/1.00; C=mid,D=n4:1.00/1.00/1.00; C=mid,D=n5:1.00/1.00/1.00; C=mid,D=n6:1.00/1.00/1.00; C=mid,D=n7:1.00/1.00/1.00
gemini-3.1-pro-preview	C=high,D=n4:.00/1.00/1.00; C=high,D=n5:.00/1.00/1.00; C=high,D=n6:.00/.75/.75; C=high,D=n7:.00/.67/.67; C=low,D=n4:.00/1.00/1.00; C=low,D=n5:.00/1.00/1.00; C=low,D=n6:.00/1.00/1.00; C=low,D=n7:.00/1.00/1.00; C=mid,D=n4:.00/1.00/1.00; C=mid,D=n5:.00/1.00/1.00; C=mid,D=n6:.00/.80/.80; C=mid,D=n7:.00/.40/.40	C=high,D=n4:.00/1.00/1.00; C=high,D=n5:.00/1.00/1.00; C=high,D=n6:.00/1.00/1.00; C=high,D=n7:.00/1.00/1.00; C=low,D=n4:.00/1.00/1.00; C=low,D=n5:.00/1.00/1.00; C=low,D=n6:.00/1.00/1.00; C=low,D=n7:.00/1.00/1.00; C=mid,D=n4:.00/1.00/1.00; C=mid,D=n5:.00/1.00/1.00; C=mid,D=n6:.00/1.00/1.00; C=mid,D=n7:.00/1.00/1.00	C=high,D=n4:.00/1.00/1.00; C=high,D=n5:.00/1.00/1.00; C=high,D=n6:.00/1.00/1.00; C=high,D=n7:.00/1.00/1.00; C=low,D=n4:.00/1.00/1.00; C=low,D=n5:.00/1.00/1.00; C=low,D=n6:.00/1.00/1.00; C=low,D=n7:.00/1.00/1.00; C=mid,D=n4:.00/1.00/1.00; C=mid,D=n5:.00/1.00/1.00; C=mid,D=n6:.00/1.00/1.00; C=mid,D=n7:.00/1.00/1.00
gemini-3.8-flash	C=high,D=n4:.00/1.00/1.00; C=high,D=n5:.50/1.00/1.00; C=high,D=n6:.00/.75/.75; C=high,D=n7:.00/1.00/1.00; C=low,D=n4:.29/1.00/1.00; C=low,D=n5:.00/1.00/1.00; C=low,D=n6:.33/1.00/1.00; C=low,D=n7:.00/1.00/1.00; C=mid,D=n4:.25/1.00/1.00; C=mid,D=n5:.25/1.00/1.00; C=mid,D=n6:.00/1.00/1.00; C=mid,D=n7:.00/1.00/1.00	C=high,D=n4:.00/1.00/1.00; C=high,D=n5:.00/1.00/1.00; C=high,D=n6:.00/.75/.75; C=high,D=n7:.00/.83/.83; C=low,D=n4:.14/1.00/1.00; C=low,D=n5:.00/1.00/1.00; C=low,D=n6:.00/1.00/1.00; C=low,D=n7:.00/1.00/1.00; C=mid,D=n4:.00/1.00/1.00; C=mid,D=n5:.00/.75/.75; C=mid,D=n6:.20/1.00/1.00; C=mid,D=n7:.00/1.00/1.00	C=high,D=n4:.00/1.00/1.00; C=high,D=n5:.00/1.00/1.00; C=high,D=n6:.00/.75/.75; C=high,D=n7:.00/.83/.83; C=low,D=n4:.00/1.00/1.00; C=low,D=n5:.50/1.00/1.00; C=low,D=n6:.00/1.00/1.00; C=low,D=n7:.00/.00/.00; C=mid,D=n4:.00/.75/.75; C=mid,D=n5:.00/.75/.75; C=mid,D=n6:.20/1.00/1.00; C=mid,D=n7:.00/.80/.80

system_id	0S: band,tier=E/P/R	3S: band,tier=E/P/R	3S+C: band,tier=E/P/R
gpt-5.6-luna	C=high,D=n4:1.00/1.00/1.00; C=high,D=n5:1.00/1.00/1.00; C=high,D=n6:.75/.75/.75; C=high,D=n7:.83/.83/.83; C=low,D=n4:1.00/1.00/1.00; C=low,D=n5:.67/.78/.78; C=low,D=n6:1.00/1.00/1.00; C=low,D=n7:.00/.00/.00; C=mid,D=n4:.75/.88/1.00; C=mid,D=n5:1.00/1.00/1.00; C=mid,D=n6:.80/.80/.80; C=mid,D=n7:.40/.70/1.00	C=high,D=n4:.00/.00/.00; C=high,D=n5:1.00/1.00/1.00; C=high,D=n6:.50/.50/.50; C=high,D=n7:.83/.83/.83; C=low,D=n4:.86/.93/1.00; C=low,D=n5:1.00/1.00/1.00; C=low,D=n6:.67/.89/1.00; C=low,D=n7:1.00/1.00/1.00; C=mid,D=n4:.75/.83/1.00; C=mid,D=n5:.75/.75/.75; C=mid,D=n6:.80/.80/.80; C=mid,D=n7:.60/.67/.80	C=high,D=n4:1.00/1.00/1.00; C=high,D=n5:1.00/1.00/1.00; C=high,D=n6:.75/.75/.75; C=high,D=n7:.67/.67/.67; C=low,D=n4:.86/.86/.86; C=low,D=n5:1.00/1.00/1.00; C=low,D=n6:.67/.89/1.00; C=low,D=n7:.00/.00/.00; C=mid,D=n4:1.00/1.00/1.00; C=mid,D=n5:1.00/1.00/1.00; C=mid,D=n6:1.00/1.00/1.00; C=mid,D=n7:1.00/1.00/1.00
grok-4.6	C=high,D=n4:1.00/1.00/1.00; C=high,D=n5:1.00/1.00/1.00; C=high,D=n6:1.00/1.00/1.00; C=high,D=n7:1.00/1.00/1.00; C=low,D=n4:1.00/1.00/1.00; C=low,D=n5:1.00/1.00/1.00; C=low,D=n6:1.00/1.00/1.00; C=low,D=n7:1.00/1.00/1.00; C=mid,D=n4:1.00/1.00/1.00; C=mid,D=n5:1.00/1.00/1.00; C=mid,D=n6:1.00/1.00/1.00; C=mid,D=n7:.80/.80/.80	C=high,D=n4:1.00/1.00/1.00; C=high,D=n5:1.00/1.00/1.00; C=high,D=n6:1.00/1.00/1.00; C=high,D=n7:1.00/1.00/1.00; C=low,D=n4:1.00/1.00/1.00; C=low,D=n5:1.00/1.00/1.00; C=low,D=n6:1.00/1.00/1.00; C=low,D=n7:1.00/1.00/1.00; C=mid,D=n4:1.00/1.00/1.00; C=mid,D=n5:1.00/1.00/1.00; C=mid,D=n6:1.00/1.00/1.00; C=mid,D=n7:1.00/1.00/1.00	C=high,D=n4:1.00/1.00/1.00; C=high,D=n5:1.00/1.00/1.00; C=high,D=n6:1.00/1.00/1.00; C=high,D=n7:1.00/1.00/1.00; C=low,D=n4:1.00/1.00/1.00; C=low,D=n5:1.00/1.00/1.00; C=low,D=n6:1.00/1.00/1.00; C=low,D=n7:1.00/1.00/1.00; C=mid,D=n4:1.00/1.00/1.00; C=mid,D=n5:1.00/1.00/1.00; C=mid,D=n6:1.00/1.00/1.00; C=mid,D=n7:1.00/1.00/1.00
mistral-large	C=high,D=n4:.00/.00/.00; C=high,D=n5:.00/.00/.00; C=high,D=n6:.00/.02/.25; C=high,D=n7:.00/.02/.33; C=low,D=n4:.00/.02/.14; C=low,D=n5:.00/.00/.00; C=low,D=n6:.00/.00/.00; C=low,D=n7:.00/.00/.00; C=mid,D=n4:.00/.00/.00; C=mid,D=n5:.00/.05/.50; C=mid,D=n6:.00/.02/.20; C=mid,D=n7:.00/.02/.20	C=high,D=n4:.00/.00/.00; C=high,D=n5:.00/.04/.50; C=high,D=n6:.00/.25/.25; C=high,D=n7:.00/.03/.17; C=low,D=n4:.00/.10/.29; C=low,D=n5:.00/.08/.17; C=low,D=n6:.00/.08/.17; C=low,D=n7:.00/.00/.00; C=mid,D=n4:.00/.00/.00; C=mid,D=n5:.00/.00/.00; C=mid,D=n6:.00/.03/.17; C=mid,D=n7:.00/.00/.00	C=high,D=n4:.00/.00/.00; C=high,D=n5:.00/.00/.00; C=high,D=n6:.00/.04/.25; C=high,D=n7:.00/.00/.00; C=low,D=n4:.00/.07/.43; C=low,D=n5:.00/.07/.17; C=low,D=n6:.00/.07/.17; C=low,D=n7:.00/.00/.00; C=mid,D=n4:.00/.00/.00; C=mid,D=n5:.00/.00/.00; C=mid,D=n6:.00/.02/.10; C=mid,D=n7:.00/.00/.00
muse-spark-1.3	C=high,D=n4:1.00/1.00/1.00; C=high,D=n5:1.00/1.00/1.00; C=high,D=n6:1.00/1.00/1.00; C=high,D=n7:1.00/1.00/1.00; C=low,D=n4:.86/.86/.86; C=low,D=n5:1.00/1.00/1.00; C=low,D=n6:1.00/1.00/1.00; C=low,D=n7:1.00/1.00/1.00; C=mid,D=n4:1.00/1.00/1.00; C=mid,D=n5:1.00/1.00/1.00; C=mid,D=n6:1.00/1.00/1.00; C=mid,D=n7:1.00/1.00/1.00	C=high,D=n4:1.00/1.00/1.00; C=high,D=n5:1.00/1.00/1.00; C=high,D=n6:1.00/1.00/1.00; C=high,D=n7:1.00/1.00/1.00; C=low,D=n4:1.00/1.00/1.00; C=low,D=n5:1.00/1.00/1.00; C=low,D=n6:1.00/1.00/1.00; C=low,D=n7:1.00/1.00/1.00; C=mid,D=n4:1.00/1.00/1.00; C=mid,D=n5:1.00/1.00/1.00; C=mid,D=n6:1.00/1.00/1.00; C=mid,D=n7:1.00/1.00/1.00	C=high,D=n4:1.00/1.00/1.00; C=high,D=n5:1.00/1.00/1.00; C=high,D=n6:1.00/1.00/1.00; C=high,D=n7:1.00/1.00/1.00; C=low,D=n4:1.00/1.00/1.00; C=low,D=n5:1.00/1.00/1.00; C=low,D=n6:1.00/1.00/1.00; C=low,D=n7:1.00/1.00/1.00; C=mid,D=n4:1.00/1.00/1.00; C=mid,D=n5:1.00/1.00/1.00; C=mid,D=n6:1.00/1.00/1.00; C=mid,D=n7:1.00/1.00/1.00
phi-4	C=high,D=n4:.00/.00/.00; C=high,D=n5:.00/.00/.00; C=high,D=n6:.00/.00/.00; C=high,D=n7:.00/.00/.17; C=low,D=n4:.00/.04/.29; C=low,D=n5:.00/.04/.33; C=low,D=n6:.00/.02/.67; C=low,D=n7:.00/.00/.00; C=mid,D=n4:.00/.05/.50; C=mid,D=n5:.00/.05/.25; C=mid,D=n6:.00/.01/.20; C=mid,D=n7:.00/.00/.20	C=high,D=n4:.00/.00/.00; C=high,D=n5:.00/.00/.00; C=high,D=n6:.00/.00/.00; C=high,D=n7:.00/.00/.00; C=low,D=n4:.00/.03/.14; C=low,D=n5:.00/.00/.00; C=low,D=n6:.00/.11/.33; C=low,D=n7:.00/.00/.00; C=mid,D=n4:.00/.00/.00; C=mid,D=n5:.00/.06/.25; C=mid,D=n6:.00/.00/.00; C=mid,D=n7:.00/.00/.00	C=high,D=n4:.00/.00/.00; C=high,D=n5:.00/.00/.00; C=high,D=n6:.00/.00/.00; C=high,D=n7:.00/.01/.50; C=low,D=n4:.00/.04/.14; C=low,D=n5:.00/.00/.00; C=low,D=n6:.00/.04/.33; C=low,D=n7:.00/.00/.00; C=mid,D=n4:.00/.00/.00; C=mid,D=n5:.00/.00/.00; C=mid,D=n6:.00/.02/.10; C=mid,D=n7:.00/.00/.00
qwen3.8-a95b	C=high,D=n4:NA/NA/NA; C=high,D=n5:1.00/1.00/1.00; C=high,D=n6:1.00/1.00/1.00; C=high,D=n7:1.00/1.00/1.00; C=low,D=n4:1.00/1.00/1.00; C=low,D=n5:1.00/1.00/1.00; C=low,D=n6:1.00/1.00/1.00; C=low,D=n7:1.00/1.00/1.00; C=mid,D=n4:1.00/1.00/1.00; C=mid,D=n5:1.00/1.00/1.00; C=mid,D=n6:1.00/1.00/1.00; C=mid,D=n7:1.00/1.00/1.00	C=high,D=n4:1.00/1.00/1.00; C=high,D=n5:1.00/1.00/1.00; C=high,D=n6:1.00/1.00/1.00; C=high,D=n7:1.00/1.00/1.00; C=low,D=n4:1.00/1.00/1.00; C=low,D=n5:1.00/1.00/1.00; C=low,D=n6:1.00/1.00/1.00; C=low,D=n7:1.00/1.00/1.00; C=mid,D=n4:1.00/1.00/1.00; C=mid,D=n5:1.00/1.00/1.00; C=mid,D=n6:1.00/1.00/1.00; C=mid,D=n7:1.00/1.00/1.00	C=high,D=n4:NA/NA/NA; C=high,D=n5:1.00/1.00/1.00; C=high,D=n6:1.00/1.00/1.00; C=high,D=n7:1.00/1.00/1.00; C=low,D=n4:1.00/1.00/1.00; C=low,D=n5:1.00/1.00/1.00; C=low,D=n6:1.00/1.00/1.00; C=low,D=n7:1.00/1.00/1.00; C=mid,D=n4:1.00/1.00/1.00; C=mid,D=n5:1.00/1.00/1.00; C=mid,D=n6:1.00/1.00/1.00; C=mid,D=n7:1.00/1.00/1.00

system_id	0S: band,tier=E/P/R	3S: band,tier=E/P/R	3S+C: band,tier=E/P/R
qwen3.8-max	C=high,D=n4:1.00/1.00/1.00; C=high,D=n5:1.00/1.00/1.00; C=high,D=n6:1.00/1.00/1.00; C=high,D=n7:1.00/1.00/1.00; C=low,D=n4:1.00/1.00/1.00; C=low,D=n5:1.00/1.00/1.00; C=low,D=n6:1.00/1.00/1.00; C=low,D=n7:1.00/1.00/1.00; C=mid,D=n4:1.00/1.00/1.00; C=mid,D=n5:1.00/1.00/1.00; C=mid,D=n6:1.00/1.00/1.00; C=mid,D=n7:1.00/1.00/1.00	C=high,D=n4:1.00/1.00/1.00; C=high,D=n5:1.00/1.00/1.00; C=high,D=n6:1.00/1.00/1.00; C=high,D=n7:1.00/1.00/1.00; C=low,D=n4:1.00/1.00/1.00; C=low,D=n5:.83/.83/.83; C=low,D=n6:1.00/1.00/1.00; C=low,D=n7:1.00/1.00/1.00; C=mid,D=n4:1.00/1.00/1.00; C=mid,D=n5:1.00/1.00/1.00; C=mid,D=n6:1.00/1.00/1.00; C=mid,D=n7:1.00/1.00/1.00	C=high,D=n4:NA/NA/NA; C=high,D=n5:1.00/1.00/1.00; C=high,D=n6:1.00/1.00/1.00; C=high,D=n7:1.00/1.00/1.00; C=low,D=n4:1.00/1.00/1.00; C=low,D=n5:1.00/1.00/1.00; C=low,D=n6:1.00/1.00/1.00; C=low,D=n7:1.00/1.00/1.00; C=mid,D=n4:1.00/1.00/1.00; C=mid,D=n5:1.00/1.00/1.00; C=mid,D=n6:1.00/1.00/1.00; C=mid,D=n7:1.00/1.00/1.00

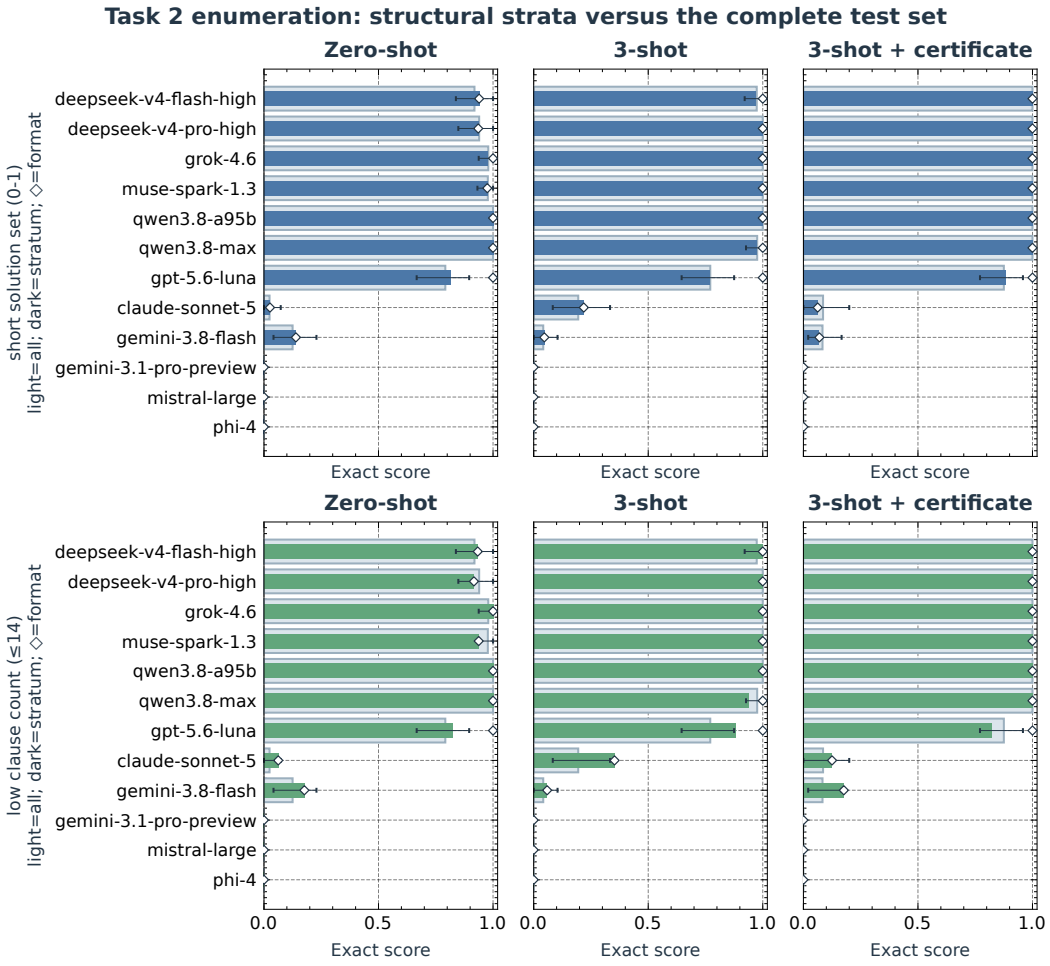

Figure 6: Enumeration exactness for the full set and prespecified solution-cardinality and clause-count strata. Diamonds show format validity.

E Complete View and Correlation Results

Table 14: Renaming-only invariance. Full machine-readable values are supplied as CSV.

system_id | task | prompt_mode	B tier:E/F	N tier:E/F
deepseek-v4-flash-high | D | 3S	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:.75/1.00	n4:.90/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00
deepseek-v4-flash-high | D | 3S+C	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:.91/1.00; n5:1.00/1.00; n6:.88/1.00; n7:1.00/1.00
deepseek-v4-flash-high | D | 0S	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00
deepseek-v4-flash-high | E | 3S	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:.89/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00
deepseek-v4-flash-high | E | 3S+C	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:.86/1.00; n7:1.00/1.00
deepseek-v4-flash-high | E | 0S	n4:.91/.91; n5:.90/.90; n6:.88/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00
deepseek-v4-flash-high | R | 3S	n4:.78/1.00; n5:.67/1.00; n6:1.00/1.00; n7:.75/1.00	n4:.75/1.00; n5:1.00/1.00; n6:.50/1.00; n7:1.00/1.00
deepseek-v4-flash-high | R | 3S+C	n4:.78/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00
deepseek-v4-flash-high | R | 0S	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:.83/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00
deepseek-v4-flash-none | D | 3S	n4:.33/1.00; n5:.42/1.00; n6:.50/1.00; n7:.42/1.00	n4:.58/1.00; n5:.50/1.00; n6:.67/1.00; n7:.50/1.00
deepseek-v4-flash-none | D | 3S+C	n4:.50/1.00; n5:.50/1.00; n6:.67/1.00; n7:.50/1.00	n4:.42/1.00; n5:.50/1.00; n6:.42/1.00; n7:.50/1.00
deepseek-v4-flash-none | D | 0S	n4:.50/1.00; n5:.42/1.00; n6:.58/1.00; n7:.58/1.00	n4:.50/1.00; n5:.42/1.00; n6:.67/1.00; n7:.50/1.00
deepseek-v4-flash-none | E | 3S	n4:.00/1.00; n5:.17/1.00; n6:.25/1.00; n7:.33/1.00	n4:.25/1.00; n5:.08/.42; n6:.25/.50; n7:.33/1.00
deepseek-v4-flash-none | E | 3S+C	n4:.08/1.00; n5:.08/1.00; n6:.33/.92; n7:.17/.92	n4:.17/.92; n5:.33/.58; n6:.25/.75; n7:.33/1.00
deepseek-v4-flash-none | E | 0S	n4:.17/.92; n5:.25/.67; n6:.25/.67; n7:.17/.42	n4:.08/.58; n5:.17/.67; n6:.25/.67; n7:.17/.67
deepseek-v4-flash-none | R | 3S	n4:.08/1.00; n5:.00/1.00; n6:.00/.92; n7:.00/1.00	n4:.00/1.00; n5:.00/1.00; n6:.00/1.00; n7:.00/1.00
deepseek-v4-flash-none | R | 3S+C	n4:.08/1.00; n5:.08/1.00; n6:.00/.92; n7:.08/1.00	n4:.17/1.00; n5:.00/1.00; n6:.08/1.00; n7:.08/1.00
deepseek-v4-flash-none | R | 0S	n4:.25/1.00; n5:.00/1.00; n6:.00/1.00; n7:.08/1.00	n4:.17/1.00; n5:.00/1.00; n6:.00/.92; n7:.00/.92
deepseek-v4-pro-high | D | 3S	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00
deepseek-v4-pro-high | D | 3S+C	n4:1.00/1.00; n5:.83/.83; n6:.75/.75; n7:1.00/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:.83/.83; n7:1.00/1.00
deepseek-v4-pro-high | D | 0S	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:.80/.80; n7:1.00/1.00
deepseek-v4-pro-high | E | 3S	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00
deepseek-v4-pro-high | E | 3S+C	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:.88/1.00; n7:1.00/1.00
deepseek-v4-pro-high | E | 0S	n4:.90/.90; n5:1.00/1.00; n6:1.00/1.00; n7:.88/.88	n4:.75/.75; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00
deepseek-v4-pro-high | R | 3S	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:NA/NA
deepseek-v4-pro-high | R | 3S+C	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00
deepseek-v4-pro-high | R | 0S	n4:1.00/1.00; n5:NA/NA; n6:1.00/1.00; n7:NA/NA	n4:1.00/1.00; n5:1.00/1.00; n6:NA/NA; n7:1.00/1.00
deepseek-v4-pro-none | D | 3S	n4:.42/1.00; n5:.33/1.00; n6:.50/1.00; n7:.50/1.00	n4:.33/1.00; n5:.33/1.00; n6:.33/1.00; n7:.58/1.00
deepseek-v4-pro-none | D | 3S+C	n4:.50/1.00; n5:.25/1.00; n6:.58/1.00; n7:.50/1.00	n4:.45/1.00; n5:.25/1.00; n6:.50/1.00; n7:.42/1.00
deepseek-v4-pro-none | D | 0S	n4:.42/.92; n5:.42/1.00; n6:.58/1.00; n7:.45/1.00	n4:.67/1.00; n5:.42/1.00; n6:.42/1.00; n7:.58/1.00
deepseek-v4-pro-none | E | 3S	n4:.00/1.00; n5:.00/1.00; n6:.00/1.00; n7:.00/.92	n4:.00/1.00; n5:.00/.83; n6:.00/1.00; n7:.00/1.00
deepseek-v4-pro-none | E | 3S+C	n4:.00/1.00; n5:.00/1.00; n6:.00/1.00; n7:.00/1.00	n4:.00/1.00; n5:.00/.92; n6:.00/1.00; n7:.00/1.00
deepseek-v4-pro-none | E | 0S	n4:.00/.17; n5:.00/.17; n6:.00/.08; n7:.00/.08	n4:.00/.00; n5:.00/.08; n6:.00/.00; n7:.00/.00
deepseek-v4-pro-none | R | 3S	n4:.25/1.00; n5:.00/1.00; n6:.00/1.00; n7:.08/1.00	n4:.17/1.00; n5:.00/1.00; n6:.00/1.00; n7:.10/1.00
deepseek-v4-pro-none | R | 3S+C	n4:.18/1.00; n5:.00/1.00; n6:.00/1.00; n7:.08/1.00	n4:.00/1.00; n5:.00/1.00; n6:.00/1.00; n7:.09/1.00
deepseek-v4-pro-none | R | 0S	n4:.17/1.00; n5:.00/.92; n6:.00/1.00; n7:.09/1.00	n4:.08/.92; n5:.00/.83; n6:.00/.91; n7:.10/1.00

Table 15: Equivalent-expression-only invariance. Full machine-readable values are supplied as CSV.

system_id | task | prompt_mode	B tier:E/F	Q tier:E/F
deepseek-v4-flash-high | D | 3S	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:.75/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00
deepseek-v4-flash-high | D | 3S+C	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00
deepseek-v4-flash-high | D | 0S	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00
deepseek-v4-flash-high | E | 3S	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:.89/1.00	n4:.90/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00
deepseek-v4-flash-high | E | 3S+C	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:.92/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00
deepseek-v4-flash-high | E | 0S	n4:.91/.91; n5:.90/.90; n6:.88/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:.91/.91; n7:1.00/1.00
deepseek-v4-flash-high | R | 3S	n4:.78/1.00; n5:.67/1.00; n6:1.00/1.00; n7:.75/1.00	n4:.83/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00
deepseek-v4-flash-high | R | 3S+C	n4:.78/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:.88/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00
deepseek-v4-flash-high | R | 0S	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:NA/NA
deepseek-v4-flash-none | D | 3S	n4:.33/1.00; n5:.42/1.00; n6:.50/1.00; n7:.42/1.00	n4:.50/1.00; n5:.50/1.00; n6:.50/1.00; n7:.58/1.00
deepseek-v4-flash-none | D | 3S+C	n4:.50/1.00; n5:.50/1.00; n6:.67/1.00; n7:.50/1.00	n4:.50/1.00; n5:.50/1.00; n6:.58/1.00; n7:.42/1.00
deepseek-v4-flash-none | D | 0S	n4:.50/1.00; n5:.42/1.00; n6:.58/1.00; n7:.58/1.00	n4:.50/1.00; n5:.50/1.00; n6:.58/1.00; n7:.50/1.00
deepseek-v4-flash-none | E | 3S	n4:.00/1.00; n5:.17/1.00; n6:.25/1.00; n7:.33/1.00	n4:.42/.92; n5:.33/1.00; n6:.25/1.00; n7:.36/.91
deepseek-v4-flash-none | E | 3S+C	n4:.08/1.00; n5:.08/1.00; n6:.33/.92; n7:.17/.92	n4:.25/.75; n5:.18/.91; n6:.42/1.00; n7:.33/1.00
deepseek-v4-flash-none | E | 0S	n4:.17/.92; n5:.25/.67; n6:.25/.67; n7:.17/.42	n4:.25/.83; n5:.27/.82; n6:.18/.64; n7:.25/.83
deepseek-v4-flash-none | R | 3S	n4:.08/1.00; n5:.00/1.00; n6:.00/.92; n7:.00/1.00	n4:.17/1.00; n5:.08/1.00; n6:.00/1.00; n7:.08/1.00
deepseek-v4-flash-none | R | 3S+C	n4:.08/1.00; n5:.08/1.00; n6:.00/.92; n7:.08/1.00	n4:.42/1.00; n5:.08/1.00; n6:.00/1.00; n7:.08/1.00
deepseek-v4-flash-none | R | 0S	n4:.25/1.00; n5:.00/1.00; n6:.00/1.00; n7:.08/1.00	n4:.08/1.00; n5:.00/1.00; n6:.00/.92; n7:.09/1.00
deepseek-v4-pro-high | D | 3S	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00
deepseek-v4-pro-high | D | 3S+C	n4:1.00/1.00; n5:.83/.83; n6:.75/.75; n7:1.00/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00
deepseek-v4-pro-high | D | 0S	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:.90/.90; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00
deepseek-v4-pro-high | E | 3S	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00
deepseek-v4-pro-high | E | 3S+C	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:.92/.92; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00
deepseek-v4-pro-high | E | 0S	n4:.90/.90; n5:1.00/1.00; n6:1.00/1.00; n7:.88/.88	n4:.91/.91; n5:.86/.86; n6:1.00/1.00; n7:1.00/1.00
deepseek-v4-pro-high | R | 3S	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:NA/NA
deepseek-v4-pro-high | R | 3S+C	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00
deepseek-v4-pro-high | R | 0S	n4:1.00/1.00; n5:NA/NA; n6:1.00/1.00; n7:NA/NA	n4:1.00/1.00; n5:NA/NA; n6:NA/NA; n7:NA/NA
deepseek-v4-pro-none | D | 3S	n4:.42/1.00; n5:.33/1.00; n6:.50/1.00; n7:.50/1.00	n4:.45/1.00; n5:.50/1.00; n6:.50/1.00; n7:.42/1.00
deepseek-v4-pro-none | D | 3S+C	n4:.50/1.00; n5:.25/1.00; n6:.58/1.00; n7:.50/1.00	n4:.50/1.00; n5:.50/1.00; n6:.50/1.00; n7:.50/1.00
deepseek-v4-pro-none | D | 0S	n4:.42/.92; n5:.42/1.00; n6:.58/1.00; n7:.45/1.00	n4:.50/1.00; n5:.50/1.00; n6:.45/1.00; n7:.42/1.00
deepseek-v4-pro-none | E | 3S	n4:.00/1.00; n5:.00/1.00; n6:.00/1.00; n7:.00/.92	n4:.33/.89; n5:.08/1.00; n6:.17/1.00; n7:.08/1.00
deepseek-v4-pro-none | E | 3S+C	n4:.00/1.00; n5:.00/1.00; n6:.00/1.00; n7:.00/1.00	n4:.50/1.00; n5:.08/.92; n6:.17/1.00; n7:.00/1.00
deepseek-v4-pro-none | E | 0S	n4:.00/.17; n5:.00/.17; n6:.00/.08; n7:.00/.08	n4:.09/.36; n5:.00/.17; n6:.00/.00; n7:.00/.08
deepseek-v4-pro-none | R | 3S	n4:.25/1.00; n5:.00/1.00; n6:.00/1.00; n7:.08/1.00	n4:.17/1.00; n5:.08/1.00; n6:.08/1.00; n7:.08/1.00
deepseek-v4-pro-none | R | 3S+C	n4:.18/1.00; n5:.00/1.00; n6:.00/1.00; n7:.08/1.00	n4:.00/1.00; n5:.00/1.00; n6:.08/1.00; n7:.08/1.00
deepseek-v4-pro-none | R | 0S	n4:.17/1.00; n5:.00/.92; n6:.00/1.00; n7:.09/1.00	n4:.25/1.00; n5:.08/1.00; n6:.08/.92; n7:.08/.75

Table 16: Distractor-only robustness. Full machine-readable values are supplied as CSV.

system_id | task | prompt_mode	B tier:E/F	X tier:E/F
deepseek-v4-flash-high | D | 3S	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:.75/1.00	n4:.92/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00
deepseek-v4-flash-high | D | 3S+C	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00
deepseek-v4-flash-high | D | 0S	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00
deepseek-v4-flash-high | E | 3S	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:.89/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00
deepseek-v4-flash-high | E | 3S+C	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:.92/1.00; n6:.89/1.00; n7:1.00/1.00
deepseek-v4-flash-high | E | 0S	n4:.91/.91; n5:.90/.90; n6:.88/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:.91/.91; n6:1.00/1.00; n7:1.00/1.00
deepseek-v4-flash-high | R | 3S	n4:.78/1.00; n5:.67/1.00; n6:1.00/1.00; n7:.75/1.00	n4:.86/1.00; n5:.80/.80; n6:NA/NA; n7:1.00/1.00
deepseek-v4-flash-high | R | 3S+C	n4:.78/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:.91/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00
deepseek-v4-flash-high | R | 0S	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00
deepseek-v4-flash-none | D | 3S	n4:.33/1.00; n5:.42/1.00; n6:.50/1.00; n7:.42/1.00	n4:.33/1.00; n5:.50/1.00; n6:.58/1.00; n7:.58/1.00
deepseek-v4-flash-none | D | 3S+C	n4:.50/1.00; n5:.50/1.00; n6:.67/1.00; n7:.50/1.00	n4:.42/1.00; n5:.42/1.00; n6:.50/1.00; n7:.67/1.00
deepseek-v4-flash-none | D | 0S	n4:.50/1.00; n5:.42/1.00; n6:.58/1.00; n7:.58/1.00	n4:.42/1.00; n5:.50/1.00; n6:.50/1.00; n7:.33/1.00
deepseek-v4-flash-none | E | 3S	n4:.00/1.00; n5:.17/1.00; n6:.25/1.00; n7:.33/1.00	n4:.00/1.00; n5:.25/1.00; n6:.08/1.00; n7:.17/.92
deepseek-v4-flash-none | E | 3S+C	n4:.08/1.00; n5:.08/1.00; n6:.33/.92; n7:.17/.92	n4:.00/1.00; n5:.25/.92; n6:.17/1.00; n7:.33/.92
deepseek-v4-flash-none | E | 0S	n4:.17/.92; n5:.25/.67; n6:.25/.67; n7:.17/.42	n4:.00/1.00; n5:.08/.67; n6:.08/1.00; n7:.27/.82
deepseek-v4-flash-none | R | 3S	n4:.08/1.00; n5:.00/1.00; n6:.00/.92; n7:.00/1.00	n4:.08/1.00; n5:.00/1.00; n6:.00/.92; n7:.00/1.00
deepseek-v4-flash-none | R | 3S+C	n4:.08/1.00; n5:.08/1.00; n6:.00/.92; n7:.08/1.00	n4:.08/1.00; n5:.00/1.00; n6:.08/1.00; n7:.08/1.00
deepseek-v4-flash-none | R | 0S	n4:.25/1.00; n5:.00/1.00; n6:.00/1.00; n7:.08/1.00	n4:.08/1.00; n5:.00/1.00; n6:.00/1.00; n7:.08/1.00
deepseek-v4-pro-high | D | 3S	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00
deepseek-v4-pro-high | D | 3S+C	n4:1.00/1.00; n5:.83/.83; n6:.75/.75; n7:1.00/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00
deepseek-v4-pro-high | D | 0S	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:.80/.80; n7:1.00/1.00
deepseek-v4-pro-high | E | 3S	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00
deepseek-v4-pro-high | E | 3S+C	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00
deepseek-v4-pro-high | E | 0S	n4:.90/.90; n5:1.00/1.00; n6:1.00/1.00; n7:.88/.88	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00
deepseek-v4-pro-high | R | 3S	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:.80/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00
deepseek-v4-pro-high | R | 3S+C	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00
deepseek-v4-pro-high | R | 0S	n4:1.00/1.00; n5:NA/NA; n6:1.00/1.00; n7:NA/NA	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00
deepseek-v4-pro-none | D | 3S	n4:.42/1.00; n5:.33/1.00; n6:.50/1.00; n7:.50/1.00	n4:.42/1.00; n5:.50/1.00; n6:.50/1.00; n7:.50/1.00
deepseek-v4-pro-none | D | 3S+C	n4:.50/1.00; n5:.25/1.00; n6:.58/1.00; n7:.50/1.00	n4:.50/1.00; n5:.33/1.00; n6:.42/1.00; n7:.50/1.00
deepseek-v4-pro-none | D | 0S	n4:.42/.92; n5:.42/1.00; n6:.58/1.00; n7:.45/1.00	n4:.50/1.00; n5:.50/1.00; n6:.50/1.00; n7:.50/1.00
deepseek-v4-pro-none | E | 3S	n4:.00/1.00; n5:.00/1.00; n6:.00/1.00; n7:.00/.92	n4:.00/1.00; n5:.00/1.00; n6:.00/.92; n7:.00/1.00
deepseek-v4-pro-none | E | 3S+C	n4:.00/1.00; n5:.00/1.00; n6:.00/1.00; n7:.00/1.00	n4:.00/1.00; n5:.00/1.00; n6:.00/1.00; n7:.00/1.00
deepseek-v4-pro-none | E | 0S	n4:.00/.17; n5:.00/.17; n6:.00/.08; n7:.00/.08	n4:.00/.17; n5:.00/.00; n6:.00/.09; n7:.00/.08
deepseek-v4-pro-none | R | 3S	n4:.25/1.00; n5:.00/1.00; n6:.00/1.00; n7:.08/1.00	n4:.27/1.00; n5:.00/1.00; n6:.00/1.00; n7:.08/1.00
deepseek-v4-pro-none | R | 3S+C	n4:.18/1.00; n5:.00/1.00; n6:.00/1.00; n7:.08/1.00	n4:.09/1.00; n5:.00/1.00; n6:.00/1.00; n7:.08/1.00
deepseek-v4-pro-none | R | 0S	n4:.17/1.00; n5:.00/.92; n6:.00/1.00; n7:.09/1.00	n4:.17/1.00; n5:.00/.91; n6:.00/1.00; n7:.08/.92

Table 17: Model-level paired correlation matrices underlying Figure 7. Each cell reports Pearson r/permutation p/paired n/Const, where Const marks an off-diagonal zero imposed because at least one paired vector is constant. Full machine-readable values are supplied as CSV.

panel | row system	deepseek-v4-flash-high r/p/n/Const	deepseek-v4-flash-none r/p/n/Const	deepseek-v4-pro-high r/p/n/Const	deepseek-v4-pro-none r/p/n/Const
Decision | deepseek-v4-flash-high	1.00/.00/33/N	.00/1.00/33/Y	.00/1.00/19/Y	.00/1.00/32/Y
Decision | deepseek-v4-flash-none	.00/1.00/33/Y	1.00/.00/48/N	.00/1.00/26/Y	-0.28/.06/47/N
Decision | deepseek-v4-pro-high	.00/1.00/19/Y	.00/1.00/26/Y	1.00/.00/26/N	.00/1.00/26/Y
Decision | deepseek-v4-pro-none	.00/1.00/32/Y	-0.28/.06/47/N	.00/1.00/26/Y	1.00/.00/47/N
Enumeration | deepseek-v4-flash-high	1.00/.00/37/N	.14/.51/37/N	-0.05/1.00/28/N	.00/1.00/37/Y
Enumeration | deepseek-v4-flash-none	.14/.51/37/N	1.00/.00/48/N	.16/.55/33/N	.00/1.00/48/Y
Enumeration | deepseek-v4-pro-high	-0.05/1.00/28/N	.16/.55/33/N	1.00/.00/33/N	.00/1.00/33/Y
Enumeration | deepseek-v4-pro-none	.00/1.00/37/Y	.00/1.00/48/Y	.00/1.00/33/Y	1.00/.00/48/N
Repair | deepseek-v4-flash-high	1.00/.00/8/N	.00/1.00/8/Y	.00/1.00/2/Y	.00/1.00/8/Y
Repair | deepseek-v4-flash-none	.00/1.00/8/Y	1.00/.00/48/N	.00/1.00/2/Y	.86/.00/46/N
Repair | deepseek-v4-pro-high	.00/1.00/2/Y	.00/1.00/2/Y	1.00/.00/2/N	.00/1.00/2/Y
Repair | deepseek-v4-pro-none	.00/1.00/8/Y	.86/.00/46/N	.00/1.00/2/Y	1.00/.00/46/N
View stability | deepseek-v4-flash-high	1.00/.00/15/N	.00/1.00/15/Y	.00/1.00/7/Y	.00/1.00/15/Y
View stability | deepseek-v4-flash-none	.00/1.00/15/Y	1.00/.00/48/N	.00/1.00/12/Y	.00/1.00/46/Y
View stability | deepseek-v4-pro-high	.00/1.00/7/Y	.00/1.00/12/Y	1.00/.00/12/N	.00/1.00/12/Y
View stability | deepseek-v4-pro-none	.00/1.00/15/Y	.00/1.00/46/Y	.00/1.00/12/Y	1.00/.00/46/N

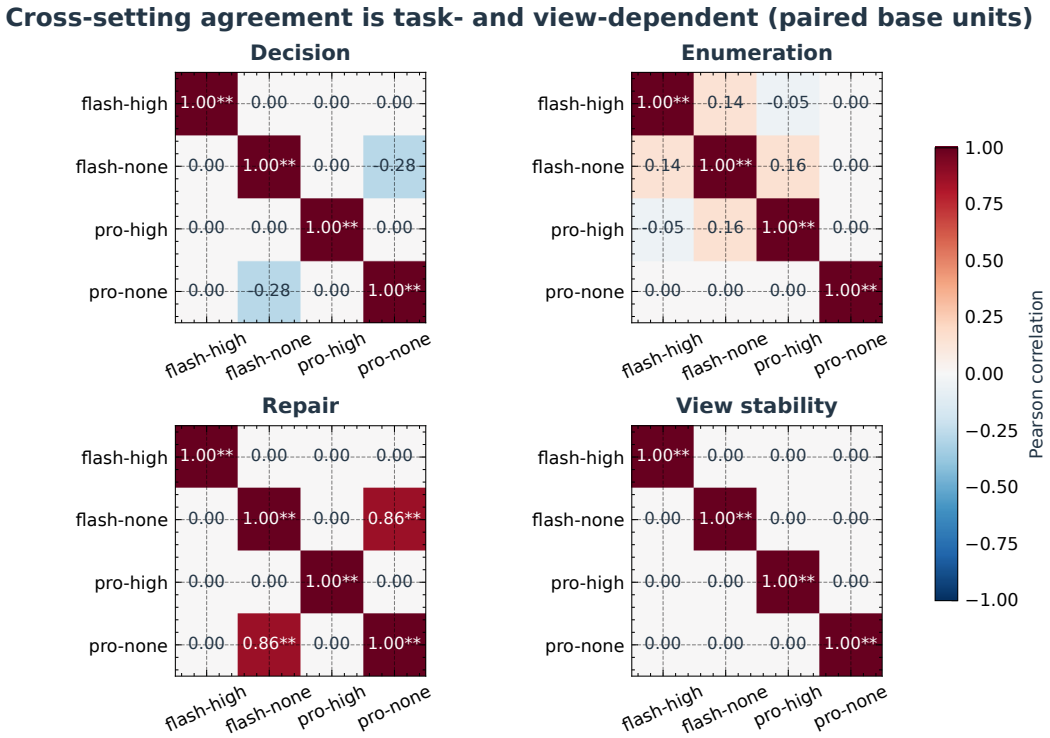

Figure 7: Paired setting correlations for decision, enumeration, repair, and view stability. Values are descriptive; constant-vector conventions and permutation counts are stated in source data.

F Interventions, Repair, and Cost

Cost accounting. Across the four frozen runs the harness issued 12,672 calls in expanded-v1, 7,344 in expanded-v2-tokenfix, 1,728 in the matched-cardinality extension, and 160 in the coarse cardinality extension, spending USD 41.8781, 45.0917, 4.1630, and 0.2010 respectively, for USD 91.3338 in total against a USD 100 ceiling. Table 23 registers a different and smaller population: the composed analysis matrix actually analysed above, which holds 12,000 cells of which 9,414 are observed, 7,344 replaced by the token-fix run and 4,656 retained from v1. Its request and cost columns therefore sum to 12,000 and USD 47.8298, and superseded expanded-v1 calls are deliberately counted once, under their replacement record, so the table is a subset rather than a ledger total. The last row of the table prints those sums directly.

Catalog pricing and reservation rule. Preflight freezes the catalog entry and its prompt- and completion-token prices for every requested slug and blocks the run if a slug, price, or requested reasoning setting is unavailable. For each request the conservative reservation is $\lceil \text{prompt characters}/4 \rceil p_{\text{in}} + L_{\text{out}} p_{\text{out}}$, where L_{out} is that task's completion cap. Before submitting work, the runner requires accumulated billed cost plus all in-flight reservations plus the next reservation to remain at or below USD 100; a completed request releases its reservation and contributes its returned billed cost. This is the catalog and reservation rule used for the worst-case totals in Section 4.

Table 18: Base-view decision outcomes by model and prompt. SC, SW, FI, and NA are mutually exclusive strict-correct, strict-wrong with valid format, format-invalid, and completion-censored classes; cells report count/share of requested calls. Full machine-readable values are supplied as CSV.

system_id	0S class=n/share	3S class=n/share	3S+C class=n/share
claude-sonnet-5	SC:14/.29; SW:.00/.00; FI:7/.15; NA:27/.56	SC:19/.40; SW:.00/.00; FI:2/.04; NA:27/.56	SC:9/.19; SW:.00/.00; FI:5/.10; NA:34/.71
deepseek-v4-flash-high	SC:33/.69; SW:.00/.00; FI:.00/.00; NA:15/.31	SC:24/.50; SW:1.00/.02; FI:.00/.00; NA:23/.48	SC:28/.58; SW:.00/.00; FI:.00/.00; NA:20/.42
deepseek-v4-pro-high	SC:26/.54; SW:.00/.00; FI:.00/.00; NA:22/.46	SC:19/.40; SW:.00/.00; FI:.00/.00; NA:29/.60	SC:20/.42; SW:.00/.00; FI:2/.04; NA:26/.54
gemini-3.1-pro-preview	SC:1.00/.02; SW:.00/.00; FI:47/.98; NA:.00/.00	SC:5/.10; SW:.00/.00; FI:43/.90; NA:.00/.00	SC:3/.06; SW:.00/.00; FI:45/.94; NA:.00/.00
gemini-3.8-flash	SC:13/.27; SW:.00/.00; FI:34/.71; NA:1.00/.02	SC:16/.33; SW:1.00/.02; FI:30/.62; NA:1.00/.02	SC:9/.19; SW:1.00/.02; FI:38/.79; NA:.00/.00
gpt-5.6-luna	SC:40/.83; SW:8/.17; FI:.00/.00; NA:.00/.00	SC:46/.96; SW:2/.04; FI:.00/.00; NA:.00/.00	SC:44/.92; SW:4/.08; FI:.00/.00; NA:.00/.00
grok-4.6	SC:48/1.00; SW:.00/.00; FI:.00/.00; NA:.00/.00	SC:48/1.00; SW:.00/.00; FI:.00/.00; NA:.00/.00	SC:48/1.00; SW:.00/.00; FI:.00/.00; NA:.00/.00
mistral-large	SC:.00/.00; SW:.00/.00; FI:48/1.00; NA:.00/.00	SC:.00/.00; SW:.00/.00; FI:48/1.00; NA:.00/.00	SC:.00/.00; SW:.00/.00; FI:48/1.00; NA:.00/.00
muse-spark-1.3	SC:19/.40; SW:.00/.00; FI:.00/.00; NA:29/.60	SC:12/.25; SW:.00/.00; FI:.00/.00; NA:36/.75	SC:11/.23; SW:.00/.00; FI:.00/.00; NA:37/.77
phi-4	SC:.00/.00; SW:.00/.00; FI:39/.81; NA:9/.19	SC:.00/.00; SW:.00/.00; FI:36/.75; NA:12/.25	SC:.00/.00; SW:.00/.00; FI:31/.65; NA:17/.35
qwen3.8-a95b	SC:32/.67; SW:.00/.00; FI:.00/.00; NA:16/.33	SC:34/.71; SW:.00/.00; FI:.00/.00; NA:14/.29	SC:30/.62; SW:.00/.00; FI:.00/.00; NA:18/.38
qwen3.8-max	SC:37/.77; SW:.00/.00; FI:.00/.00; NA:11/.23	SC:28/.58; SW:.00/.00; FI:.00/.00; NA:20/.42	SC:29/.60; SW:.00/.00; FI:.00/.00; NA:19/.40

Table 19: Certificate intervention on enumeration. Cells report E/P/R/Fmt; F1 is omitted because it is not defined in this aggregate table. Full machine-readable values are supplied as CSV.

system_id	3S: tier=E/P/R/Fmt	3S+C: tier=E/P/R/Fmt
claude-sonnet-5	D=n4:.55/1.00/1.00/.55; D=n5:.09/1.00/1.00/.09; D=n6:.00/.71/.71/.00; D=n7:.00/1.00/1.00/.00	D=n4:.17/1.00/1.00/.17; D=n5:.10/1.00/1.00/.10; D=n6:.00/1.00/1.00/.00; D=n7:.00/1.00/1.00/.00
deepseek-v4-flash-high	D=n4:1.00/1.00/1.00/1.00; D=n5:1.00/1.00/1.00/1.00; D=n6:1.00/1.00/1.00/1.00; D=n7:.89/.89/.89/1.00	D=n4:1.00/1.00/1.00/1.00; D=n5:1.00/1.00/1.00/1.00; D=n6:1.00/1.00/1.00/1.00; D=n7:1.00/1.00/1.00/1.00
deepseek-v4-pro-high	D=n4:1.00/1.00/1.00/1.00; D=n5:1.00/1.00/1.00/1.00; D=n6:1.00/1.00/1.00/1.00; D=n7:1.00/1.00/1.00/1.00	D=n4:1.00/1.00/1.00/1.00; D=n5:1.00/1.00/1.00/1.00; D=n6:1.00/1.00/1.00/1.00; D=n7:1.00/1.00/1.00/1.00
gemini-3.1-pro-preview	D=n4:.00/1.00/1.00/.00; D=n5:.00/1.00/1.00/.00; D=n6:.00/1.00/1.00/.00; D=n7:.00/1.00/1.00/.00	D=n4:.00/1.00/1.00/.00; D=n5:.00/1.00/1.00/.00; D=n6:.00/1.00/1.00/.00; D=n7:.00/1.00/1.00/.00
gemini-3.8-flash	D=n4:.08/1.00/1.00/.08; D=n5:.00/.92/.92/.00; D=n6:.08/.92/.92/.08; D=n7:.00/.92/.92/.00	D=n4:.00/.92/.92/.00; D=n5:.25/.92/.92/.25; D=n6:.08/.92/.92/.08; D=n7:.00/.75/.75/.00
gpt-5.6-luna	D=n4:.75/.82/.92/1.00; D=n5:.92/.92/.92/1.00; D=n6:.67/.72/.75/1.00; D=n7:.75/.78/.83/1.00	D=n4:.92/.92/.92/1.00; D=n5:1.00/1.00/1.00/1.00; D=n6:.83/.89/.92/1.00; D=n7:.75/.75/.75/1.00
grok-4.6	D=n4:1.00/1.00/1.00/1.00; D=n5:1.00/1.00/1.00/1.00; D=n6:1.00/1.00/1.00/1.00; D=n7:1.00/1.00/1.00/1.00	D=n4:1.00/1.00/1.00/1.00; D=n5:1.00/1.00/1.00/1.00; D=n6:1.00/1.00/1.00/1.00; D=n7:1.00/1.00/1.00/1.00
mistral-large	D=n4:.00/.06/.17/.00; D=n5:.00/.05/.17/.00; D=n6:.00/.12/.19/.00; D=n7:.00/.01/.08/.00	D=n4:.00/.04/.25/.00; D=n5:.00/.04/.08/.00; D=n6:.00/.04/.17/.00; D=n7:.00/.00/.00/.00
muse-spark-1.3	D=n4:1.00/1.00/1.00/1.00; D=n5:1.00/1.00/1.00/1.00; D=n6:1.00/1.00/1.00/1.00; D=n7:1.00/1.00/1.00/1.00	D=n4:1.00/1.00/1.00/1.00; D=n5:1.00/1.00/1.00/1.00; D=n6:1.00/1.00/1.00/1.00; D=n7:1.00/1.00/1.00/1.00
phi-4	D=n4:.00/.02/.08/.00; D=n5:.00/.02/.08/.00; D=n6:.00/.03/.09/.00; D=n7:.00/.00/.00/.00	D=n4:.00/.02/.08/.00; D=n5:.00/.00/.00/.00; D=n6:.00/.02/.12/.00; D=n7:.00/.00/.25/.00
qwen3.8-a95b	D=n4:1.00/1.00/1.00/1.00; D=n5:1.00/1.00/1.00/1.00; D=n6:1.00/1.00/1.00/1.00; D=n7:1.00/1.00/1.00/1.00	D=n4:1.00/1.00/1.00/1.00; D=n5:1.00/1.00/1.00/1.00; D=n6:1.00/1.00/1.00/1.00; D=n7:1.00/1.00/1.00/1.00
qwen3.8-max	D=n4:1.00/1.00/1.00/1.00; D=n5:.91/.91/.91/1.00; D=n6:1.00/1.00/1.00/1.00; D=n7:1.00/1.00/1.00/1.00	D=n4:1.00/1.00/1.00/1.00; D=n5:1.00/1.00/1.00/1.00; D=n6:1.00/1.00/1.00/1.00; D=n7:1.00/1.00/1.00/1.00

Table 20: Cross-task prompt intervention. Full machine-readable values are supplied as CSV.

system_id	D-0S: tier=E/F	D-3S: tier=E/F	D-3S+C: tier=E/F	E-0S: tier=E/F	E-3S: tier=E/F	E-3S+C: tier=E/F	R-0S: tier=E/F	R-3S: tier=E/F	R-3S+C: tier=E/F
claude-sonnet-5	D=n4:.67/.67; D=n5:.50/.50; D=n6:.67/.67; D=n7:1.00/1.00	D=n4:.89/.89; D=n5:.80/.80; D=n6:1.00/1.00; D=n7:1.00/1.00	D=n4:.71/.71; D=n5:1.00/1.00; D=n6:.60/.60; D=n7:.00/.00	D=n4:.08/.08; D=n5:.00/.00; D=n6:.00/.00; D=n7:.00/.00	D=n4:.55/.55; D=n5:.09/.09; D=n6:.00/.00; D=n7:.00/.00	D=n4:.17/.17; D=n5:.10/.10; D=n6:.00/.00; D=n7:.00/.00	D=n4:.17/.17; D=n5:.00/.00; D=n6:.00/.00; D=n7:.00/.00	--	--
deepseek-v4-flash-high	D=n4:1.00/1.00; D=n5:1.00/1.00; D=n6:1.00/1.00; D=n7:1.00/1.00	D=n4:1.00/1.00; D=n5:1.00/1.00; D=n6:1.00/1.00; D=n7:.75/1.00	D=n4:1.00/1.00; D=n5:1.00/1.00; D=n6:1.00/1.00; D=n7:1.00/1.00	D=n4:.91/.91; D=n5:.90/.90; D=n6:.88/1.00; D=n7:1.00/1.00	D=n4:1.00/1.00; D=n5:1.00/1.00; D=n6:1.00/1.00; D=n7:.89/1.00	D=n4:1.00/1.00; D=n5:1.00/1.00; D=n6:1.00/1.00; D=n7:1.00/1.00	D=n4:1.00/1.00; D=n5:1.00/1.00; D=n6:1.00/1.00; D=n7:1.00/1.00	D=n4:.78/1.00; D=n5:.67/1.00; D=n6:1.00/1.00; D=n7:.75/1.00	D=n4:.78/1.00; D=n5:1.00/1.00; D=n6:1.00/1.00; D=n7:1.00/1.00
deepseek-v4-pro-high	D=n4:1.00/1.00; D=n5:1.00/1.00; D=n6:1.00/1.00; D=n7:1.00/1.00	D=n4:1.00/1.00; D=n5:1.00/1.00; D=n6:1.00/1.00; D=n7:1.00/1.00	D=n4:1.00/1.00; D=n5:.83/.83; D=n6:.75/.75; D=n7:1.00/1.00	D=n4:.90/.90; D=n5:1.00/1.00; D=n6:1.00/1.00; D=n7:.88/.88	D=n4:1.00/1.00; D=n5:1.00/1.00; D=n6:1.00/1.00; D=n7:1.00/1.00	D=n4:1.00/1.00; D=n5:1.00/1.00; D=n6:1.00/1.00; D=n7:1.00/1.00	D=n4:1.00/1.00; D=n5:NA/NA; D=n6:1.00/1.00; D=n7:NA/NA	D=n4:1.00/1.00; D=n5:1.00/1.00; D=n6:1.00/1.00; D=n7:1.00/1.00	D=n4:1.00/1.00; D=n5:1.00/1.00; D=n6:1.00/1.00; D=n7:1.00/1.00
gemini-3.1-pro-preview	D=n4:.08/.08; D=n5:.00/.00; D=n6:.00/.00; D=n7:.00/.00	D=n4:.08/.08; D=n5:.17/.17; D=n6:.08/.08; D=n7:.08/.08	D=n4:.00/.00; D=n5:.00/.00; D=n6:.17/.17; D=n7:.08/.08	D=n4:.00/.00; D=n5:.00/.00; D=n6:.00/.00; D=n7:.00/.00	D=n4:.00/.00; D=n5:.00/.00; D=n6:.00/.00; D=n7:.00/.00	D=n4:.00/.00; D=n5:.00/.00; D=n6:.00/.00; D=n7:.00/.00	D=n4:.08/.17; D=n5:.00/.00; D=n6:.00/.00; D=n7:.08/.08	--	--
gemini-3.8-flash	D=n4:.25/.25; D=n5:.50/.50; D=n6:.08/.08; D=n7:.27/.27	D=n4:.33/.33; D=n5:.33/.33; D=n6:.36/.36; D=n7:.33/.42	D=n4:.17/.17; D=n5:.33/.33; D=n6:.08/.17; D=n7:.17/.17	D=n4:.25/.25; D=n5:.17/.17; D=n6:.08/.08; D=n7:.00/.00	D=n4:.08/.08; D=n5:.00/.00; D=n6:.08/.08; D=n7:.00/.00	D=n4:.00/.00; D=n5:.25/.25; D=n6:.08/.08; D=n7:.00/.00	D=n4:.00/.00; D=n5:.17/.17; D=n6:.08/.08; D=n7:.18/.18	--	--
gpt-5.6-luna	D=n4:.75/1.00; D=n5:1.00/1.00; D=n6:.92/1.00; D=n7:.67/1.00	D=n4:1.00/1.00; D=n5:1.00/1.00; D=n6:.92/1.00; D=n7:.92/1.00	D=n4:.92/1.00; D=n5:1.00/1.00; D=n6:.92/1.00; D=n7:.83/1.00	D=n4:.92/1.00; D=n5:.83/1.00; D=n6:.83/1.00; D=n7:.58/1.00	D=n4:.75/1.00; D=n5:.92/1.00; D=n6:.67/1.00; D=n7:.75/1.00	D=n4:.92/1.00; D=n5:1.00/1.00; D=n6:.83/1.00; D=n7:.75/1.00	D=n4:.67/1.00; D=n5:.91/1.00; D=n6:.67/.92; D=n7:1.00/1.00	--	--
grok-4.6	D=n4:1.00/1.00; D=n5:1.00/1.00; D=n6:1.00/1.00; D=n7:1.00/1.00	D=n4:1.00/1.00; D=n5:1.00/1.00; D=n6:1.00/1.00; D=n7:1.00/1.00	D=n4:1.00/1.00; D=n5:1.00/1.00; D=n6:1.00/1.00; D=n7:1.00/1.00	D=n4:1.00/1.00; D=n5:1.00/1.00; D=n6:1.00/1.00; D=n7:.92/1.00	D=n4:1.00/1.00; D=n5:1.00/1.00; D=n6:1.00/1.00; D=n7:1.00/1.00	D=n4:1.00/1.00; D=n5:1.00/1.00; D=n6:1.00/1.00; D=n7:1.00/1.00	D=n4:1.00/1.00; D=n5:1.00/1.00; D=n6:1.00/1.00; D=n7:1.00/1.00	--	--
mistral-large	D=n4:.00/.00; D=n5:.00/.00; D=n6:.00/.00; D=n7:.00/.00	D=n4:.00/.00; D=n5:.00/.00; D=n6:.00/.00; D=n7:.00/.00	D=n4:.00/.00; D=n5:.00/.00; D=n6:.00/.00; D=n7:.00/.00	D=n4:.00/.00; D=n5:.00/.00; D=n6:.00/.00; D=n7:.00/.00	D=n4:.00/.00; D=n5:.00/.00; D=n6:.00/.00; D=n7:.00/.00	D=n4:.00/.00; D=n5:.00/.00; D=n6:.00/.00; D=n7:.00/.00	D=n4:.00/.00; D=n5:.00/.00; D=n6:.00/.00; D=n7:.00/.00	--	--
muse-spark-1.3	D=n4:1.00/1.00; D=n5:1.00/1.00; D=n6:1.00/1.00; D=n7:1.00/1.00	D=n4:1.00/1.00; D=n5:1.00/1.00; D=n6:1.00/1.00; D=n7:1.00/1.00	D=n4:1.00/1.00; D=n5:1.00/1.00; D=n6:NA/NA; D=n7:1.00/1.00	D=n4:.91/.91; D=n5:1.00/1.00; D=n6:1.00/1.00; D=n7:1.00/1.00	D=n4:1.00/1.00; D=n5:1.00/1.00; D=n6:1.00/1.00; D=n7:1.00/1.00	D=n4:1.00/1.00; D=n5:1.00/1.00; D=n6:1.00/1.00; D=n7:1.00/1.00	D=n4:1.00/1.00; D=n5:1.00/1.00; D=n6:1.00/1.00; D=n7:1.00/1.00	--	--
phi-4	D=n4:.00/.00; D=n5:.00/.00; D=n6:.00/.00; D=n7:.00/.00	D=n4:.00/.00; D=n5:.00/.00; D=n6:.00/.00; D=n7:.00/.00	D=n4:.00/.00; D=n5:.00/.00; D=n6:.00/.00; D=n7:.00/.00	D=n4:.00/.00; D=n5:.00/.00; D=n6:.00/.00; D=n7:.00/.00	D=n4:.00/.00; D=n5:.00/.00; D=n6:.00/.00; D=n7:.00/.00	D=n4:.00/.00; D=n5:.00/.00; D=n6:.00/.00; D=n7:.00/.00	D=n4:.00/.00; D=n5:.00/.00; D=n6:.00/.00; D=n7:.00/.00	--	--
qwen3.8-a95b	D=n4:1.00/1.00; D=n5:1.00/1.00; D=n6:1.00/1.00; D=n7:1.00/1.00	D=n4:1.00/1.00; D=n5:1.00/1.00; D=n6:1.00/1.00; D=n7:1.00/1.00	D=n4:1.00/1.00; D=n5:1.00/1.00; D=n6:1.00/1.00; D=n7:1.00/1.00	D=n4:1.00/1.00; D=n5:1.00/1.00; D=n6:1.00/1.00; D=n7:1.00/1.00	D=n4:1.00/1.00; D=n5:1.00/1.00; D=n6:1.00/1.00; D=n7:1.00/1.00	D=n4:1.00/1.00; D=n5:1.00/1.00; D=n6:1.00/1.00; D=n7:1.00/1.00	D=n4:.86/1.00; D=n5:1.00/1.00; D=n6:.50/1.00; D=n7:1.00/1.00	--	--

system_id	D-0S: tier=E/F	D-3S: tier=E/F	D-3S+C: tier=E/F	E-0S: tier=E/F	E-3S: tier=E/F	E-3S+C: tier=E/F	R-0S: tier=E/F	R-3S: tier=E/F	R-3S+C: tier=E/F
qwen3.8-max	D=n4:1.00/1.00; D=n5:1.00/1.00; D=n6:1.00/1.00; D=n7:1.00/1.00	D=n4:1.00/1.00; D=n5:1.00/1.00; D=n6:1.00/1.00; D=n7:1.00/1.00	D=n4:1.00/1.00; D=n5:1.00/1.00; D=n6:1.00/1.00; D=n7:1.00/1.00	D=n4:1.00/1.00; D=n5:1.00/1.00; D=n6:1.00/1.00; D=n7:1.00/1.00	D=n4:1.00/1.00; D=n5:.91/1.00; D=n6:1.00/1.00; D=n7:1.00/1.00	D=n4:1.00/1.00; D=n5:1.00/1.00; D=n6:1.00/1.00; D=n7:1.00/1.00	D=n4:1.00/1.00; D=n5:1.00/1.00; D=n6:1.00/1.00; D=n7:NA/NA	--	--

Table 21: Third-task conditioned re-enumeration by corruption type. Flag counts responses whose error flag is true, which the prompt already guarantees for every instance; E is the exact-set endpoint and P, R, Fmt follow the convention of Table 4. F1 is omitted because it is not defined in this aggregate table. Full machine-readable values are supplied as CSV.

system_id	0S: kind,tier=Flag/E/P/R/Fmt	3S: kind,tier=Flag/E/P/R/Fmt	3S+C: kind,tier=Flag/E/P/R/Fmt
claude-sonnet-5	C=missing_solution,D=n5:NA/NA/NA/NA/NA; C=missing_solution,D=n6:NA/NA/NA/NA/NA; C=missing_solution,D=n7:NA/NA/NA/NA/NA; C=replaced_solution,D=n4:1.00/.17/1.00/1.00/.17; C=replaced_solution,D=n5:1.00/.00/1.00/1.00/.00; C=replaced_solution,D=n6:1.00/.00/1.00/1.00/.00; C=replaced_solution,D=n7:.67/.00/.67/.67/.00; C=spurious_solution,D=n4:NA/NA/NA/NA/NA; C=spurious_solution,D=n5:1.00/.00/1.00/1.00/.00; C=spurious_solution,D=n6:NA/NA/NA/NA/NA; C=spurious_solution,D=n7:NA/NA/NA/NA/NA	--	--
deepseek-v4-flash-high	C=missing_solution,D=n5:NA/NA/NA/NA/NA; C=missing_solution,D=n6:NA/NA/NA/NA/NA; C=missing_solution,D=n7:NA/NA/NA/NA/NA; C=replaced_solution,D=n4:1.00/1.00/1.00/1.00/1.00; C=replaced_solution,D=n5:1.00/1.00/1.00/1.00/1.00; C=replaced_solution,D=n6:1.00/1.00/1.00/1.00/1.00; C=replaced_solution,D=n7:1.00/1.00/1.00/1.00/1.00; C=spurious_solution,D=n4:NA/NA/NA/NA/NA; C=spurious_solution,D=n5:NA/NA/NA/NA/NA; C=spurious_solution,D=n6:NA/NA/NA/NA/NA; C=spurious_solution,D=n7:NA/NA/NA/NA/NA	C=missing_solution,D=n5:NA/NA/NA/NA/NA; C=missing_solution,D=n6:NA/NA/NA/NA/NA; C=missing_solution,D=n7:NA/NA/NA/NA/NA; C=replaced_solution,D=n4:1.00/.83/.83/.83/1.00; C=replaced_solution,D=n5:1.00/.00/.00/.00/1.00; C=replaced_solution,D=n6:1.00/1.00/1.00/1.00/1.00; C=replaced_solution,D=n7:1.00/1.00/1.00/1.00/1.00; C=spurious_solution,D=n4:1.00/.67/.67/.67/1.00; C=spurious_solution,D=n5:1.00/1.00/1.00/1.00/1.00; C=spurious_solution,D=n6:NA/NA/NA/NA/NA; C=spurious_solution,D=n7:1.00/.50/.50/.50/1.00	C=missing_solution,D=n5:NA/NA/NA/NA/NA; C=missing_solution,D=n6:1.00/1.00/1.00/1.00/1.00; C=missing_solution,D=n7:NA/NA/NA/NA/NA; C=replaced_solution,D=n4:1.00/.60/.60/.60/1.00; C=replaced_solution,D=n5:1.00/1.00/1.00/1.00/1.00; C=replaced_solution,D=n6:1.00/1.00/1.00/1.00/1.00; C=replaced_solution,D=n7:1.00/1.00/1.00/1.00/1.00; C=spurious_solution,D=n4:1.00/1.00/1.00/1.00/1.00; C=spurious_solution,D=n5:1.00/1.00/1.00/1.00/1.00; C=spurious_solution,D=n6:1.00/1.00/1.00/1.00/1.00; C=spurious_solution,D=n7:1.00/1.00/1.00/1.00/1.00
deepseek-v4-pro-high	C=missing_solution,D=n5:NA/NA/NA/NA/NA; C=missing_solution,D=n6:NA/NA/NA/NA/NA; C=missing_solution,D=n7:NA/NA/NA/NA/NA; C=replaced_solution,D=n4:1.00/1.00/1.00/1.00/1.00; C=replaced_solution,D=n5:NA/NA/NA/NA/NA; C=replaced_solution,D=n6:1.00/1.00/1.00/1.00/1.00; C=replaced_solution,D=n7:NA/NA/NA/NA/NA; C=spurious_solution,D=n4:NA/NA/NA/NA/NA; C=spurious_solution,D=n5:NA/NA/NA/NA/NA; C=spurious_solution,D=n6:NA/NA/NA/NA/NA; C=spurious_solution,D=n7:NA/NA/NA/NA/NA	C=missing_solution,D=n5:NA/NA/NA/NA/NA; C=missing_solution,D=n6:NA/NA/NA/NA/NA; C=missing_solution,D=n7:NA/NA/NA/NA/NA; C=replaced_solution,D=n4:1.00/1.00/1.00/1.00/1.00; C=replaced_solution,D=n5:1.00/1.00/1.00/1.00/1.00; C=replaced_solution,D=n6:1.00/1.00/1.00/1.00/1.00; C=replaced_solution,D=n7:NA/NA/NA/NA/NA; C=spurious_solution,D=n4:1.00/1.00/1.00/1.00/1.00; C=spurious_solution,D=n5:1.00/1.00/1.00/1.00/1.00; C=spurious_solution,D=n6:NA/NA/NA/NA/NA; C=spurious_solution,D=n7:1.00/1.00/1.00/1.00/1.00	C=missing_solution,D=n5:NA/NA/NA/NA/NA; C=missing_solution,D=n6:1.00/1.00/1.00/1.00/1.00; C=missing_solution,D=n7:NA/NA/NA/NA/NA; C=replaced_solution,D=n4:1.00/1.00/1.00/1.00/1.00; C=replaced_solution,D=n5:1.00/1.00/1.00/1.00/1.00; C=replaced_solution,D=n6:1.00/1.00/1.00/1.00/1.00; C=replaced_solution,D=n7:NA/NA/NA/NA/NA; C=spurious_solution,D=n4:1.00/1.00/1.00/1.00/1.00; C=spurious_solution,D=n5:1.00/1.00/1.00/1.00/1.00; C=spurious_solution,D=n6:1.00/1.00/1.00/1.00/1.00; C=spurious_solution,D=n7:1.00/1.00/1.00/1.00/1.00

system_id	0S: kind,tier=Flag/E/P/R/Fmt	3S: kind,tier=Flag/E/P/R/Fmt	3S+C: kind,tier=Flag/E/P/R/Fmt
gemini-3.1-pro-preview	C=missing_solution,D=n5:1.00/.00/1.00/1.00/.00; C=missing_solution,D=n6:1.00/.00/1.00/1.00/.00; C=missing_solution,D=n7:1.00/.00/1.00/1.00/.00; C=replaced_solution,D=n4:.83/.00/.83/.83/.00; C=replaced_solution,D=n5:1.00/.00/1.00/1.00/.00; C=replaced_solution,D=n6:.67/.00/.67/.67/.00; C=replaced_solution,D=n7:1.00/.00/1.00/1.00/.00; C=spurious_solution,D=n4:1.00/.17/.67/.67/.33; C=spurious_solution,D=n5:.67/.00/.17/.17/.00; C=spurious_solution,D=n6:.33/.00/.17/.17/.00; C=spurious_solution,D=n7:.83/.17/.33/.33/.17	--	--
gemini-3.8-flash	C=missing_solution,D=n5:1.00/.00/1.00/1.00/.00; C=missing_solution,D=n6:.67/.00/.67/.67/.00; C=missing_solution,D=n7:1.00/.00/1.00/1.00/.00; C=replaced_solution,D=n4:1.00/.00/1.00/1.00/.00; C=replaced_solution,D=n5:1.00/.40/1.00/1.00/.40; C=replaced_solution,D=n6:1.00/.33/1.00/1.00/.33; C=replaced_solution,D=n7:1.00/.40/1.00/1.00/.40; C=spurious_solution,D=n4:.83/.00/.83/.83/.00; C=spurious_solution,D=n5:.83/.00/.83/.83/.00; C=spurious_solution,D=n6:.83/.00/.67/.67/.00; C=spurious_solution,D=n7:.20/.00/.20/.20/.00	--	--
gpt-5.6-luna	C=missing_solution,D=n5:1.00/1.00/1.00/1.00/1.00; C=missing_solution,D=n6:1.00/1.00/1.00/1.00/1.00; C=missing_solution,D=n7:NA/NA/NA/NA/NA; C=replaced_solution,D=n4:1.00/.83/.92/1.00/1.00; C=replaced_solution,D=n5:1.00/.80/.90/1.00/1.00; C=replaced_solution,D=n6:1.00/.67/.83/1.00/.67; C=replaced_solution,D=n7:1.00/1.00/1.00/1.00/1.00; C=spurious_solution,D=n4:1.00/.50/.50/.50/1.00; C=spurious_solution,D=n5:1.00/1.00/1.00/1.00/1.00; C=spurious_solution,D=n6:1.00/.50/.50/.50/1.00; C=spurious_solution,D=n7:1.00/1.00/1.00/1.00/1.00	--	--
grok-4.6	C=missing_solution,D=n5:1.00/1.00/1.00/1.00/1.00; C=missing_solution,D=n6:1.00/1.00/1.00/1.00/1.00; C=missing_solution,D=n7:1.00/1.00/1.00/1.00/1.00; C=replaced_solution,D=n4:1.00/1.00/1.00/1.00/1.00; C=replaced_solution,D=n5:1.00/1.00/1.00/1.00/1.00; C=replaced_solution,D=n6:1.00/1.00/1.00/1.00/1.00; C=replaced_solution,D=n7:1.00/1.00/1.00/1.00/1.00; C=spurious_solution,D=n4:1.00/1.00/1.00/1.00/1.00; C=spurious_solution,D=n5:1.00/1.00/1.00/1.00/1.00; C=spurious_solution,D=n6:1.00/1.00/1.00/1.00/1.00; C=spurious_solution,D=n7:1.00/1.00/1.00/1.00/1.00	--	--

system_id	0S: kind,tier=Flag/E/P/R/Fmt	3S: kind,tier=Flag/E/P/R/Fmt	3S+C: kind,tier=Flag/E/P/R/Fmt
mistral-large	C=missing_solution,D=n5:1.00/.00/1.00/.33/.00; C=missing_solution,D=n6:.67/.00/.33/.22/.00; C=missing_solution,D=n7:1.00/.00/.00/.00/.00; C=replaced_solution,D=n4:.67/.00/.00/.00/.00; C=replaced_solution,D=n5:.20/.00/.00/.00/.00; C=replaced_solution,D=n6:.33/.00/.00/.00/.00; C=replaced_solution,D=n7:.80/.00/.00/.00/.00; C=spurious_solution,D=n4:.50/.00/.00/.00/.00; C=spurious_solution,D=n5:.00/.00/.00/.00/.00; C=spurious_solution,D=n6:.17/.00/.00/.00/.00; C=spurious_solution,D=n7:1.00/.00/.00/.00/.00	--	--
muse-spark-1.3	C=missing_solution,D=n5:1.00/1.00/1.00/1.00/1.00; C=missing_solution,D=n6:NA/NA/NA/NA/NA; C=missing_solution,D=n7:1.00/1.00/1.00/1.00/1.00; C=replaced_solution,D=n4:1.00/1.00/1.00/1.00/1.00; C=replaced_solution,D=n5:1.00/1.00/1.00/1.00/1.00; C=replaced_solution,D=n6:1.00/1.00/1.00/1.00/1.00; C=replaced_solution,D=n7:1.00/1.00/1.00/1.00/1.00; C=spurious_solution,D=n4:1.00/1.00/1.00/1.00/1.00; C=spurious_solution,D=n5:1.00/1.00/1.00/1.00/1.00; C=spurious_solution,D=n6:NA/NA/NA/NA/NA; C=spurious_solution,D=n7:1.00/1.00/1.00/1.00/1.00	--	--
phi-4	C=missing_solution,D=n5:1.00/.00/.50/.33/.00; C=missing_solution,D=n6:1.00/.00/.83/.44/.00; C=missing_solution,D=n7:1.00/.00/.00/.00/.00; C=replaced_solution,D=n4:1.00/.00/.03/.33/.00; C=replaced_solution,D=n5:1.00/.00/.23/.60/.00; C=replaced_solution,D=n6:1.00/.00/.00/.00/.00; C=replaced_solution,D=n7:1.00/.00/.50/.50/.00; C=spurious_solution,D=n4:1.00/.00/.00/.00/.00; C=spurious_solution,D=n5:1.00/.00/.00/.00/.00; C=spurious_solution,D=n6:1.00/.00/.00/.00/.00; C=spurious_solution,D=n7:1.00/.00/.00/.00/.00	--	--
qwen3.8-a95b	C=missing_solution,D=n5:1.00/1.00/1.00/1.00/1.00; C=missing_solution,D=n6:NA/NA/NA/NA/NA; C=missing_solution,D=n7:NA/NA/NA/NA/NA; C=replaced_solution,D=n4:1.00/1.00/1.00/1.00/1.00; C=replaced_solution,D=n5:1.00/1.00/1.00/1.00/1.00; C=replaced_solution,D=n6:1.00/1.00/1.00/1.00/1.00; C=replaced_solution,D=n7:1.00/1.00/1.00/1.00/1.00; C=spurious_solution,D=n4:1.00/.50/.50/.50/1.00; C=spurious_solution,D=n5:1.00/1.00/1.00/1.00/1.00; C=spurious_solution,D=n6:1.00/.00/.00/.00/1.00; C=spurious_solution,D=n7:NA/NA/NA/NA/NA	--	--

system_id	0S: kind,tier=Flag/E/P/R/Fmt	3S: kind,tier=Flag/E/P/R/Fmt	3S+C: kind,tier=Flag/E/P/R/Fmt
qwen3.8-max	C=missing_solution,D=n5:1.00/1.00/1.00/1.00/1.00; C=missing_solution,D=n6:NA/NA/NA/NA/NA; C=missing_solution,D=n7:NA/NA/NA/NA/NA; C=replaced_solution,D=n4:1.00/1.00/1.00/1.00/1.00; C=replaced_solution,D=n5:1.00/1.00/1.00/1.00/1.00; C=replaced_solution,D=n6:1.00/1.00/1.00/1.00/1.00; C=replaced_solution,D=n7:NA/NA/NA/NA/NA; C=spurious_solution,D=n4:NA/NA/NA/NA/NA; C=spurious_solution,D=n5:NA/NA/NA/NA/NA; C=spurious_solution,D=n6:NA/NA/NA/NA/NA; C=spurious_solution,D=n7:NA/NA/NA/NA/NA	--	--

Table 22: Focal thinking setting transfer. Full machine-readable values are supplied as CSV.

system_id | task | prompt_mode	B tier:E/F	X tier:E/F	Q tier:E/F	N tier:E/F	O tier:E/F
deepseek-v4-flash-high | D | 3S	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:.75/1.00	n4:.92/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:.90/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00
deepseek-v4-flash-high | D | 3S+C	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:.91/1.00; n5:1.00/1.00; n6:.88/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00
deepseek-v4-flash-high | D | 0S	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:.89/.89
deepseek-v4-flash-high | E | 3S	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:.89/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:.90/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:.83/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:.88/1.00
deepseek-v4-flash-high | E | 3S+C	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:.92/1.00; n6:.89/1.00; n7:1.00/1.00	n4:.92/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:.86/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:.92/1.00; n6:1.00/1.00; n7:1.00/1.00
deepseek-v4-flash-high | E | 0S	n4:.91/.91; n5:.90/.90; n6:.88/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:.91/.91; n6:1.00/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:.91/.91; n7:1.00/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:.91/.91; n6:1.00/1.00; n7:1.00/1.00
deepseek-v4-flash-high | R | 3S	n4:.78/1.00; n5:.67/1.00; n6:1.00/1.00; n7:.75/1.00	n4:.86/1.00; n5:.80/.80; n6:NA/NA; n7:1.00/1.00	n4:.83/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:.75/1.00; n5:1.00/1.00; n6:.50/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:.75/1.00; n6:1.00/1.00; n7:1.00/1.00
deepseek-v4-flash-high | R | 3S+C	n4:.78/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:.91/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:.88/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00
deepseek-v4-flash-high | R | 0S	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:NA/NA	n4:.83/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:.83/.83; n5:.00/1.00; n6:NA/NA; n7:1.00/1.00
deepseek-v4-flash-none | D | 3S	n4:.33/1.00; n5:.42/1.00; n6:.50/1.00; n7:.42/1.00	n4:.33/1.00; n5:.50/1.00; n6:.58/1.00; n7:.58/1.00	n4:.50/1.00; n5:.50/1.00; n6:.50/1.00; n7:.58/1.00	n4:.58/1.00; n5:.50/1.00; n6:.67/1.00; n7:.50/1.00	n4:.50/1.00; n5:.50/1.00; n6:.58/1.00; n7:.50/1.00
deepseek-v4-flash-none | D | 3S+C	n4:.50/1.00; n5:.50/1.00; n6:.67/1.00; n7:.50/1.00	n4:.42/1.00; n5:.42/1.00; n6:.50/1.00; n7:.67/1.00	n4:.50/1.00; n5:.50/1.00; n6:.58/1.00; n7:.42/1.00	n4:.42/1.00; n5:.50/1.00; n6:.42/1.00; n7:.50/1.00	n4:.50/1.00; n5:.33/1.00; n6:.50/1.00; n7:.50/1.00
deepseek-v4-flash-none | D | 0S	n4:.50/1.00; n5:.42/1.00; n6:.58/1.00; n7:.58/1.00	n4:.42/1.00; n5:.50/1.00; n6:.50/1.00; n7:.33/1.00	n4:.50/1.00; n5:.50/1.00; n6:.58/1.00; n7:.50/1.00	n4:.50/1.00; n5:.42/1.00; n6:.67/1.00; n7:.50/1.00	n4:.42/1.00; n5:.25/1.00; n6:.17/1.00; n7:.58/1.00
deepseek-v4-flash-none | E | 3S	n4:.00/1.00; n5:.17/1.00; n6:.25/1.00; n7:.33/1.00	n4:.00/1.00; n5:.25/1.00; n6:.08/1.00; n7:.17/.92	n4:.42/.92; n5:.33/1.00; n6:.25/1.00; n7:.36/.91	n4:.25/1.00; n5:.08/.42; n6:.25/.50; n7:.33/1.00	n4:.08/1.00; n5:.17/.92; n6:.17/1.00; n7:.33/1.00
deepseek-v4-flash-none | E | 3S+C	n4:.08/1.00; n5:.08/1.00; n6:.33/.92; n7:.17/.92	n4:.00/1.00; n5:.25/.92; n6:.17/1.00; n7:.33/.92	n4:.25/.75; n5:.18/.91; n6:.42/1.00; n7:.33/1.00	n4:.17/.92; n5:.33/.58; n6:.25/.75; n7:.33/1.00	n4:.08/1.00; n5:.17/1.00; n6:.42/1.00; n7:.17/.83
deepseek-v4-flash-none | E | 0S	n4:.17/.92; n5:.25/.67; n6:.25/.67; n7:.17/.42	n4:.00/1.00; n5:.08/.67; n6:.08/1.00; n7:.27/.82	n4:.25/.83; n5:.27/.82; n6:.18/.64; n7:.25/.83	n4:.08/.58; n5:.17/.67; n6:.25/.67; n7:.17/.67	n4:.17/1.00; n5:.25/.83; n6:.25/.92; n7:.08/.50
deepseek-v4-flash-none | R | 3S	n4:.08/1.00; n5:.00/1.00; n6:.00/.92; n7:.00/1.00	n4:.08/1.00; n5:.00/1.00; n6:.00/.92; n7:.00/1.00	n4:.17/1.00; n5:.08/1.00; n6:.00/1.00; n7:.08/1.00	n4:.00/1.00; n5:.00/1.00; n6:.00/1.00; n7:.00/1.00	n4:.00/1.00; n5:.08/1.00; n6:.00/.92; n7:.00/1.00
deepseek-v4-flash-none | R | 3S+C	n4:.08/1.00; n5:.08/1.00; n6:.00/.92; n7:.08/1.00	n4:.08/1.00; n5:.00/1.00; n6:.08/1.00; n7:.08/1.00	n4:.42/1.00; n5:.08/1.00; n6:.00/1.00; n7:.08/1.00	n4:.17/1.00; n5:.00/1.00; n6:.08/1.00; n7:.08/1.00	n4:.08/1.00; n5:.00/1.00; n6:.17/1.00; n7:.08/1.00
deepseek-v4-flash-none | R | 0S	n4:.25/1.00; n5:.00/1.00; n6:.00/1.00; n7:.08/1.00	n4:.08/1.00; n5:.00/1.00; n6:.00/1.00; n7:.08/1.00	n4:.08/1.00; n5:.00/1.00; n6:.00/.92; n7:.09/1.00	n4:.17/1.00; n5:.00/1.00; n6:.00/.92; n7:.00/.92	n4:.17/1.00; n5:.00/1.00; n6:.00/.92; n7:.00/1.00
deepseek-v4-pro-high | D | 3S	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00
deepseek-v4-pro-high | D | 3S+C	n4:1.00/1.00; n5:.83/.83; n6:.75/.75; n7:1.00/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:.83/.83; n7:1.00/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00
deepseek-v4-pro-high | D | 0S	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:.80/.80; n7:1.00/1.00	n4:.90/.90; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:.80/.80; n7:1.00/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00
deepseek-v4-pro-high | E | 3S	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00
deepseek-v4-pro-high | E | 3S+C	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:.92/.92; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:.88/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00

system_id | task | prompt_mode	B tier:E/F	X tier:E/F	Q tier:E/F	N tier:E/F	O tier:E/F
deepseek-v4-pro-high | E | 0S	n4:.90/.90; n5:1.00/1.00; n6:1.00/1.00; n7:.88/.88	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:.91/.91; n5:.86/.86; n6:1.00/1.00; n7:1.00/1.00	n4:.75/.75; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:.88/1.00
deepseek-v4-pro-high | R | 3S	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:.80/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:NA/NA	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:NA/NA	n4:1.00/1.00; n5:1.00/1.00; n6:.67/.67; n7:1.00/1.00
deepseek-v4-pro-high | R | 3S+C	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00
deepseek-v4-pro-high | R | 0S	n4:1.00/1.00; n5:NA/NA; n6:1.00/1.00; n7:NA/NA	n4:1.00/1.00; n5:1.00/1.00; n6:1.00/1.00; n7:1.00/1.00	n4:1.00/1.00; n5:NA/NA; n6:NA/NA; n7:NA/NA	n4:1.00/1.00; n5:1.00/1.00; n6:NA/NA; n7:1.00/1.00	n4:1.00/1.00; n5:NA/NA; n6:.50/.50; n7:NA/NA
deepseek-v4-pro-none | D | 3S	n4:.42/1.00; n5:.33/1.00; n6:.50/1.00; n7:.50/1.00	n4:.42/1.00; n5:.50/1.00; n6:.50/1.00; n7:.50/1.00	n4:.45/1.00; n5:.50/1.00; n6:.50/1.00; n7:.42/1.00	n4:.33/1.00; n5:.33/1.00; n6:.33/1.00; n7:.58/1.00	n4:.58/1.00; n5:.42/1.00; n6:.25/1.00; n7:.42/1.00
deepseek-v4-pro-none | D | 3S+C	n4:.50/1.00; n5:.25/1.00; n6:.58/1.00; n7:.50/1.00	n4:.50/1.00; n5:.33/1.00; n6:.42/1.00; n7:.50/1.00	n4:.50/1.00; n5:.50/1.00; n6:.50/1.00; n7:.50/1.00	n4:.45/1.00; n5:.25/1.00; n6:.50/1.00; n7:.42/1.00	n4:.58/1.00; n5:.33/1.00; n6:.58/1.00; n7:.50/1.00
deepseek-v4-pro-none | D | 0S	n4:.42/.92; n5:.42/1.00; n6:.58/1.00; n7:.45/1.00	n4:.50/1.00; n5:.50/1.00; n6:.50/1.00; n7:.50/1.00	n4:.50/1.00; n5:.50/1.00; n6:.45/1.00; n7:.42/1.00	n4:.67/1.00; n5:.42/1.00; n6:.42/1.00; n7:.58/1.00	n4:.50/1.00; n5:.42/1.00; n6:.25/1.00; n7:.50/1.00
deepseek-v4-pro-none | E | 3S	n4:.00/1.00; n5:.00/1.00; n6:.00/1.00; n7:.00/.92	n4:.00/1.00; n5:.00/1.00; n6:.00/.92; n7:.00/1.00	n4:.33/.89; n5:.08/1.00; n6:.17/1.00; n7:.08/1.00	n4:.00/1.00; n5:.00/.83; n6:.00/1.00; n7:.00/1.00	n4:.00/1.00; n5:.00/1.00; n6:.00/.92; n7:.00/1.00
deepseek-v4-pro-none | E | 3S+C	n4:.00/1.00; n5:.00/1.00; n6:.00/1.00; n7:.00/1.00	n4:.00/1.00; n5:.00/1.00; n6:.00/1.00; n7:.00/1.00	n4:.50/1.00; n5:.08/.92; n6:.17/1.00; n7:.00/1.00	n4:.00/1.00; n5:.00/.92; n6:.00/1.00; n7:.00/1.00	n4:.00/1.00; n5:.00/1.00; n6:.00/1.00; n7:.00/1.00
deepseek-v4-pro-none | E | 0S	n4:.00/.17; n5:.00/.17; n6:.00/.08; n7:.00/.08	n4:.00/.17; n5:.00/.00; n6:.00/.09; n7:.00/.08	n4:.09/.36; n5:.00/.17; n6:.00/.00; n7:.00/.08	n4:.00/.00; n5:.00/.08; n6:.00/.00; n7:.00/.00	n4:.00/.17; n5:.00/.08; n6:.00/.00; n7:.00/.00
deepseek-v4-pro-none | R | 3S	n4:.25/1.00; n5:.00/1.00; n6:.00/1.00; n7:.08/1.00	n4:.27/1.00; n5:.00/1.00; n6:.00/1.00; n7:.08/1.00	n4:.17/1.00; n5:.08/1.00; n6:.08/1.00; n7:.08/1.00	n4:.17/1.00; n5:.00/1.00; n6:.00/1.00; n7:.10/1.00	n4:.18/1.00; n5:.00/1.00; n6:.00/1.00; n7:.08/.92
deepseek-v4-pro-none | R | 3S+C	n4:.18/1.00; n5:.00/1.00; n6:.00/1.00; n7:.08/1.00	n4:.09/1.00; n5:.00/1.00; n6:.00/1.00; n7:.08/1.00	n4:.00/1.00; n5:.00/1.00; n6:.08/1.00; n7:.08/1.00	n4:.00/1.00; n5:.00/1.00; n6:.00/1.00; n7:.09/1.00	n4:.00/1.00; n5:.00/1.00; n6:.00/1.00; n7:.08/1.00
deepseek-v4-pro-none | R | 0S	n4:.17/1.00; n5:.00/.92; n6:.00/1.00; n7:.09/1.00	n4:.17/1.00; n5:.00/.91; n6:.00/1.00; n7:.08/.92	n4:.25/1.00; n5:.08/1.00; n6:.08/.92; n7:.08/.75	n4:.08/.92; n5:.00/.83; n6:.00/.91; n7:.10/1.00	n4:.08/1.00; n5:.00/1.00; n6:.00/1.00; n7:.08/1.00

Table 23: Composed analysis matrix: model registry, canonical and actual model slugs, tokens, median latency in seconds, and cost; the route field is recorded on every response but the serving stack returned no provider value, so that column reads unavailable throughout. The final row is the column-sum total for the composed population, not the four-run spend. Full machine-readable values are supplied as CSV.

system_id	requested_slug	canonical_slug	actual_models	providers	requests	prompt_tokens	completion_tokens	median_latency_seconds	cost_usd	reasoning
claude-sonnet-5	anthropic/claude-sonnet-5	anthropic/claude-sonnet-5-20260630	anthropic/claude-sonnet-5	unavailable	336	365724	539983	14.839	6.1313	high
deepseek-v4-flash-high	deepseek/deepseek-v4-flash-0731	deepseek/deepseek-v4-flash-20260731	deepseek/deepseek-v4-flash-0731	unavailable	2160	2057064	3744987	26.4425	0.6551	high
deepseek-v4-flash-none	deepseek/deepseek-v4-flash-0731	deepseek/deepseek-v4-flash-20260731	deepseek/deepseek-v4-flash-0731	unavailable	2160	1909750	81038	2.8625	0.1536	none
deepseek-v4-pro-high	deepseek/deepseek-v4-pro-0813	deepseek/deepseek-v4-pro-20260813	deepseek/deepseek-v4-pro-0813	unavailable	2160	2031447	3931111	20.218	13.5366	high
deepseek-v4-pro-none	deepseek/deepseek-v4-pro-0813	deepseek/deepseek-v4-pro-20260813	deepseek/deepseek-v4-pro-0813	unavailable	2160	1909890	135228	2.653	1.9294	none
gemini-3.1-pro-preview	google/gemini-3.1-pro-preview	google/gemini-3.1-pro-preview-20260219	google/gemini-3.1-pro-preview	unavailable	336	294267	555497	12.1885	7.2545	high
gemini-3.8-flash	google/gemini-3.8-flash	google/gemini-3.8-flash-20260902	google/gemini-3.8-flash	unavailable	336	292006	601447	12.5715	2.4757	high
gpt-5.6-luna	openai/gpt-5.6-luna	openai/gpt-5.6-luna-20260709	openai/gpt-5.6-luna	unavailable	336	276788	196885	6.533	0.3004	high
grok-4.6	x-ai/grok-4.6	x-ai/grok-4.6-20260810	x-ai/grok-4.6	unavailable	336	344468	613278	27.8715	4.3173	high
mistral-large	mistralai/mistral-large	mistralai/mistral-large	mistralai/mistral-large	unavailable	336	282849	21426	2.336	0.6551	default
muse-spark-1.3	meta/muse-spark-1.3	meta/muse-spark-1.3-20260902	meta/muse-spark-1.3	unavailable	336	281278	554147	9.231	2.7048	high
phi-4	microsoft/phi-4	microsoft/phi-4	microsoft/phi-4	unavailable	336	278132	230124	12.8265	0.0517	default
qwen3.8-a95b	qwen/qwen3.8-2.4t-a95b	qwen/qwen3.8-2.4t-a95b-20260812	qwen/qwen3.8-2.4t-a95b	unavailable	336	301139	558890	17.4385	3.9805	high
qwen3.8-max	qwen/qwen3.8-max-0902	qwen/qwen3.8-max-20260902	qwen/qwen3.8-max-0902	unavailable	336	312843	509675	27.994	3.6837	high
TOTAL (composed matrix)					12000	10937645	12273716		47.8298	

G Paired Protocol Correction and Third-Task Construct Evidence

Table 24 turns the completion-budget correction into a paired measurement. v2 replaced complete matrices, so each listed cell was issued once under the v1 caps (96/1,024/1,152) and once under the v2 caps (1,280/2,560/2,560) for the same configuration, instance, view, task, and prompt. Rows give the v1 state and columns the v2 state, where a cell is *censored* when its completion usage reached its task cap without producing a parseable object, an *interface* failure when it is observable but violates the whole-content JSON contract, and otherwise *wrong* or *correct* against the exact oracle. The delta rows give the paired change with a base-cluster interval over the 48 base identifiers. The replacement is an intervention of convenience rather than a randomised one: it is confounded with time and with anything the serving stack changed in between, so the table quantifies how much the correction moved the matrix without attributing the whole movement to the cap.

Table 25 pairs the third task with enumeration on the same configuration and instance. Because both endpoints are read on the same cells, the comparison cannot be explained by differences in support or by which instances each endpoint reached. Its columns count the instances where both responses satisfy the strict contract (the denominator of the identical-set rate, printed per configuration and in pooled total), the instances where both endpoints return the same set, the instances where each endpoint alone is correct, the discordant pairs entering the exact McNemar test, and the recovered error flags that came back `false`; the last column counts strict predictions that echoed the corrupted candidate that was shown in the prompt. Table 26 reports the structural quantities of the matched ladder: matching variable count, clause count, and clause-length profile does not hold the free and constant variable counts fixed, which is why the extension is reported as a descriptive stress test of exact-set scaling under matched visible dimensions rather than as an isolated effect of solution count.

Table 27 provides the previously missing per-configuration protocol-correction breakdown. Each row uses the same complete pairing key as Table 24 and reports both directions of censoring movement together with the configuration's strict-correct change and base-cluster interval; it therefore exposes heterogeneous correction magnitudes without treating the non-random cross-time intervention as a causal cap effect.

Table 24: Paired protocol correction. Cells are matched on configuration, instance, view, task, and prompt mode, so every cell listed here was issued once under the v1 caps (96/1,024/1,152) and again under the v2 caps (1,280/2,560/2,560). Rows give the v1 state and the columns the v2 state; the delta rows report the paired change with a base-cluster interval. The intervention is non-random and crosses time, so these are paired descriptions of the correction and not an isolated effect of the completion cap. Full machine-readable values are supplied as CSV.

task	v1 state	n	v2 censored	v2 interface	v2 wrong	v2 correct	observed delta pp	strict delta pp	strict delta CI pp
decision	censored	2263	941	130	19	1173			
decision	interface	284	2	233	2	47			
decision	wrong	0	0	0	0	0			
decision	correct	189	16	1	0	172			
decision	delta v2 minus v1	2736					47.66	43.97	[+37.57, +50.48]
enumeration	censored	1801	402	114	14	1271			
enumeration	interface	389	3	366	1	19			
enumeration	wrong	38	0	7	6	25			
enumeration	correct	508	10	46	21	431			
enumeration	delta v2 minus v1	2736					50.66	45.25	[+42.00, +48.36]
repair	censored	1607	1144	49	19	395			
repair	interface	118	10	98	1	9			
repair	wrong	6	0	1	1	4			
repair	correct	141	21	3	6	111			
repair	delta v2 minus v1	1872					23.08	20.19	[+15.97, +24.68]
all	censored	5671	2487	293	52	2839			
all	interface	791	15	697	4	75			
all	wrong	44	0	8	7	29			
all	correct	838	47	50	27	714			
all	delta v2 minus v1	7344					42.51	38.39	[+34.55, +42.10]

Table 25: Third-task construct evidence. The candidate-conditioned task is paired with enumeration on the same configuration and instance, so the two endpoints cannot be separated by support. 'both parsed' counts instances where both responses satisfy the strict contract and is the denominator of the identical-set rate; 'identical returned set' counts instances where both parsed responses return the same set. All four correctness columns and the McNemar p are restricted to those parsed pairs, so an unparseable answer is missing rather than scored as incorrect; no discordant count can exceed the row's both-parsed value, and the per-row McNemar is NA where no complete pair exists. 'flag returned false' counts recovered error flags that were false, which the prompt's single-error guarantee makes a constant. The task is therefore reported as conditioned re-enumeration and not as detection, localisation, or repair. Full machine-readable values are supplied as CSV.

configuration	paired cells	both parsed	identical returned set	identical rate	repair correct (parsed)	enumeration correct (parsed)	repair only correct (parsed)	enumeration only correct (parsed)	McNemar p	flag returned false
claude-sonnet-5	15	1	1	100	1	1	0	0	1	0
deepseek-v4-flash-high	6	4	4	100	4	4	0	0	1	0
deepseek-v4-pro-high	2	2	2	100	2	2	0	0	1	0
gemini-3.1-pro-preview	48	0	0	NA	0	0	0	0	NA	0
gemini-3.8-flash	47	1	1	100	1	1	0	0	1	0
gpt-5.6-luna	45	44	31	70.45	36	36	6	6	1	0
grok-4.6	48	48	47	97.92	48	47	1	0	1	0
mistral-large	48	0	0	NA	0	0	0	0	NA	0
muse-spark-1.3	24	24	24	100	24	24	0	0	1	0
phi-4	42	0	0	NA	0	0	0	0	NA	0
qwen3.8-a95b	15	15	14	93.33	14	15	0	1	1	0
qwen3.8-max	8	8	8	100	8	8	0	0	1	0
all configurations	348	147	132	89.8	138	138	7	7	1	0

Table 26: Structural drift of the matched solution-count ladder. The six variants per family hold variable count, clause count, and clause-length profile fixed, but releasing a selector from zero to five also changes the free and constant variable counts; clause count, mean variable occurrence, and the node count of a plain unit-propagating DPLL search do not move. The fixed DPLL proxy does not remove the autoregressive burden of producing 32 unique sorted strings, so the ladder is reported as a descriptive stress test under matched visible dimensions rather than an isolated search-difficulty effect. Full machine-readable values are supplied as CSV.

solution count K	released selectors	families	clauses	free variables	constant variables	mean variable occurrence	DPLL nodes
1	0	24	16	0	8	4	3
2	1	24	16	1	7	4	3
4	2	24	16	2	6	4	3
8	3	24	16	3	5	4	3
16	4	24	16	4	4	4	3
32	5	24	16	5	3	4	3

Table 27: Per-configuration paired protocol correction. Every row matches v1 and v2 on configuration, instance, view, task, and prompt mode and reports both directions of censoring movement, censored-to-correct recovery, and the strict-correct change with a base-cluster interval. The intervention is non-random and crosses time, so these are configuration-specific paired descriptions rather than causal cap effects. Full machine-readable values are supplied as CSV.

configuration	paired n	v1 censored	v2 censored	censored to observed	censored to correct	observed to censored	v1 to v2 correct	observed delta pp	strict delta pp [95% CI]
claude-sonnet-5	336	316	150	167	52	1	11 to 54	49.4	+12.80 [+8.04, +18.15]
deepseek-v4-flash-high	2160	1970	903	1099	1065	32	182 to 1218	49.4	+47.96 [+43.19, +52.45]
deepseek-v4-pro-high	2160	2087	1106	994	977	13	55 to 1036	45.42	+45.42 [+38.98, +51.67]
gemini-3.1-pro-preview	336	0	0	0	0	0	35 to 11	0	-7.14 [-10.12, -4.17]
gemini-3.8-flash	336	6	3	6	1	3	4 to 55	0.89	+15.18 [+10.42, +19.94]
gpt-5.6-luna	336	169	3	166	146	0	136 to 283	49.4	+43.75 [+38.39, +49.11]
grok-4.6	336	0	0	0	0	0	336 to 335	0	-0.30 [-0.89, +0.00]
muse-spark-1.3	336	332	139	193	192	0	4 to 196	57.44	+57.14 [+50.89, +63.39]
phi-4	336	194	45	151	0	2	0 to 0	44.35	+0.00 [+0.00, +0.00]
qwen3.8-a95b	336	280	95	196	195	11	56 to 239	55.06	+54.46 [+46.73, +62.20]
qwen3.8-max	336	317	105	212	211	0	19 to 230	63.1	+62.80 [+56.55, +68.75]