INVARIANTBENCH: Can Large Language Models Exhibit Inherent Reasoning Across Equivalent Transformations?

Azmine Toushik Wasi [♦1 2] **Mahir Absar Khan** [♦◊1 2] **Abdur Rahman** [♦◊1 2] **Wahid Faisal** [◊1 2] **Sukanta Saha** [◊1 2] **Saimon Bhuiyan** [◊1 3] **Mahdiya Rahman Sukanya** [◊1 3] **Rahatun Nesa Priti** [◊1 2] **Md. Iqramul Hoque** [◊1 2] **Munem Shahriar** [◊1 4] **Raima Islam** [1 5] **Sanatan Sushil** [◊1 6] **Shahriyar Zaman Ridoy** [◊1 7] **Kazi Rajwan Sultan** [◊8] **MD Shafikul Islam** [1 9] **Md Manjurul Ahsan** [1 10] **Md Rizwan Parvez** [♦11]

Abstract

Reasoning is often attributed to large language models (LLMs), yet it remains unclear whether they operate over underlying semantics or rely on surface-form patterns. Existing benchmarks evaluate correctness on fixed problem instances, but overlook a fundamental property of reasoning: invariance under semantics-preserving transformations. If a model truly understands a problem, its predictions should remain consistent across equivalent representations. We introduce INVARIANTBENCH, a benchmark of 1,200 seed problems $\times$ 4 invariant forms, spanning 16 tasks across 3 reasoning families and 12 fine-grained invariance axes. Each problem is paired with multiple semantically equivalent variants under a strict invariance contract, enabling evaluation beyond accuracy to measure consistency across representations. Experiments on 15 frontier and open-weight LLMs reveal a persistent invariance gap: accuracy drops of $1-5\%$ for strong models and >10 pp for smaller systems under transformation, with full-consistency rates remaining below 5% for most models, compared to $>48\%$ for the best systems and $>98\%$ for humans. These results show that high base accuracy substantially overestimates reasoning ability, and establish invariance as a necessary axis for evaluating and improving robust language understanding.

[♦]Core Contributors [◊]Data Collection and Annotation Contributors [1]Computational Intelligence and Operations Laboratory (CIOL) [2]Shahjalal University of Science and Technology, Bangladesh [3]University of Dhaka, Bangladesh [4]BRAC University, Bangladesh [5]Harvard University, USA [6]KSRM Steel Plant Limited, Bangladesh [7]North South University, Bangladesh [8]Faridpur Engineering College, Bangladesh [9]Louisiana State University, USA [10]University of Oklahoma, USA [11]Qatar Computing Research Institute (QCRI), Qatar. Correspondence to: Azmine Toushik Wasi <azminetoushik.wasi@gmail.com>.

The 3rd *AI for Math Workshop* at the 43rd International Conference on Machine Learning (ICML), Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

1. Introduction

Large language models (LLMs) are now routinely described as performing *reasoning*, multi-step deductive, mathematical, algorithmic, and semantic inference across a wide range of benchmarks (Wei et al., 2022b; Bubeck et al., 2023; OpenAI et al., 2024; Anthropic, 2024). Headline scores on GSM8K (Cobbe et al., 2021), MATH (Hendrycks et al., 2021b), BIG-BENCH HARD (Suzgun et al., 2022), and MMLU (Hendrycks et al., 2021a) have risen so quickly that several recent reports interpret them as evidence for emergent general-purpose reasoning (Bubeck et al., 2023; Wei et al., 2022a). A natural question follows: *do these scores measure reasoning, or do they measure pattern-matching against a particular surface form of the problem?* If the former, a model that solves a problem under one phrasing should solve it under any semantically-equivalent phrasing. If the latter, we should expect surface perturbations, even those preserving every mathematical and logical entailment, to systematically degrade performance.

The empirical evidence is increasingly that the latter possibility is non-trivial. Mirzadeh *et al.* (Mirzadeh et al., 2025) show that simply replacing the proper nouns and numerical values in GSM8K items reduces accuracy across every model they evaluate, with degradations as large as 65% when a single, semantically-irrelevant clause is added. Turpin *et al.* (Turpin et al., 2023) establish that chain-of-thought rationales can be systematically biased by features the rationale itself does not mention, undermining the interpretation of the chain as a window into the model's reasoning. Ribeiro *et al.* (Ribeiro et al., 2020) and Naik *et al.* (Naik et al., 2018) have long documented brittleness to surface paraphrase in NLP systems. Concerns about benchmark contamination (Xu et al., 2024; Sainz et al., 2023; Golchin & Surdeanu, 2024) sharpen the worry: if a model can recall an answer because it has seen the exact item, then accuracy on the original instance is uninformative about its

competence on *any* instance. The community has responded with dynamic benchmarks (Kiela et al., 2021; et al., 2023), contamination-resistant evaluations (Zhu et al., 2024; White et al., 2024), and proposals for testing reasoning under controlled perturbations (Mirzadeh et al., 2025; Li et al., 2024; Gulati et al., 2024). None of these efforts has produced a unified, multi-axis framework that makes *representational invariance* a first-class evaluation primitive.

Reasoning, in the classical sense, is closed under semantically-preserving transformations: an inference is valid in virtue of the relations between its propositions, not in virtue of how those propositions are spelled (Russell & Norvig, 2010; Marcus, 2020). Renaming the variables of a logical formula, permuting independent premises, switching from a fraction to its decimal expansion, or rephrasing a deduction as an equivalent contrapositive, none of these should change the answer for an agent that operates over the underlying semantics. Conversely, any perturbation that does change the answer reveals a feature of the model's behaviour that is *not* a feature of the problem. *Invariance is therefore a necessary condition for reasoning, and a measurable one.* It is also *robust to data contamination*: even if an item is in the training set, its semantics-preserving variants need not be, and so the gap between accuracy on the original and accuracy on its variants provides a clean lower bound on reasoning that is not memorisation.

We introduce INVARIANTBENCH, a benchmark of seed reasoning problems paired with three semantics-preserving transformations drawn from twelve invariance axes (Table 6). The transformations fall into three families: *symbolic and linguistic invariance* (variable renaming, entity relabeling, style/register shifts, synonym or semantic paraphrasing), *structural and representational invariance* (premise reordering, redundant restatement, distractor injection or removal, equivalent logical or mathematical reformulations, and representation or encoding swaps), and *abstraction and information-structure invariance* (abstraction-level shifts and perspective or role reframing). All transformations preserve the underlying semantics, including the correct answer and entailment structure, as defined in § 2. INVARIANTBENCH contains 1,200 seed problems across three reasoning families: Formal Logic and Discrete Structure, Quantitative Reasoning and Optimization, and Relational, Planning, Algorithmic and Semantic Reasoning, further divided into 15 sub-types spanning logic, constraint satisfaction, algebraic systems, probabilistic reasoning, graph search, isomorphism, and language-based reasoning. Each seed has three invariant variants, yielding 4,800 evaluated items per model. Evaluation uses an LLM-judged rubric grounded in verifiable solutions, supporting both multiple-choice and open-ended formats. Our core contributions are:

1. We introduce a **formal notion of representational invariance** for LLM evaluation, together with a semantic-preservation contract that all transformations in our work satisfy (§ 2).
2. We present **INVARIANTBENCH**, a large-scale reasoning benchmark of 1,200 seed problems $\times$ 4 invariant forms, spanning 16 tasks across 3 reasoning families and 12 fine-grained invariance axes, with verifiable answers and an automated evaluation protocol (§3).
3. We conduct a **systematic empirical study** of 15 frontier and open-weight LLMs, revealing a persistent invariance gap: accuracy drops of $1-5\%$ for strong models and $> 10\%$ for smaller ones, with very low full-consistency rates ($< 5\%$ for most models), showing that high base accuracy overestimates true reasoning ability (§4).
4. We provide **comprehensive failure analysis and diagnostic interventions**, introducing eight main failure modes from model outputs that link errors to invariant axes and localize breakdowns across perception, transformation, and composition, and further evaluating **SFT** and **GRPO**, showing SFT induces memorization while GRPO improves reasoning and cross-form generalization without closing the invariance gap.

2. Preliminaries

Reasoning under Representational Invariance. A reasoning problem is a pair $(P, a^\star)$ where P is a problem statement and $a^\star \in \mathcal{Y}$ is the verifiable ground-truth answer. A surface form $s \in \Sigma^\star$ denotes P under a fixed semantics $[\![\cdot]\!]$, i.e., $[\![s]\!] = (P, a^\star)$. Two forms s, s' are semantically equivalent, $s \equiv s'$, if $[\![s]\!] = [\![s']\!]$. A transformation $\tau : \Sigma^\star \to \Sigma^\star$ is semantics-preserving if

$$[\![\tau(s)]\!] = [\![s]\!] = (P, a^\star), \tag{1}$$

i.e., $\tau(s) \equiv s$. INVARIANTBENCH defines a fixed set of $K = 12$ transformations $\mathcal{T}$ (Table 6). Each seed s_0 yields variants $s_k = \tau_k(s_0)$. A model $M : \Sigma^\star \to \mathcal{Y}$ predicts answers from surface forms. For a seed and its variants, we define $\hat{a}_M(P) = \big(M(s_0), M(s_1), \ldots, M(s_K)\big)$. Representation invariance requires

$$M(s) = M(s') \quad \forall\, s \equiv s'. \tag{2}$$

This property is independent of correctness, isolating sensitivity to representation from accuracy.
We identify two failure modes: (i) *accuracy fragility*, where $M(s_0) = a^\star$ but $M(s_k) \neq a^\star$ for some k; and (ii) *inconsistency*, where predictions vary across equivalent forms even when some outputs match $a^\star$. INVARIANTBENCH measures both, separating stability from seed-specific performance.

The Twelve Invariance Axes. The twelve transformations used in INVARIANTBENCH are grouped into three families: symbolic & linguistic, structural & representational, and

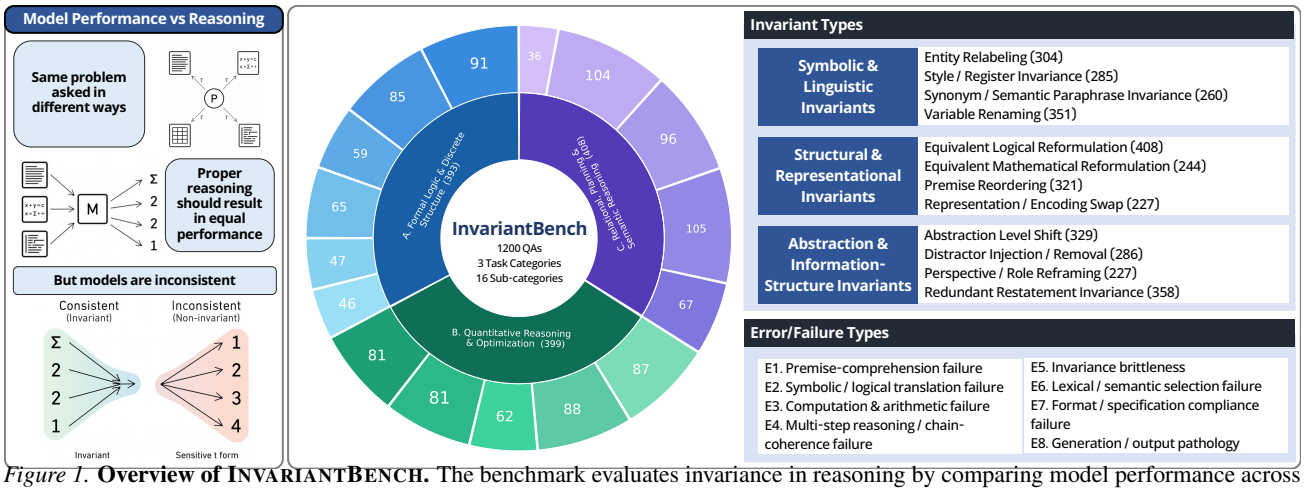

Figure 1. **Overview of INVARIANTBENCH.** The benchmark evaluates invariance in reasoning by comparing model performance across semantically equivalent transformations of the same problem (left), and consists of 1,200 seed questions spanning 3 reasoning families and 16 subcategories (center). It applies 12 controlled invariance transformations grouped into symbolic, structural, and abstraction-level shifts, and analyzes failures through 8 distinct error modes capturing where reasoning breaks under representation change (right).

abstraction & information-structure, as outlined in Figure 1 and Table 6. The symbolic & linguistic family includes entity relabeling, style/register variation, synonym/paraphrase substitution, and variable renaming. The structural & representational family covers equivalent logical reformulation, equivalent mathematical reformulation, premise reordering, and representation/encoding swaps. The abstraction & information-structure family includes abstraction level shifts, distractor injection/removal, perspective/role reframing, and redundant restatement invariance. Each transformation $\tau \in \mathcal{T}$ is designed to preserve semantics (Eq. equation 1) while altering surface form, structure, or information organisation, and all variants are human-verified to ensure answer equivalence.

Task Design and Distribution. INVARIANTBENCH spans three principal reasoning families: Formal Logic & Discrete Structure (393), Quantitative Reasoning & Optimization (399), and Relational, Planning, Algorithmic & Semantic Reasoning (408), ensuring balanced coverage across symbolic, numerical, and structural regimes. Each family is further decomposed into fine-grained sub-types capturing key reasoning primitives, including logic-based constraints (e.g., multi-quantifier and set algebra), algebraic and probabilistic systems, and graph- and language-based reasoning tasks. This hierarchical design isolates distinct reasoning capabilities while maintaining controlled overlap across related modes, enabling systematic analysis of performance degradation under invariant transformations. Full task design details, category definitions, and explanations are provided in § B.

Metrics: Accuracy, Consistency, and Stability. To evaluate whether models exhibit genuine representation-invariant reasoning rather than surface-level pattern matching, we use a compact set of metrics operating at both aggregate and conditional levels (full details in § F). Each sample is evaluated across four invariant variants, forming a correctness vector $\mathbf{x}_i = (x_{i,v})_{v\in\mathcal{V}} \in \{0,1\}^4$ with $c_i = \sum_{v\in\mathcal{V}} x_{i,v}$. At the *aggregate level*, we measure the invariance degradation gap $\Delta = \mathrm{Acc}(\text{base}) - \frac{1}{3}\sum_{j\in\mathcal{I}} \mathrm{Acc}(j)$, which quantifies sensitivity to representation shifts beyond standard accuracy. We further characterize robustness via the invariant mass index $\mathrm{IMI} = P(c = 4)$, capturing the fraction of samples solved correctly under all transformations. To capture *conditional stability*, we define the conditional invariance retention $\mathrm{CIR} = P\left(\bigwedge_{j\in\mathcal{I}} x_{i,j} = 1 \;\middle|\; x_{i,\text{base}} = 1\right)$, and the invariance collapse rate $\mathrm{ICR} = P\left(x_{i,\text{base}} = 1 \wedge \sum_{j\in\mathcal{I}} x_{i,j} \leq 1\right)$, which together distinguish transferable correctness from brittle, non-generalizing solutions. Pairwise dependence is measured via $\mathrm{PIDI}(j,k) = P(x_{i,j} = 0 \mid x_{i,k} = 0)$, indicating whether failures are independent or structurally coupled. We also summarize *structural behavior* using consistency entropy $\mathrm{CE} = \frac{1}{N}\sum_{i=1}^{N} H\left(\frac{c_i}{4}\right)$, and bilateral symmetry $\mathrm{BSS} = 1 - \frac{|F^+ - F^-|}{F^+ + F^-}$, which together distinguish deterministic, symmetric reasoning from unstable or directionally biased behavior. Finally, we define invariance sensitivity through degradation gradients $g_{12} = \mathrm{Acc}(\mathcal{I}_1) - \mathrm{Acc}(\mathcal{I}_2)$, $g_{23} = \mathrm{Acc}(\mathcal{I}_2) - \mathrm{Acc}(\mathcal{I}_3)$, where violations of monotonicity indicate non-semantic artifacts in representation shifts. Together, these metrics define a unified evaluation framework that captures not only correctness, but also stability, transferability, and structural coherence under semantically equivalent transformations.

Scope and limitations of the contract. We focus on *open-ended* evaluation, where the answer space $\mathcal{Y}$ admits multiple mathematically equivalent forms (e.g., $\frac{12}{25} \equiv 0.48$). To ensure that Acc reflects semantic correctness rather than no-

Base: Look at this long rule: (p ∧ ¬q) ∨ q ∨ (¬p ∧ q). Which of these short rules means the exact same thing?
Answer: p ∨ q; **Task Type:** A. Formal Logic & Discrete Structure, **Task Sub-Type:** A5. Inductive Rule + Formal Deduction

Symbolic and Linguistic Invariant Type: Entity Relabeling
Question: A traffic light controller outputs GREEN when: (Road-clear AND no-pedestrian) OR pedestrian-present OR (road-not-clear AND pedestrian-present). Using R for road-clear and W for pedestrian waiting, this is (R∧¬W)∨W∨(¬R∧W). What is the simplest equivalent Boolean expression?

Structural and Representational Invariant Type: Premise Reordering
Question: Look at this long rule: q ∨ (¬p ∧ q) ∨ (p ∧ ¬q). Which of these short rules means the exact same thing?

Abstraction and Information-Structure Invariant Type: Abstraction Level Shift
Question: A machine turns on if ('Power is on' AND 'No error') OR 'Error is present' OR ('Power is off' AND 'Error is present'). If P represents 'Power is on' and Q represents 'Error is present', which simplified boolean rule represents the exact same machine state?

Base: Three volunteers produce kits at constant rates. Worker A+B produce 33 per hour, B+C produce 36, and A+C produce 25. In one run, A works 3 hours, B works 2 hours, and C works 3 hours, but worker C loses 4 units/hour due to overhead. How many kits are produced?
Answer: 107; **Task Type:** B. Quantitative Reasoning & Optimization, **Task Sub-Type:** B4. Interdependent Equation Systems with Structural Constraints

Symbolic and Linguistic I.Type: Style / Register Invariance
Question: Three volunteers produce kits at constant rates. Let these rates be a, b, and c. Given that a+b=33 per hour, b+c=36 per hour, and a+c=25 per hour, calculate the total production if volunteer A works 3 hours, B works 2 hours, and C works 3 hours, keeping in mind that C's effective rate is reduced by 4 units per hour due to overhead.

Structural and Representational I. Type: Representation / Encoding Swap
Question: Production data: (A+C)=25 kits/hr, (A+B)=33 kits/hr, (B+C)=36 kits/hr. Labor allocation: A: 3 hours, B: 2 hours, C: 3 hours. Constraint: Worker C suffers a productivity loss of 4 kits/hr. Total kits produced?

Abstraction and Information-Structure Invariant Type: Redundant Restatement Invariance
Question: Find the rate vector (a, b, c) from the pair sums a+b=33, b+c=36, and a+c=25. Apply a decrement of 4 to the c-component to account for overhead. Then, compute the dot product of this modified rate vector with the hour vector (3, 2, 3) to find the total kits produced.

Figure 2. Two representative examples from INVARIANTBENCH.

tation, we employ a verifier (Cobbe et al., 2021; Lightman et al., 2023) (§E.2) that normalizes numerical and symbolic expressions prior to comparison; residual notation-sensitive mismatches are tracked separately (§ I). Importantly, invariance is *not* a proxy for correctness: $\mathrm{Cons}(M) = 1$ is necessary but not sufficient, as a model may be consistently wrong across all variants. We therefore interpret correctness under consistency (e.g., IMI/CIR) as the strongest verifiable signal. Conversely, any $\mathrm{Cons}(M) < 1$ provides a constructive counterexample to representation-invariant reasoning under Eq. 2. Finally, invariance is *axis-dependent*: different transformations preserve semantics at different granularities (from strict logical equivalence to shifts in pragmatics or information structure). Accordingly, we report per-axis behaviour to distinguish superficial robustness from deeper representation-invariant competence.

3. INVARIANTBENCH Benchmark

We construct INVARIANTBENCH as a controlled evaluation suite for representational invariance in large language models. The benchmark is designed to systematically vary surface form while preserving semantic content, enabling precise measurement of whether models maintain consistent reasoning under equivalent reformulations. The dataset comprises 1,200 seed problems spanning 16 core tasks (Table 6), organized into 3 principal reasoning categories: symbolic and linguistic reasoning, structural and representational reasoning, and abstraction and information-structure reasoning. Each seed problem is instantiated into four semantically equivalent surface forms, generated via a structured transformation pipeline that ensures invariance preservation under a formal semantic contract.

Invariance design. Each problem is transformed under a strict semantic invariance contract, ensuring that all variants preserve the underlying solution despite surface-level or representational perturbations (details in § D.2), as shown in Figure 2. All transformations are expert-verified.

Expert annotation protocol. Annotation is performed by domain-expert annotators trained in formal reasoning, logic, and mathematical problem design. Each seed problem is first independently validated for unambiguous solvability and then manually expanded into invariant variants under strict equivalence constraints. Annotators ensure that transformations do not introduce new solution paths or alter difficulty through extraneous reasoning shortcuts. A secondary expert review stage verifies semantic equivalence across all variants, while a final adjudication pass resolves edge cases involving ambiguity in logical or mathematical reformulations.

Quality control and verification. To ensure correctness, each problem–variant set is validated using a two-stage verification pipeline. First, automated consistency checks verify structural equivalence and detect invalid transformations (e.g., accidental constraint changes). Second, expert reviewers manually inspect sampled instances for semantic drift and unintended ambiguity. Only items passing both stages are included in the final benchmark. This results in a high-confidence dataset of invariantly paired reasoning tasks suitable for measuring representation-sensitive degradation in modern LLMs.

4. Experiments and Evaluation

Models. We categorize the evaluated models into three groups. **Small open-source models** (below 30B) comprise Llama-3.2-1B-Instruct (Grattafiori et al., 2024), Ministral-3B-Instruct-2512 (Liu et al., 2026), Gemma-3-4B-Instruct (Team et al., 2025), Qwen3-8B (Yang et al., 2025), and Gemma-4-26B-A4B-Instruct (Google, 2025). **Large open-source models** (more than 30B) comprise Olmo-3.1-32B-Instruct (Olmo et al., 2026), Aya-Expanse-32B (Dang et al., 2024), Llama-3.3-70B-Instruct (Grattafiori et al., 2024), Qwen3.6-Plus (Yang et al., 2025), and DeepSeek-V3.1 (DeepSeek-AI et al., 2025). Lastly, **proprietary models** (closed source) include o4-mini (OpenAI, 2025b), GPT-

Table 1. **Accuracy for all evaluated models.** Base = original question; $\mathcal{I}_1/\mathcal{I}_2/\mathcal{I}_3$ = Symbolic & Linguistic / Structural & Representational / Abstraction & Information-Structure invariants. $\overline{\mathcal{I}}$ = mean invariant accuracy; Δ = Base $-\overline{\mathcal{I}}$ (positive = degradation under invariance).

	Accuracy Across Different Invariants (%) (↑)				*Mean (↑)*	*Base*$-\overline{\mathcal{I}}$
Model	Base	$\mathcal{I}_1$	$\mathcal{I}_2$	$\mathcal{I}_3$	$\overline{\mathcal{I}}$	Δ
Open-Source Models (Small) ≤30B						
Llama-3.2-1B-I	6.92	6.00	7.67	7.58	7.08	-0.16
Ministral-3B-2512	10.83	9.75	11.75	10.58	10.69	+0.14
Gemma-3-4B-I	9.92	10.25	11.00	9.58	10.28	-0.36
Qwen3-8B	55.08	57.83	57.75	56.58	57.39	**-2.31**
Gemma-4-26B-A4B-I	29.25	26.58	26.83	26.00	26.47	+2.78
Open-Source Models (Large) ¿30B						
Olmo-3.1-32B-I	17.50	13.75	17.08	14.00	14.94	+2.56
Aya-Expanse-32B	15.67	12.00	16.83	14.33	14.39	+1.28
Llama-3.3-70B-I	25.08	24.25	27.58	23.50	25.11	-0.03
Qwen3.6-Plus	69.75	65.67	65.33	66.17	65.72	+4.03
DeepSeek-V3.1	19.33	16.17	20.17	15.92	17.42	+1.91
Closed-Source Models						
Command-A	20.83	20.58	21.17	21.17	20.97	-0.14
o4-mini	21.50	15.17	18.67	14.92	16.25	+5.25
GPT-5.2	69.75	67.25	68.67	69.25	68.39	+1.36
Gemini-3-Flash	41.92	37.92	37.75	38.92	38.20	+3.72
Claude Haiku 4.5	48.67	49.58	48.83	47.67	48.69	-0.02
Human Baseline						
Human (Base)	90.00	90.42	90.17	89.50	90.03	-0.03
Human (Expert)	**99.92**	**100.00**	**99.25**	**99.33**	**99.53**	+0.39

5.2 (OpenAI, 2025a), Gemini-3-Flash (Team, 2025), and Claude 4.5 Haiku (Anthropic, 2025).

Human Baselines. Human baselines are derived from two groups: *base* (high-school to early undergraduate level) and *expert* (STEM graduates, Master's, and PhD students) (detailed in § E.3), with three annotators per group and majority voting used to get final responses for each question.

Improving Invariant Reasoning Capabilities. We evaluate supervised fine-tuning (SFT) and GRPO-based RL on GEMMA-3-1B-IT. Under SFT, a 40:60 stratified split leads to strong memorization, with gains driven by leakage across invariant variants. In the constrained `base` + `inv1` setting, unseen variants remain low (17.42, 14.58) while seen variants reach 41.42 and 41.75, indicating poor invariance. GRPO improves unseen variants to 19.17 and 20.69 within 5 epochs, a gain of $\sim$ 3–6 points over SFT, while seen variants reach 29.72 and 31.53. This reflects modest but limited generalization.

5. Findings and Analysis

5.1. Overall Results

Overall Accuracy. As shown in Table 1, across all model families, accuracy consistently degrades under invariant transformations, confirming that representation shift remains a primary failure mode even for high-performing systems. Smaller open-source models such as LLAMA-3.2-1B-I and MINISTRAL-3B-2512 show low absolute performance and unstable invariance behaviour, with small fluctuations around $\overline{\mathcal{I}}$, indicating weak semantic grounding. QWEN3-8B notably outperforms several larger models (e.g., LLAMA-3.3-70B-INSTRUCT, DEEPSEEK-V3.1, OLMO-3.1-32B, and O4-MINI), achieving substantially higher base accuracy despite its smaller parameter scale. Larger open-source and closed-source models improve substantially in base accuracy, but still show persistent and often amplified degradation gaps ($\Delta > 0$), suggesting that scaling alone does not eliminate invariance sensitivity. Notably, strong models such as QWEN3.6-PLUS, O4-MINI, and GEMINI-3-FLASH exhibit large positive Δ, implying that improvements in base performance do not translate proportionally to invariant robustness. In contrast, human performance (HUMAN (BASE) and HUMAN (EXPERT)) remains near ceiling with negligible degradation and near-zero Δ, establishing a clear upper bound on invariance stability that current models have not yet matched.

Invariance Stability. As shown in Table 2, across models, invariance stability metrics reveal a clear separation between scaling gains and true robustness, with most systems showing low IMI and CIR alongside non-trivial ICR, indicating that full consistency across variants remains rare. Smaller models exhibit uniformly weak stability, with low IMI and CIR values and high CE, reflecting fragmented and largely unstructured reasoning under reformulation. Larger open-source and closed-source models improve substantially in CIR and IMI (e.g., QWEN3.6-PLUS and GPT-5.2), but still retain non-zero ICR and elevated CE, showing that correct

Table 2. **Invariance stability metrics (Overall)**. **IMI** = Invariant Mass Index; **CIR** = Conditional Invariance Reliability; **ICR** = Invariance Collapse Rate; **BSS** = Bilateral Symmetry Score (1 = symmetric failures, 0 = all failures in one direction); **CE** = Consistency Entropy (lower = more structured reasoning); **PIDI** = pairwise failure co-occurrence probability.

	Global Stability (%)			*Novel (%)*		*Pairwise Failure Dep. (%)*		
Model	IMI↑	CIR↑	ICR↓	BSS↑	CE↓	$P(\mathcal{I}_1{=}0\|\mathcal{I}_2{=}0)$	$P(\mathcal{I}_1{=}0\|\mathcal{I}_3{=}0)$	$P(\mathcal{I}_2{=}0\|\mathcal{I}_3{=}0)$
Open-Source Models (Small) ≤30B								
Llama-3.2-1B-I	1.67	24.10	**3.08**	75.45	**11.87**	96.75	96.66	95.13
Ministral-3B-2512	1.17	10.77	6.42	81.98	20.45	93.11	93.38	91.24
Gemma-3-4B-I	1.92	19.33	5.17	71.91	18.93	93.07	92.90	91.71
Qwen3-8B	32.17	58.40	7.00	96.23	38.33	72.39	71.79	66.03
Gemma-4-26B-A4B-I	7.25	24.79	10.42	92.02	35.34	84.28	84.35	82.55
Open-Source Models (Large) ¿30B								
Olmo-3.1-32B-I	3.00	17.14	9.42	**97.48**	25.85	91.86	90.60	88.18
Aya-Expanse-32B	2.67	17.02	8.08	90.17	24.90	92.59	92.70	88.23
Llama-3.3-70B-I	5.50	21.93	10.67	88.68	38.24	83.08	84.31	80.50
Qwen3.6-Plus	**48.58**	**69.65**	5.67	65.25	26.93	74.28	71.92	70.69
DeepSeek-V3.1	2.17	11.21	11.08	94.93	31.53	89.77	87.91	85.73
Closed-Source Models								
Command-A	4.33	20.80	8.50	86.27	32.89	85.62	86.05	86.15
o4-mini	3.67	17.05	11.67	88.54	27.77	90.88	90.60	87.07
GPT-5.2	48.50	69.53	4.67	72.18	28.33	71.01	73.71	69.92
Gemini-3-Flash	17.00	40.56	10.92	85.00	37.51	80.19	80.08	80.22
Claude Haiku 4.5	17.42	35.79	12.92	95.98	52.24	66.12	66.88	64.01
Human Baseline								
Human (Base)	65.33	72.59	2.42	57.69	28.93	11.02	8.73	11.11
Human (Expert)	**98.58**	**98.67**	**0.00**	11.76	**1.17**	0.00	0.00	0.00

reasoning often fails to persist across all invariant axes. Pairwise dependency scores (PIDI) remain high for most models, indicating strongly correlated failure modes rather than independent errors, though they are notably lower in stronger systems, suggesting partial decoupling of invariant weaknesses. In contrast, human baselines achieve near-perfect CIR and IMI with negligible ICR and near-zero PIDI, establishing that robust invariance is achievable in principle but not yet matched by current models.

Symmetry, Entropy and Recovery. As shown in Table 3, across models, bilateral symmetry (BSS) is generally high, indicating that most systems exhibit broadly balanced forward and backward failure modes, although this symmetry often coexists with substantial error rates in both directions. Small open-source models show relatively structured but weak recovery behaviour, with low IRR $*$ full and consistently negative IRR $*$ Δ, suggesting that correct reasoning on invariants is rarely recovered when the base fails. Larger models improve in partial recovery (IRR $*$ part), but still exhibit strongly negative IRR $*$ Δ in most cases, indicating that invariants are not reliably easier than the base and often fail independently. Fragility Gradient (FG) scores remain mostly at 1 across systems, showing that most models do not respect the intended monotonic difficulty ordering, with only a few exceptions achieving FG = 2. In contrast, human baselines achieve markedly higher full recovery rates and more meaningful gradient structure, reinforcing that human reasoning exhibits both stronger recoverability and more coherent sensitivity to representational complexity.

Invariance sensitivity Profile. As shown in Table 4, across models, conditional invariance retention (TCIR) increases substantially with scale, with larger systems such as QWEN3.6-PLUS and GPT-5.2 achieving consistently high retention across all invariant types, indicating stronger stability once a problem is partially solvable. However, collapse contributions (TCC) remain non-uniform across invariant axes, with $\mathcal{I}_2$ (structural transformations) often contributing disproportionately to partial collapses, suggesting that representational changes in structure remain the most failure-inducing perturbation. Relative sensitivity (RIS) further reveals a systematic asymmetry: symbolic and linguistic invariants are generally easier for most models (positive RIS), while abstraction-level transformations tend to be harder (negative RIS), indicating a consistent gradient of difficulty across invariant families. Despite strong TCIR in large models, persistent negative RIS values in $\mathcal{I}_3$ show that abstraction shifts remain a structural bottleneck even when base competence is high. In contrast, human baselines exhibit near-ceiling TCIR with low and balanced collapse contributions, demonstrating stable invariance across all transformation types with minimal sensitivity drift.

Correctness Profile. As shown in §G.1, model performance drops sharply from $\geq$1 to $\geq$4 correct variants, unlike hu-

Table 3. **Performance on invariance analysis (symmetry, entropy, recovery) metrics. BSS** = Bilateral Symmetry Score; F^+ = forward failure rate (base ✓ $\rightarrow$ inv ✗); F^- = backward failure rate (base ✗ $\rightarrow$ inv ✓); **CE** = Consistency Entropy (0 = fully structured, 1 = random); IRR_{full} = full invariant recovery given base failure; IRR_{part} = partial recovery ($\geq$1 invariant correct); $\text{IRR}_{\Delta} = \text{IRR}_{\text{full}} - (1 - \text{CIR})$ (positive = base harder than invariants); **FG** = Fragility Gradient monotonicity score (0–2, 2 = expected ordering); $\bar{g}$ = mean gradient magnitude in pp.

	Bilateral Symmetry			*Entropy*	*Invariant Recovery (%)*			*Fragility Gradient*	
Model	BSS↑	$F^+(\%)$	$F^-(\%)$	CE↓	IRR_{full} ↑	IRR_{part} ↑	IRR_{Δ}	FG	$\bar{g}$
Open-Source Models (Small) $\leq$30B									
Llama-3.2-1B-I	75.45	5.25	8.67	**11.87**	0.09	9.31	-75.81	1	-0.79
Ministral-3B-2512	81.98	9.67	13.92	20.45	0.65	15.61	-88.58	1	-0.42
Gemma-3-4B-I	71.91	8.00	14.25	18.93	0.28	15.82	-80.39	1	+0.34
Qwen3-8B	96.23	22.92	21.25	38.33	10.39	47.31	-31.21	**2**	+0.63
Gemma-4-26B-A4B-I	92.02	22.00	18.75	35.34	2.59	26.50	-72.62	1	+0.29
Open-Source Models (Large) ¿30B									
Olmo-3.1-32B-I	**97.48**	14.50	15.25	25.85	0.81	18.48	-82.05	1	-0.13
Aya-Expanse-32B	90.17	13.00	15.83	24.90	0.89	18.77	-82.09	1	-1.17
Llama-3.3-70B-I	88.68	19.58	24.58	38.24	2.45	32.81	-75.62	1	+0.37
Qwen3.6-Plus	65.25	21.17	10.25	26.93	4.13	33.88	-26.22	1	-0.25
DeepSeek-V3.1	94.93	17.17	19.00	31.53	1.86	23.55	-86.93	1	+0.13
Closed-Source Models									
Command-A	86.27	16.50	21.75	32.89	1.47	27.47	-77.73	1	-0.29
o4-mini	88.54	17.83	14.17	27.77	1.06	18.05	-81.89	1	+0.12
GPT-5.2	72.18	21.25	12.00	28.33	**12.12**	39.67	-18.35	0	-1.00
Gemini-3-Flash	85.00	24.92	18.42	37.51	4.02	31.71	-55.42	1	-0.50
Claude Haiku 4.5	95.98	31.25	28.83	52.24	9.58	**56.17**	-54.63	2	+0.95
Human Baseline									
Human (Base)	57.69	24.67	10.00	28.93	**76.67**	**100.00**	+49.26	2	+0.46
Human (Expert)	11.76	1.33	0.08	**1.17**	0.00	100.00	-1.33	1	+0.33

mans, which remain nearly perfect invariants, while even strong models like QWEN3.6-PLUS and GPT-5.2 show improved but still clearly sub-human high-consistency retention.

§ H provides the complete category- and subcategory-wise breakdown of the experimental results.

5.2. Error and Reasoning Failure Mode Analysis

We conducted a systematic two-pass qualitative coding of all failures produced during evaluation across all scored question–response pairs, yielding an eight-category failure mode taxonomy aligned to the reasoning pipeline (§ I): premise comprehension (E1), problem formalisation (E2), local computation (E3), chain coherence (E4), invariance robustness (E5), lexical-semantic selection (E6), output specification (E7), and generation pathology (E8). Aggregate failure rates span from 31.3% (GPT-5) to 84.4% (OLMO-3.1-32B), with a clear tier structure: top models (GPT-5, QWEN3.6-PLUS) are primarily limited by E5 and E4 failures, while lower-tier models are dominated by E1 and E3 errors, indicating that constraint integration and basic computation break down before higher-order reasoning effects are observable. The most prevalent single failure mode is *dropped or ignored constraints* (E1.1), which accounts for a disproportionate share of errors in algebraic and multi-constraint.

Across all systems, *Abstraction Level Shift* is the most uniformly damaging invariant transformation, with breakage rates in the 11–21% range even for top-performing models—consistently exceeding all other invariance axes and confirming that failures reflect genuine representational fragility rather than memorization artifacts. Mitigation analysis identifies five systemic root causes: fragile input representation binding, weak cross-form consistency, limited multi-step state persistence, over-reliance on surface priors, and decoding instability under structured constraints. Effective interventions therefore require joint enforcement across invariance families rather than axis-specific augmentation, explicit constraint-aware attention regularization, and structured intermediate state tracking. Full error taxonomy, per-model breakage rates by sub-type and invariant axis, concrete failure examples, and detailed mitigation strategies are provided in § I.

5.3. Performance Improvement Approaches

We investigate how different optimization strategies influence invariance learning and cross-form generalization under identical evaluation conditions.

Supervised Fine-Tuning (SFT). We investigate how different optimization strategies influence invariance learning and cross-form generalization under identical evaluation

Table 4. **Invariance sensitivity profile (Overall)**. **TCIR** = conditional invariance retention; **TCC** = collapse contribution rate; **RIS** = relative sensitivity (pp deviation from model mean; positive = easier, negative = harder). Subscripts 1/2/3 denote $\mathcal{I}_1/\mathcal{I}_2/\mathcal{I}_3$.

	Cond. Inv. Retention (%)			*Collapse Contribution (%)*			*Rel. Sensitivity (pp)*		
Model	$TCIR_1$	$TCIR_2$	$TCIR_3$	$TCC_1 \downarrow$	$TCC_2 \downarrow$	$TCC_3 \downarrow$	RIS_1	RIS_2	RIS_3
Open-Source Models (Small) ≤30B									
Llama-3.2-1B-I	38.50	49.20	48.66	78.57	60.00	61.43	-1.04	+0.63	+0.54
Ministral-3B-2512	39.39	47.47	42.76	69.20	59.92	67.93	-0.98	+1.02	-0.15
Gemma-3-4B-I	42.41	45.52	39.66	69.13	61.74	71.74	+0.06	+0.81	-0.61
Qwen3-8B	75.76	75.66	74.13	**60.07**	59.36	54.06	+1.02	+0.94	-0.23
Gemma-4-26B-A4B-I	55.38	55.90	54.17	66.77	62.80	64.33	-0.59	-0.34	-1.17
Open-Source Models (Large) ¿30B									
Olmo-3.1-32B-I	41.98	52.16	42.75	71.18	60.07	70.14	-1.83	+1.50	-1.58
Aya-Expanse-32B	38.10	53.44	45.50	75.54	57.19	67.63	-2.71	+2.12	-0.38
Llama-3.3-70B-I	48.83	55.54	47.32	67.83	57.36	68.58	-0.85	+2.48	-1.60
Qwen3.6-Plus	82.08	81.67	82.71	69.89	62.50	**53.98**	-1.06	-1.40	-0.56
DeepSeek-V3.1	42.17	52.61	41.52	68.22	60.06	69.97	-1.73	+2.27	-1.98
Closed-Source Models									
Command-A	48.34	49.71	49.71	65.33	64.76	65.33	-0.36	+0.23	+0.23
o4-mini	42.52	52.34	41.82	73.00	62.33	73.00	-2.39	+1.11	-2.64
GPT-5.2	**82.26**	**84.00**	**84.71**	69.87	**55.13**	55.77	-1.48	-0.06	+0.52
Gemini-3-Flash	62.85	62.57	64.50	65.12	66.05	62.96	-1.21	-1.37	-0.20
Claude Haiku 4.5	63.98	63.01	61.51	61.54	58.82	64.48	+0.89	+0.14	-1.02
Human Baseline									
Human (Base)	90.42	90.17	89.50	47.37	59.65	54.39	+0.40	+0.15	-0.52
Human (Expert)	**100.00**	**99.25**	**99.33**	**0.00**	100.00	**0.00**	+0.38	-0.37	-0.30

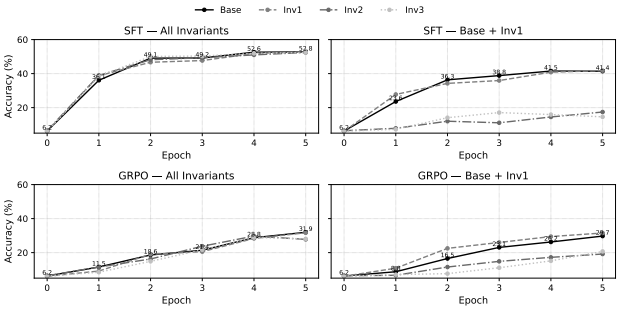

Figure 3. **Learning dynamics (60% test-set accuracy) under SFT and GRPO across training regimes.** SFT shows rapid gains but strong regime sensitivity and poor generalization under partial invariants, while GRPO exhibits tighter base–invariant coupling, indicating improved robustness.

conditions. As shown in Figure 3, SFT exhibits a sharp and uniform rise under the *All Invariants* regime, with all four variants converging tightly to $\sim$ 52–53% (Base: 52.75, Inv1: 52.75, Inv2: 52.42, Inv3: 52.42), indicating strong cross-variant leakage and memorization of overlapping surface forms rather than invariant reasoning. In contrast, under *Base + Inv1* training, performance improves primarily on seen variants (Base: 41.42, Inv1: 41.75), while unseen Inv2/Inv3 remain low (17.42 and 14.58), creating a persistent $\sim$ 24–27 point gap. This divergence demonstrates that SFT learns form-specific mappings that fail to transfer across semantically equivalent reformulations. Full training details and epoch-wise results are provided in § J.1.

Reinforcement Learning (RL) with GRPO (Group Relative Policy Optimization). In Figure 3, GRPO shows a slower but more differentiated learning trajectory, with early divergence across variants that gradually narrows under *All Invariants*, reaching $\sim$ 32% on Base/Inv1 and $\sim$ 28% on Inv2/Inv3 by epoch 5 (Base: 31.94, Inv1: 31.67, Inv2: 27.78, Inv3: 28.06). Under *Base + Inv1*, GRPO improves generalization to unseen variants more than SFT but still remains limited, reaching 19.17 (Inv2) and 20.69 (Inv3), while seen variants achieve 29.72 (Base) and 31.53 (Inv1). This indicates that reward-driven optimization partially reduces dependence on surface form, enabling modest invariance transfer, but does not achieve strong generalization. Additional details and results are reported in § J.2.

Overall, both SFT and GRPO improve performance but are insufficient to close the invariance gap. This highlights the need for stronger mechanisms to induce true invariant reasoning.

6. Concluding Remarks

In this work, we introduced INVARIANTBENCH, a principled framework for evaluating reasoning through invariance under semantics-preserving transformations. Our results demonstrate that high accuracy on fixed benchmarks substantially overestimates true reasoning ability, as models exhibit consistent and measurable degradation across equivalent representations. The observed invariance gap, together with a detailed failure taxonomy, reveals that current LLMs rely heavily on surface-form cues rather than stable, abstraction-level reasoning. While methods such as GRPO show partial improvements, robust cross-form generalization remains an open challenge. We hope this work establishes invariance as a core evaluation axis and motivates future research toward representation-robust reasoning systems.

References

Anthropic. The claude 3 model family: Opus, sonnet, haiku. Technical report, Anthropic, 2024. URL https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pd.

Anthropic. Introducing claude haiku 4.5, 2025. URL https://www.anthropic.com/news/claude-haiku-4-5.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., and Zhang, Y. Sparks of artificial general intelligence: Early experiments with gpt-4, 2023. URL https://arxiv.org/abs/2303.12712.

Chen, M., Tworek, J., Jun, H., Yuan, Q., de Oliveira Pinto, H. P., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., Ryder, N., Pavlov, M., Power, A., Kaiser, L., Bavarian, M., Winter, C., Tillet, P., Such, F. P., Cummings, D., Plappert, M., Chantzis, F., Barnes, E., Herbert-Voss, A., Guss, W. H., Nichol, A., Paino, A., Tezak, N., Tang, J., Babuschkin, I., Balaji, S., Jain, S., Saunders, W., Hesse, C., Carr, A. N., Leike, J., Achiam, J., Misra, V., Morikawa, E., Radford, A., Knight, M., Brundage, M., Murati, M., Mayer, K., Welinder, P., McGrew, B., Amodei, D., McCandlish, S., Sutskever, I., and Zaremba, W. Evaluating large language models trained on code, 2021. URL https://arxiv.org/abs/2107.03374.

Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., and Tafjord, O. Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018. URL https://arxiv.org/abs/1803.05457.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., and Schulman, J. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

Dang, J., Singh, S., D'souza, D., Ahmadian, A., Salamanca, A., Smith, M., Peppin, A., Hong, S., Govindassamy, M., Zhao, T., Kublik, S., Amer, M., Aryabumi, V., Campos, J. A., Tan, Y.-C., Kocmi, T., Strub, F., Grinsztajn, N., Flet-Berliac, Y., Locatelli, A., Lin, H., Talupuru, D., Venkitesh, B., Cairuz, D., Yang, B., Chung, T., Ko, W.-Y., Shi, S. S., Shukayev, A., Bae, S., Piktus, A., Castagné, R., Cruz-Salinas, F., Kim, E., Crawhall-Stein, L., Morisot, A., Roy, S., Blunsom, P., Zhang, I., Gomez, A., Frosst, N., Fadaee, M., Ermis, B., Üstün, A., and Hooker, S. Aya expanse: Combining research breakthroughs for a new multilingual frontier, 2024. URL https://arxiv.org/abs/2412.04261.

DeepSeek-AI, Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., Dai, D., Guo, D., Yang, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., Luo, F., Hao, G., Chen, G., Li, G., Zhang, H., Bao, H., Xu, H., Wang, H., Zhang, H., Ding, H., Xin, H., Gao, H., Li, H., Qu, H., Cai, J. L., Liang, J., Guo, J., Ni, J., Li, J., Wang, J., Chen, J., Chen, J., Yuan, J., Qiu, J., Li, J., Song, J., Dong, K., Hu, K., Gao, K., Guan, K., Huang, K., Yu, K., Wang, L., Zhang, L., Xu, L., Xia, L., Zhao, L., Wang, L., Zhang, L., Li, M., Wang, M., Zhang, M., Zhang, M., Tang, M., Li, M., Tian, N., Huang, P., Wang, P., Zhang, P., Wang, Q., Zhu, Q., Chen, Q., Du, Q., Chen, R. J., Jin, R. L., Ge, R., Zhang, R., Pan, R., Wang, R., Xu, R., Zhang, R., Chen, R., Li, S. S., Lu, S., Zhou, S., Chen, S., Wu, S., Ye, S., Ye, S., Ma, S., Wang, S., Zhou, S., Yu, S., Zhou, S., Pan, S., Wang, T., Yun, T., Pei, T., Sun, T., Xiao, W. L., Zeng, W., Zhao, W., An, W., Liu, W., Liang, W., Gao, W., Yu, W., Zhang, W., Li, X. Q., Jin, X., Wang, X., Bi, X., Liu, X., Wang, X., Shen, X., Chen, X., Zhang, X., Chen, X., Nie, X., Sun, X., Wang, X., Cheng, X., Liu, X., Xie, X., Liu, X., Yu, X., Song, X., Shan, X., Zhou, X., Yang, X., Li, X., Su, X., Lin, X., Li, Y. K., Wang, Y. Q., Wei, Y. X., Zhu, Y. X., Zhang, Y., Xu, Y., Xu, Y., Huang, Y., Li, Y., Zhao, Y., Sun, Y., Li, Y., Wang, Y., Yu, Y., Zheng, Y., Zhang, Y., Shi, Y., Xiong, Y., He, Y., Tang, Y., Piao, Y., Wang, Y., Tan, Y., Ma, Y., Liu, Y., Guo, Y., Wu, Y., Ou, Y., Zhu, Y., Wang, Y., Gong, Y., Zou, Y., He, Y., Zha, Y., Xiong, Y., Ma, Y., Yan, Y., Luo, Y., You, Y., Liu, Y., Zhou, Y., Wu, Z. F., Ren, Z. Z., Ren, Z., Sha, Z., Fu, Z., Xu, Z., Huang, Z., Zhang, Z., Xie, Z., Zhang, Z., Hao, Z., Gou, Z., Ma, Z., Yan, Z., Shao, Z., Xu, Z., Wu, Z., Zhang, Z., Li, Z., Gu,

Z., Zhu, Z., Liu, Z., Li, Z., Xie, Z., Song, Z., Gao, Z., and Pan, Z. Deepseek-v3 technical report, 2025. URL https://arxiv.org/abs/2412.19437.

et al., S. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Trans. Mach. Learn. Res.*, 2023, 2023. URL https://api.semanticscholar.org/CorpusID:271601672.

Geva, M., Khashabi, D., Segal, E., Khot, T., Roth, D., and Berant, J. ¡i¿did aristotle use a laptop?¡/i¿a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361, 2021. ISSN 2307-387X. doi: 10.1162/tacl_a_00370. URL http://dx.doi.org/10.1162/tacl_a_00370.

Golchin, S. and Surdeanu, M. Time travel in LLMs: Tracing data contamination in large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=2Rwq6c3tvr.

Google. Gemma 4 model card, 2025. URL http://ai.google.dev/gemma/docs/core/model_card_4.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., Rodriguez, A., Gregerson, A., Spataru, A., Roziere, B., Biron, B., Tang, B., Chern, B., Caucheteux, C., Nayak, C., Bi, C., Marra, C., McConnell, C., Keller, C., Touret, C., Wu, C., Wong, C., Ferrer, C. C., Nikolaidis, C., Allonsius, D., Song, D., Pintz, D., Livshits, D., Wyatt, D., Esiobu, D., Choudhary, D., Mahajan, D., Garcia-Olano, D., Perino, D., Hupkes, D., Lakomkin, E., AlBadawy, E., Lobanova, E., Dinan, E., Smith, E. M., Radenovic, F., Guzmán, F., Zhang, F., Synnaeve, G., Lee, G., Anderson, G. L., Thattai, G., Nail, G., Mialon, G., Pang, G., Cucurell, G., Nguyen, H., Korevaar, H., Xu, H., Touvron, H., Zarov, I., Ibarra, I. A., Kloumann, I., Misra, I., Evtimov, I., Zhang, J., Copet, J., Lee, J., Geffert, J., Vranes, J., Park, J., Mahadeokar, J., Shah, J., van der Linde, J., Billock, J., Hong, J., Lee, J., Fu, J., Chi, J., Huang, J., Liu, J., Wang, J., Yu, J., Bitton, J., Spisak, J., Park, J., Rocca, J., Johnstun, J., Saxe, J., Jia, J., Alwala, K. V., Prasad, K., Upasani, K., Plawiak, K., Li, K., Heafield, K., Stone, K., El-Arini, K., Iyer, K., Malik, K., Chiu, K., Bhalla, K., Lakhotia, K., Rantala-Yeary, L., van der Maaten, L., Chen, L., Tan, L., Jenkins, L., Martin, L., Madaan, L., Malo, L., Blecher, L., Landzaat, L., de Oliveira, L., Muzzi, M., Pasupuleti, M., Singh, M., Paluri, M., Kardas, M., Tsimpoukelli, M., Oldham, M., Rita, M., Pavlova, M., Kambadur, M., Lewis, M., Si, M., Singh, M. K., Hassan, M., Goyal, N., Torabi, N., Bashlykov, N., Bogoychev, N., Chatterji, N., Zhang, N., Duchenne, O., Çelebi, O., Alrassy, P., Zhang, P., Li, P., Vasic, P., Weng, P., Bhargava, P., Dubal, P., Krishnan, P., Koura, P. S., Xu, P., He, Q., Dong, Q., Srinivasan, R., Ganapathy, R., Calderer, R., Cabral, R. S., Stojnic, R., Raileanu, R., Maheswari, R., Girdhar, R., Patel, R., Sauvestre, R., Polidoro, R., Sumbaly, R., Taylor, R., Silva, R., Hou, R., Wang, R., Hosseini, S., Chennabasappa, S., Singh, S., Bell, S., Kim, S. S., Edunov, S., Nie, S., Narang, S., Raparthy, S., Shen, S., Wan, S., Bhosale, S., Zhang, S., Vandenhende, S., Batra, S., Whitman, S., Sootla, S., Collot, S., Gururangan, S., Borodinsky, S., Herman, T., Fowler, T., Sheasha, T., Georgiou, T., Scialom, T., Speckbacher, T., Mihaylov, T., Xiao, T., Karn, U., Goswami, V., Gupta, V., Ramanathan, V., Kerkez, V., Gonguet, V., Do, V., Vogeti, V., Albiero, V., Petrovic, V., Chu, W., Xiong, W., Fu, W., Meers, W., Martinet, X., Wang, X., Wang, X., Tan, X. E., Xia, X., Xie, X., Jia, X., Wang, X., Goldschlag, Y., Gaur, Y., Babaei, Y., Wen, Y., Song, Y., Zhang, Y., Li, Y., Mao, Y., Coudert, Z. D., Yan, Z., Chen, Z., Papakipos, Z., Singh, A., Srivastava, A., Jain, A., Kelsey, A., Shajnfeld, A., Gangidi, A., Victoria, A., Goldstand, A., Menon, A., Sharma, A., Boesenberg, A., Baevski, A., Feinstein, A., Kallet, A., Sangani, A., Teo, A., Yunus, A., Lupu, A., Alvarado, A., Caples, A., Gu, A., Ho, A., Poulton, A., Ryan, A., Ramchandani, A., Dong, A., Franco, A., Goyal, A., Saraf, A., Chowdhury, A., Gabriel, A., Bharambe, A., Eisenman, A., Yazdan, A., James, B., Maurer, B., Leonhardi, B., Huang, B., Loyd, B., Paola, B. D., Paranjape, B., Liu, B., Wu, B., Ni, B., Hancock, B., Wasti, B., Spence, B., Stojkovic, B., Gamido, B., Montalvo, B., Parker, C., Burton, C., Mejia, C., Liu, C., Wang, C., Kim, C., Zhou, C., Hu, C., Chu, C.-H., Cai, C., Tindal, C., Feichtenhofer, C., Gao, C., Civin, D., Beaty, D., Kreymer, D., Li, D., Adkins, D., Xu, D., Testuggine, D., David, D., Parikh, D., Liskovich, D., Foss, D., Wang, D., Le, D., Holland, D., Dowling, E., Jamil, E., Montgomery, E., Presani, E., Hahn, E., Wood, E., Le, E.-T., Brinkman, E., Arcaute, E., Dunbar, E., Smothers, E., Sun, F., Kreuk, F., Tian, F., Kokkinos, F., Ozgenel, F., Caggioni, F., Kanayet, F., Seide, F., Florez, G. M., Schwarz, G., Badeer, G., Swee, G., Halpern, G., Herman, G., Sizov, G., Guangyi, Zhang, Lakshminarayanan, G., Inan, H., Shojanazeri, H., Zou, H., Wang, H., Zha, H., Habeeb, H., Rudolph, H., Suk, H., Aspegren, H., Goldman, H., Zhan, H., Damlaj, I., Molybog, I., Tufanov, I., Leontiadis, I., Veliche, I.-E., Gat, I., Weissman, J., Geboski, J., Kohli, J., Lam, J., Asher, J., Gaya, J.-B., Marcus, J., Tang, J., Chan, J., Zhen, J., Reizenstein, J., Teboul, J., Zhong, J., Jin, J., Yang, J., Cummings, J., Carvill, J., Shepard, J., McPhie, J., Torres, J., Ginsburg, J., Wang, J., Wu, K., U, K. H., Saxena, K., Khandelwal, K., Zand, K., Matosich,

K., Veeraraghavan, K., Michelena, K., Li, K., Jagadeesh, K., Huang, K., Chawla, K., Huang, K., Chen, L., Garg, L., A, L., Silva, L., Bell, L., Zhang, L., Guo, L., Yu, L., Moshkovich, L., Wehrstedt, L., Khabsa, M., Avalani, M., Bhatt, M., Mankus, M., Hasson, M., Lennie, M., Reso, M., Groshev, M., Naumov, M., Lathi, M., Keneally, M., Liu, M., Seltzer, M. L., Valko, M., Restrepo, M., Patel, M., Vyatskov, M., Samvelyan, M., Clark, M., Macey, M., Wang, M., Hermoso, M. J., Metanat, M., Rastegari, M., Bansal, M., Santhanam, N., Parks, N., White, N., Bawa, N., Singhal, N., Egebo, N., Usunier, N., Mehta, N., Laptev, N. P., Dong, N., Cheng, N., Chernoguz, O., Hart, O., Salpekar, O., Kalinli, O., Kent, P., Parekh, P., Saab, P., Balaji, P., Rittner, P., Bontrager, P., Roux, P., Dollar, P., Zvyagina, P., Ratanchandani, P., Yuvraj, P., Liang, Q., Alao, R., Rodriguez, R., Ayub, R., Murthy, R., Nayani, R., Mitra, R., Parthasarathy, R., Li, R., Hogan, R., Battey, R., Wang, R., Howes, R., Rinott, R., Mehta, S., Siby, S., Bondu, S. J., Datta, S., Chugh, S., Hunt, S., Dhillon, S., Sidorov, S., Pan, S., Mahajan, S., Verma, S., Yamamoto, S., Ramaswamy, S., Lindsay, S., Lindsay, S., Feng, S., Lin, S., Zha, S. C., Patil, S., Shankar, S., Zhang, S., Zhang, S., Wang, S., Agarwal, S., Sajuyigbe, S., Chintala, S., Max, S., Chen, S., Kehoe, S., Satterfield, S., Govindaprasad, S., Gupta, S., Deng, S., Cho, S., Virk, S., Subramanian, S., Choudhury, S., Goldman, S., Remez, T., Glaser, T., Best, T., Koehler, T., Robinson, T., Li, T., Zhang, T., Matthews, T., Chou, T., Shaked, T., Vontimitta, V., Ajayi, V., Montanez, V., Mohan, V., Kumar, V. S., Mangla, V., Ionescu, V., Poenaru, V., Mihailescu, V. T., Ivanov, V., Li, W., Wang, W., Jiang, W., Bouaziz, W., Constable, W., Tang, X., Wu, X., Wang, X., Wu, X., Gao, X., Kleinman, Y., Chen, Y., Hu, Y., Jia, Y., Qi, Y., Li, Y., Zhang, Y., Zhang, Y., Adi, Y., Nam, Y., Yu, Wang, Zhao, Y., Hao, Y., Qian, Y., Li, Y., He, Y., Rait, Z., DeVito, Z., Rosnbrick, Z., Wen, Z., Yang, Z., Zhao, Z., and Ma, Z. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Gulati, A., Miranda, B., Chen, E., Xia, E., Fronsdal, K., de Moraes Dumont, B., and Koyejo, S. PutnamAXIOM: A functional and static benchmark for measuring higher level mathematical reasoning. In *The 4th Workshop on Mathematical Reasoning and AI at NeurIPS'24*, 2024. URL https://openreview.net/forum?id=YXnwlZe0yf.

Han, S., Schoelkopf, H., Zhao, Y., Qi, Z., Riddell, M., Zhou, W., Coady, J., Peng, D., Qiao, Y., Benson, L., Sun, L., Wardle-Solano, A., Szabó, H., Zubova, E., Burtell, M., Fan, J., Liu, Y., Wong, B., Sailor, M., Ni, A., Nan, L., Kasai, J., Yu, T., Zhang, R., Fabbri, A., Kryscinski, W. M., Yavuz, S., Liu, Y., Lin, X. V., Joty, S., Zhou, Y., Xiong, C., Ying, R., Cohan, A., and Radev, D. FOLIO: Natural language reasoning with first-order logic. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 22017–22031, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1229. URL https://aclanthology.org/2024.emnlp-main.1229/.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., and Steinhardt, J. Measuring mathematical problem solving with the math dataset. In Vanschoren, J. and Yeung, S. (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021b. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/be83ab3ecd0db773eb2dc1b0a17836a1-Paper-round2.pdf.

Karia, R., Bramblett, D., Dobhal, D., Verma, P., and Srivastava, S. ∀uto∃val: Autonomous assessment of llms in formal synthesis and interpretation tasks, 2024. URL https://arxiv.org/abs/2403.18327.

Keysers, D., Schärli, N., Scales, N., Buisman, H., Furrer, D., Kashubin, S., Momchev, N., Sinopalnikov, D., Stafiniak, L., Tihon, T., Tsarkov, D., Wang, X., van Zee, M., and Bousquet, O. Measuring compositional generalization: A comprehensive method on realistic data. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SygcCnNKwr.

Kiela, D., Bartolo, M., Nie, Y., Kaushik, D., Geiger, A., Wu, Z., Vidgen, B., Prasad, G., Singh, A., Ringshia, P., Ma, Z., Thrush, T., Riedel, S., Waseem, Z., Stenetorp, P., Jia, R., Bansal, M., Potts, C., and Williams, A. Dynabench: Rethinking benchmarking in NLP. In Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., and Zhou, Y. (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4110–4124, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.324. URL https://aclanthology.org/2021.naacl-main.324/.

Kim, N. and Linzen, T. COGS: A compositional generalization challenge based on semantic interpretation. In

Webber, B., Cohn, T., He, Y., and Liu, Y. (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9087–9105, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.731. URL https://aclanthology.org/2020.emnlp-main.731/.

Lake, B. M. and Baroni, M. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International Conference on Machine Learning*, 2017. URL https://api.semanticscholar.org/CorpusID:46761158.

Lanham, T., Chen, A., Radhakrishnan, A., Steiner, B., Denison, C., Hernandez, D., Li, D., Durmus, E., Hubinger, E., Kernion, J., Lukošiūtė, K., Nguyen, K., Cheng, N., Joseph, N., Schiefer, N., Rausch, O., Larson, R., McCandlish, S., Kundu, S., Kadavath, S., Yang, S., Henighan, T., Maxwell, T., Telleen-Lawton, T., Hume, T., Hatfield-Dodds, Z., Kaplan, J., Brauner, J., Bowman, S. R., and Perez, E. Measuring faithfulness in chain-of-thought reasoning, 2023. URL https://arxiv.org/abs/2307.13702.

Li, Q., Cui, L., Zhao, X., Kong, L., and Bi, W. GSM-plus: A comprehensive benchmark for evaluating the robustness of LLMs as mathematical problem solvers. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2961–2984, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.163. URL https://aclanthology.org/2024.acl-long.163/.

Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., and Cobbe, K. Let's verify step by step, 2023. URL https://arxiv.org/abs/2305.20050.

Liu, A. H., Khandelwal, K., Subramanian, S., Jouault, V., Rastogi, A., Sadé, A., Jeffares, A., Jiang, A., Cahill, A., Gavaudan, A., Sablayrolles, A., Héliou, A., You, A., Ehrenberg, A., Lo, A., Eliseev, A., Calvi, A., Sooriyarachchi, A., Bout, B., Rozière, B., Monicault, B. D., Lanfranchi, C., Barreau, C., Courtot, C., Grattarola, D., Dabert, D., de las Casas, D., Chane-Sane, E., Ahmed, F., Berrada, G., Ecrepont, G., Guinet, G., Novikov, G., Kunsch, G., Lample, G., Martin, G., Gupta, G., Ludziejewski, J., Rute, J., Studnia, J., Amar, J., Delas, J., Roberts, J. S., Yadav, K., Chandu, K., Jain, K., Aitchison, L., Fainsin, L., Blier, L., Zhao, L., Martin, L., Saulnier, L., Gao, L., Buyl, M., Jennings, M., Pellat, M., Prins, M., Poirée, M., Guillaumin, M., Dinot, M., Futeral, M., Darrin, M., Augustin, M., Chiquier, M., Schimpf, M., Grinsztajn, N., Gupta, N., Raghuraman, N., Bousquet, O., Duchenne, O., Wang, P., von Platen, P., Jacob, P., Wambergue, P., Kurylowicz, P., Muddireddy, P. R., Chagniot, P., Stock, P., Agrawal, P., Torroba, Q., Sauvestre, R., Soletskyi, R., Menneer, R., Vaze, S., Barry, S., Gandhi, S., Waghjale, S., Gandhi, S., Ghosh, S., Mishra, S., Aithal, S., Antoniak, S., Scao, T. L., Cachet, T., Sorg, T. S., Lavril, T., Saada, T. N., Chabal, T., Foubert, T., Robert, T., Wang, T., Lawson, T., Bewley, T., Bewley, T., Edwards, T., Jamil, U., Tomasini, U., Nemychnikova, V., Phung, V., Maladière, V., Richard, V., Bouaziz, W., Li, W.-D., Marshall, W., Li, X., Yang, X., Ouahidi, Y. E., Wang, Y., Tang, Y., and Ramzi, Z. Ministral 3, 2026. URL https://arxiv.org/abs/2601.08584.

Liu, J., Cui, L., Liu, H., Huang, D., Wang, Y., and Zhang, Y. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning, 2020. URL https://arxiv.org/abs/2007.08124.

Marcus, G. The next decade in ai: Four steps towards robust artificial intelligence, 2020. URL https://arxiv.org/abs/2002.06177.

Mirzadeh, I., Alizadeh, K., Shahrokhi, H., Tuzel, O., Bengio, S., and Farajtabar, M. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models, 2025. URL https://arxiv.org/abs/2410.05229.

Naik, A., Ravichander, A., Sadeh, N., Rose, C., and Neubig, G. Stress test evaluation for natural language inference. In Bender, E. M., Derczynski, L., and Isabelle, P. (eds.), *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 2340–2353, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics. URL https://aclanthology.org/C18-1198/.

Olmo, T., :, Ettinger, A., Bertsch, A., Kuehl, B., Graham, D., Heineman, D., Groeneveld, D., Brahman, F., Timbers, F., Ivison, H., Morrison, J., Poznanski, J., Lo, K., Soldaini, L., Jordan, M., Chen, M., Noukhovitch, M., Lambert, N., Walsh, P., Dasigi, P., Berry, R., Malik, S., Shah, S., Geng, S., Arora, S., Gupta, S., Anderson, T., Xiao, T., Murray, T., Romero, T., Graf, V., Asai, A., Bhagia, A., Wettig, A., Liu, A., Rangapur, A., Anastasiades, C., Huang, C., Schwenk, D., Trivedi, H., Magnusson, I., Lochner, J., Liu, J., Miranda, L. J. V., Sap, M., Morgan, M., Schmitz, M., Guerquin, M., Wilson, M., Huff, R., Bras, R. L., Xin, R., Shao, R., Skjonsberg, S., Shen, S. Z., Li, S. S., Wilde, T., Pyatkin, V., Merrill, W., Chang, Y., Gu, Y., Zeng, Z., Sabharwal, A., Zettlemoyer, L., Koh, P. W., Farhadi, A., Smith, N. A., and Hajishirzi, H. Olmo 3, 2026. URL https://arxiv.org/abs/2512.13961.

OpenAI. Update to gpt-5 system card: Gpt-5.2, December 2025a. URL https://deploymentsafety.openai.com/gpt-5-2.

OpenAI. Openai o3 and o4-mini system card, April 2025b. URL https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf.

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.-L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H. W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S. P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S. S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Łukasz Kaiser, Kamali, A., Kanitscheider, I., Keskar, N. S., Khan, T., Kilpatrick, L., Kim, J. W., Kim, C., Kim, Y., Kirchner, J. H., Kiros, J., Knight, M., Kokotajlo, D., Łukasz Kondraciuk, Kondrich, A., Konstantinidis, A., Kosic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C. M., Lim, R., Lin, M., Lin, S., Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S. M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D., Mu, T., Murati, M., Murk, O., Mély, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Ouyang, L., O'Keefe, C., Pachocki, J., Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., de Avila Belbute Peres, F., Petrov, M., de Oliveira Pinto, H. P., Michael, Pokorny, Pokrass, M., Pong, V. H., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotsted, B., Roussez, H., Ryder, N., Saltarelli, M., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B., Song, Y., Staudacher, N., Such, F. P., Summers, N., Sutskever, I., Tang, J., Tezak, N., Thompson, M. B., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J. F. C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J. J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W., and Zoph, B. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Paul, D., West, R., Bosselut, A., and Faltings, B. Making reasoning matter: Measuring and improving faithfulness of chain-of-thought reasoning, 2024. URL https://arxiv.org/abs/2402.13950.

Ribeiro, M. T., Wu, T., Guestrin, C., and Singh, S. Beyond accuracy: Behavioral testing of NLP models with CheckList. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J. (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.442. URL https://aclanthology.org/2020.acl-main.442/.

Russell, S. J. and Norvig, P. *Artificial Intelligence: A Modern Approach*. Prentice Hall, Upper Saddle River, NJ, 3 edition, 2010. URL http://repo.darmajaya.ac.id/5272/1/Artificial%20Intelligence-A%20Modern%20Approach%20%283rd%20Edition%29%20%28%20PDFDrive%20%29.pdf. Accessed: 2026-04-30.

Sainz, O., Campos, J., García-Ferrero, I., Etxaniz, J., de Lacalle, O. L., and Agirre, E. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 10776–10787, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.722. URL https://aclanthology.org/2023.findings-emnlp.722/.

Stolfo, A., Jin, Z., Shridhar, K., Schölkopf, B., and Sachan, M. A causal framework to quantify the robustness of mathematical reasoning with language models. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Proceedings of the 61st Annual Meeting of the Association for*

Computational Linguistics (Volume 1: Long Papers), pp. 545–561, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.32. URL https://aclanthology.org/2023.acl-long.32/.

Suzgun, M., Scales, N., Scharli, N., Gehrmann, S., Tay, Y., Chung, H. W., Chowdhery, A., Le, Q. V., Chi, E. H., Zhou, D., and Wei, J. Challenging big-bench tasks and whether chain-of-thought can solve them. In *Annual Meeting of the Association for Computational Linguistics*, 2022. URL https://api.semanticscholar.org/CorpusID:252917648.

Tafjord, O., Dalvi, B., and Clark, P. ProofWriter: Generating implications, proofs, and abductive statements over natural language. In Zong, C., Xia, F., Li, W., and Navigli, R. (eds.), *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 3621–3634, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.317. URL https://aclanthology.org/2021.findings-acl.317/.

Team, G. Gemini: A family of highly capable multimodal models, 2025. URL https://arxiv.org/abs/2312.11805.

Team, G., Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., Rouillard, L., Mesnard, T., Cideron, G., bastien Grill, J., Ramos, S., Yvinec, E., Casbon, M., Pot, E., Penchev, I., Liu, G., Visin, F., Kenealy, K., Beyer, L., Zhai, X., Tsitsulin, A., Busa-Fekete, R., Feng, A., Sachdeva, N., Coleman, B., Gao, Y., Mustafa, B., Barr, I., Parisotto, E., Tian, D., Eyal, M., Cherry, C., Peter, J.-T., Sinopalnikov, D., Bhupatiraju, S., Agarwal, R., Kazemi, M., Malkin, D., Kumar, R., Vilar, D., Brusilovsky, I., Luo, J., Steiner, A., Friesen, A., Sharma, A., Sharma, A., Gilady, A. M., Goedeckemeyer, A., Saade, A., Feng, A., Kolesnikov, A., Bendebury, A., Abdagic, A., Vadi, A., György, A., Pinto, A. S., Das, A., Bapna, A., Miech, A., Yang, A., Paterson, A., Shenoy, A., Chakrabarti, A., Piot, B., Wu, B., Shahriari, B., Petrini, B., Chen, C., Lan, C. L., Choquette-Choo, C. A., Carey, C., Brick, C., Deutsch, D., Eisenbud, D., Cattle, D., Cheng, D., Paparas, D., Sreepathihalli, D. S., Reid, D., Tran, D., Zelle, D., Noland, E., Huizenga, E., Kharitonov, E., Liu, F., Amirkhanyan, G., Cameron, G., Hashemi, H., Klimczak-Plucińska, H., Singh, H., Mehta, H., Lehri, H. T., Hazimeh, H., Ballantyne, I., Szpektor, I., Nardini, I., Pouget-Abadie, J., Chan, J., Stanton, J., Wieting, J., Lai, J., Orbay, J., Fernandez, J., Newlan, J., yeong Ji, J., Singh, J., Black, K., Yu, K., Hui, K., Vodrahalli, K., Greff, K., Qiu, L., Valentine, M., Coelho, M., Ritter, M., Hoffman, M., Watson, M., Chaturvedi, M., Moynihan, M., Ma, M., Babar, N., Noy, N., Byrd, N., Roy, N., Momchev, N., Chauhan, N., Sachdeva, N., Bunyan, O., Botarda, P., Caron, P., Rubenstein, P. K., Culliton, P., Schmid, P., Sessa, P. G., Xu, P., Stanczyk, P., Tafti, P., Shivanna, R., Wu, R., Pan, R., Rokni, R., Willoughby, R., Vallu, R., Mullins, R., Jerome, S., Smoot, S., Girgin, S., Iqbal, S., Reddy, S., Sheth, S., Põder, S., Bhatnagar, S., Panyam, S. R., Eiger, S., Zhang, S., Liu, T., Yacovone, T., Liechty, T., Kalra, U., Evci, U., Misra, V., Roseberry, V., Feinberg, V., Kolesnikov, V., Han, W., Kwon, W., Chen, X., Chow, Y., Zhu, Y., Wei, Z., Egyed, Z., Cotruta, V., Giang, M., Kirk, P., Rao, A., Black, K., Babar, N., Lo, J., Moreira, E., Martins, L. G., Sanseviero, O., Gonzalez, L., Gleicher, Z., Warkentin, T., Mirrokni, V., Senter, E., Collins, E., Barral, J., Ghahramani, Z., Hadsell, R., Matias, Y., Sculley, D., Petrov, S., Fiedel, N., Shazeer, N., Vinyals, O., Dean, J., Hassabis, D., Kavukcuoglu, K., Farabet, C., Buchatskaya, E., Alayrac, J.-B., Anil, R., Dmitry, Lepikhin, Borgeaud, S., Bachem, O., Joulin, A., Andreev, A., Hardin, C., Dadashi, R., and Hussenot, L. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Turpin, M., Michael, J., Perez, E., and Bowman, S. R. Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Valmeekam, K., Marquez, M., Sreedharan, S., and Kambhampati, S. On the planning abilities of large language models: a critical investigation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Wallace, E., Feng, S., Kandpal, N., Gardner, M., and Singh, S. Universal adversarial triggers for attacking and analyzing nlp. In *Conference on Empirical Methods in Natural Language Processing*, 2019. URL https://api.semanticscholar.org/CorpusID:201698258.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., and Fedus, W. Emergent abilities of large language models, 2022a. URL https://arxiv.org/abs/2206.07682.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., and Zhou, D. Chain-of-thought prompting elicits reasoning in large language

models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022b. Curran Associates Inc. ISBN 9781713871088.

White, C., Dooley, S., Roberts, M., Pal, A., Feuer, B., Jain, S., Shwartz-Ziv, R., Jain, N., Saifullah, K., Dey, S., Shubh-Agrawal, Sandha, S. S., Naidu, S., Hegde, C., LeCun, Y., Goldstein, T., Neiswanger, W., and Goldblum, M. Livebench: A challenging, contamination-limited llm benchmark, 2024. URL https://arxiv.org/abs/2406.19314.

Xu, C., Guan, S., Greene, D., and Kechadi, M.-T. Benchmark data contamination of large language models: A survey, 2024. URL https://arxiv.org/abs/2406.04244.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., and Qiu, Z. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Yang, S., Chiang, W.-L., Zheng, L., Gonzalez, J. E., and Stoica, I. Rethinking benchmark and contamination for language models with rephrased samples, 2023. URL https://arxiv.org/abs/2311.04850.

Zhu, K., Chen, J., Wang, J., Gong, N. Z., Yang, D., and Xie, X. Dyval: Dynamic evaluation of large language models for reasoning tasks, 2024. URL https://arxiv.org/abs/2309.17167.

Zhuo, T. Y., Li, Z., Huang, Y., Shiri, F., Wang, W., Haffari, G., and Li, Y.-F. On robustness of prompt-based semantic parsing with large pre-trained language model: An empirical study on codex. In Vlachos, A. and Augenstein, I. (eds.), *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 1090–1102, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.77. URL https://aclanthology.org/2023.eacl-main.77/.

Ethics Statement

We have taken multiple steps to ensure ethical compliance throughout the development of this work. The dataset was carefully curated to avoid copyright concerns, with diligent attention paid to sourcing and licensing, though minor oversights may occur despite best efforts. The benchmark construction process was designed with transparency and fairness in mind, ensuring that no demographic groups were singled out or negatively represented. We adhered to the NeurIPS Code of Ethics and confirm that no human subjects or sensitive personal data were involved.

A. Related Work

A.1. LLM reasoning benchmarks

The dominant tradition treats reasoning as accuracy on a held-out test set of fixed instances. GSM8K (Cobbe et al., 2021) introduced 8.5K grade-school math word problems; MATH (Hendrycks et al., 2021b) extended this to 12.5K competition-level items; MMLU (Hendrycks et al., 2021a) aggregates 57 multiple-choice domains. BIG-BENCH (et al., 2023) contributed 204 tasks with broad topical coverage, distilled by Suzgun *et al.* (Suzgun et al., 2022) into the 23-task BIG-BENCH HARD subset that requires non-trivial multi-step reasoning. Domain-specific benchmarks include HUMANEVAL (Chen et al., 2021) for code, STRATEGYQA (Geva et al., 2021) and ARC (Clark et al., 2018) for commonsense and science, LOGIQA (Liu et al., 2020) and FOLIO (Han et al., 2024) for formal logic, and PLANBENCH (Valmeekam et al., 2023) for planning. All of these score a single fixed phrasing per instance; none probe whether the model would have answered the same way under a paraphrased, reordered, or otherwise semantics-preserving rewrite.

A.2. Robustness and surface-form sensitivity

A separate literature has documented that NLP systems are sensitive to surface form: Ribeiro *et al.* (Ribeiro et al., 2020) use minimum-pair perturbations to expose behavioural failures; Naik *et al.* (Naik et al., 2018) stress-test NLI systems with antonym, length, and word-overlap perturbations; Wallace *et al.* (Wallace et al., 2019) show that universal adversarial triggers can flip predictions; and Zhuo *et al.* (Zhuo et al., 2023) report that small prompt perturbations damage chain-of-thought accuracy. For reasoning specifically, Mirzadeh, *et al.* (Mirzadeh et al., 2025) produce GSM-SYMBOLIC by templating 100 GSM8K items and varying the names and numerical values, finding that all models tested exhibit non-trivial performance variance and that adding a clause that "looks relevant" drops accuracy by up to 65%. Li *et al.* (Li et al., 2024) introduce GSM-PLUS, an adversarial variant of GSM8K with eight perturbation types. Stolfo *et al.* (Stolfo et al., 2023) present a causal framework for diagnosing arithmetic-reasoning robustness. INVARIANTBENCH differs from these in three respects: (i) it is not restricted to grade-school arithmetic; (ii) it covers twelve invariance axes spanning symbolic, structural, equivalence-based, and stylistic families; and (iii) every transformation is required to be *semantics-preserving* in a formal sense, so a degradation on a variant is unambiguously attributable to representation-dependence in the model rather than to a change in problem difficulty.

A.3. Faithfulness of chain-of-thought reasoning

Turpin *et al.* (Turpin et al., 2023) demonstrate that CoT explanations systematically misrepresent the actual reason for a prediction: biasing a multiple-choice prompt without altering the question can flip the answer while leaving the rationale apparently coherent. Lanham *et al.* (Lanham et al., 2023) measure CoT faithfulness by truncation, mistake-injection, and paraphrase, and find substantial unfaithfulness across model scales. Paul *et al.* (Paul et al., 2024) survey the broader landscape. These results imply that even when a model produces a fluent chain that arrives at the correct answer, the chain may not be the mechanism by which the answer was derived, an interpretation that *predicts* brittleness under semantics-preserving rewrites and that our results are consistent with.

A.4. Contamination and dynamic evaluation

The risk that strong benchmark scores reflect memorisation of seen instances rather than transferable competence is now well documented (Xu et al., 2024; Sainz et al., 2023; Yang et al., 2023; Golchin & Surdeanu, 2024). Dynamic benchmarks respond by regenerating items at evaluation time (Kiela et al., 2021; Zhu et al., 2024; White et al., 2024; Gulati et al., 2024). INVARIANTBENCH contributes to this line by treating each invariant variant as a contamination-resistant probe of the same underlying problem: a model that has memorised the base item still has to solve the variant, and the gap between the two is a lower bound on its reasoning above memorisation.

A.5. Compositional and systematic generalisation

A long line of work, beginning with Lake and Baroni (Lake & Baroni, 2017) and continued by Keysers *et al.*, and Kim and Linzen (Keysers et al., 2020; Kim & Linzen, 2020), asks whether neural models can generalise compositionally, for example, by handling novel combinations of seen primitives. INVARIANTBENCH sits adjacent to this literature but probes a different property: not generalisation across compositional recombinations of primitives, but *invariance under recoding of the same problem*. The two properties are related but distinct: a model can in principle be compositionally generalising yet representation-fragile, or representation-invariant yet non-compositional. Our experiments target the first axis directly.

A.6. Logical and symbolic evaluation

For first-order and modal logic, FOLIO (Han et al., 2024) and PROOFWRITER (Tafjord et al., 2021) provide formal deduction items; LOGIQA (Liu et al., 2020) provides multiple-choice logical items in natural language. The ∀uto∃val framework (Karia et al., 2024) synthesises items from formal grammars. None of these explicitly enforces invariance across surface forms of the same logical content. INVARIANTBENCH's *equivalent logical reformulation* axis (e.g. contrapositive, De Morgan's, prenex normal form) and *variable renaming* axis together enable the precise version of the question *"does the model solve the proposition or recognise the string?"*

A.7. Position of INVARIANTBENCH

INVARIANTBENCH unifies four threads of prior work, robustness testing (Ribeiro et al., 2020; Mirzadeh et al., 2025), contamination-resistant evaluation (Kiela et al., 2021; Zhu et al., 2024), faithfulness analysis (Turpin et al., 2023; Lanham et al., 2023), and reasoning-domain coverage (et al., 2023; Suzgun et al., 2022; Han et al., 2024), into a single framework whose primitive is the (seed, variant) pair. The result is a benchmark that scores a model not on whether it answered a particular instance, but on whether *the answer is the same across every semantics-preserving variant of that instance*.

B. Task Design and Descriptions

INVARIANTBENCH partitions its 1,200 evaluation instances across three high-level task families, Formal Logic & Discrete Structure (A), Quantitative Reasoning & Optimization (B), and Relational, Planning, Algorithmic and Semantic Reasoning (C), each decomposed into fine-grained sub-types that target distinct competency dimensions of language model reasoning, as per Table 5.

Table 5. Distribution of invariant types across the benchmark. Counts indicate the number of samples per transformation category.

Invariant Family	Transformation Type	Count
Symbolic & Linguistic	Entity Relabeling	304
	Style / Register Invariance	285
	Synonym / Semantic Paraphrase Invariance	260
	Variable Renaming	351
Structural & Representational	Equivalent Logical Reformulation	408
	Equivalent Mathematical Reformulation	244
	Premise Reordering	321
	Representation / Encoding Swap	227
Abstraction & Information-Structure	Abstraction Level Shift	329
	Distractor Injection / Removal	286
	Perspective / Role Reframing	227
	Redundant Restatement Invariance	358

B.1. Formal Logic & Discrete Structure (Category A)

Category A encompasses problems whose solution requires manipulating symbolic structures according to well-defined inference rules, set-algebraic identities, or constraint satisfaction procedures. Its 393 instances are distributed across five subtypes (A1–A5) and a physics-inspired representational subtype (A6), collectively probing whether models can maintain deductive soundness under surface-level reformulation.

B.1.1. A1, MULTI-QUANTIFIER FIRST-ORDER LOGIC (91 INSTANCES)

Sub-type A1 presents problems whose correct resolution demands chaining two or more quantifier alternations (∀/∃) within first-order logical formulas, requiring models to track variable scope, substitution legality, and the semantic distinction between universal and existential commitments simultaneously. Problems are constructed so that incorrect quantifier handling, collapsing $\forall x\, \exists y$ into $\exists y\, \forall x$, for instance, produces plausible but logically distinct conclusions, making shallow pattern matching insufficient. This design specifically emphasizes the capacity for scope-sensitive symbolic manipulation, which is well-established as a failure mode of large language models (**?**). The sub-type tests whether a model's stated conclusion is invariant to logically equivalent premise reorderings and quantifier-preserving reformulations, revealing whether the model has internalized quantifier semantics or relies on surface lexical cues.

B.1.2. A2. SET ALGEBRA + CARDINALITY WITH CONSTRAINTS (85 INSTANCES)

Sub-type A2 tasks models with reasoning about intersections, unions, complements, and power sets while simultaneously satisfying externally imposed cardinality constraints that couple the sizes of different set regions. The combinatorial structure is deliberately non-trivial: constraints are stated in ways that require models to construct implicit Venn-diagram partitions, enumerate mutually exclusive cases, and propagate numerical bounds through set-algebraic identities before arriving at a determinate answer. This sub-type isolates a reasoning layer distinct from pure arithmetic, cardinality reasoning requires understanding set containment and exclusion, not merely addition, and tests whether that layer is robust to notation changes (e.g., $|A \cap B|$ versus "the number of elements common to A and B"). Failures here indicate that models conflate the cardinality of a union with the sum of cardinalities without correctly subtracting intersection counts, a well-documented error type.

B.1.3. A3. CONDITIONAL OVERRIDE TRANSFORMATION SYSTEMS (59 INSTANCES)

Sub-type A3 defines small rule systems in which a base transformation is conditionally overridden by one or more priority rules that fire when specified guard conditions are satisfied, requiring models to simulate the application of a layered rule hierarchy to a given input state. The structure mirrors production-rule systems and priority-based rewriting, making it a proxy for understanding how models handle exception-driven logic in programs or regulatory texts. Correct reasoning requires identifying which rules are applicable, resolving conflicts via the stated priority ordering, and verifying that no applicable override has been overlooked, a form of exhaustive case analysis. Sub-type A3 thereby tests whether models can maintain a complete and consistent mental representation of a stateful rule system, a capacity that underlies reliable code reasoning and contract interpretation.

B.1.4. A4, CONSTRAINT SATISFACTION IN PROPOSITIONAL LOGIC (65 INSTANCES)

Sub-type A4 instantiates classical propositional constraint satisfaction problems in which a set of Boolean variables must be assigned truth values consistent with a conjunction of clauses, presented in varying surface encodings (CNF, natural language conditionals, truth tables, or scenario descriptions). The problems are designed to have small but non-trivial solution spaces, typically requiring case analysis or unit propagation to resolve, so that both systematic search and shallow heuristics are distinguishable by their outputs. This sub-type directly probes unit propagation, pure literal elimination, and backtracking, which are the core algorithmic primitives of SAT solving, albeit at a scale where explicit enumeration remains feasible. The invariance dimension tests whether a model reaches the same satisfying assignment (or correctly identifies unsatisfiability) regardless of whether clauses are presented in conjunctive normal form, as nested conditionals, or as a truth-table specification.

B.1.5. A5. INDUCTIVE RULE + FORMAL DEDUCTION (47 INSTANCES)

Sub-type A5 supplies models with a finite base case and one or more inductive rules, then asks them to derive properties of the induced structure or to determine the status of a query element after a specified number of rule applications, combining inductive definition with forward-chaining deduction. The construction mirrors Datalog-style reasoning and property verification in program analysis, positioning this sub-type as a test of recursive symbolic inference rather than pattern interpolation from training data. Correct solutions require recognizing the fixed-point semantics of the rule system and understanding which facts can be derived versus which remain undetermined, a distinction that naive models frequently collapse. Sub-type A5 thus evaluates whether models can faithfully simulate a bounded inference engine, a capability critical for formal verification, theorem proving assistance, and code analysis tasks.

B.1.6. A6. UNIT/REPRESENTATION INVARIANCE IN PHYSICS-STYLE WORD PROBLEMS (46 INSTANCES)

Sub-type A6 constructs quantitative word problems drawn from elementary physics or engineering contexts, motion, energy, pressure, rate, where the same physical scenario is described using different unit systems or measurement conventions (e.g., SI versus imperial, radians versus degrees). The expected answer is numerically identical after appropriate unit conversion, and correctness hinges on recognizing and applying those conversions before performing algebraic manipulations. This sub-type is specifically designed to test *representational invariance*: whether a model's computed answer changes spuriously when surface quantities are re-expressed in equivalent but syntactically distinct forms, a failure mode that exposes over-reliance on memorized numerical patterns rather than principled dimensional analysis. Problems are calibrated so that dimensional confusion produces answers differing by well-known conversion factors (e.g., $\times 1000$, $\times 3.281$), making incorrect outputs interpretable and systematically analyzable.

B.2. Quantitative Reasoning & Optimization (Category B)

Category B assembles 399 instances targeting numerical reasoning in structured algebraic, probabilistic, and optimization settings where multiple interdependent quantities must be jointly tracked and manipulated. Sub-types B1–B5 span sequential constraint propagation, coupled equation solving, probabilistic inference, and structured optimization, collectively probing whether quantitative competence is robust to reformulations that preserve mathematical content while varying presentation.

B.2.1. B1. SEQUENTIAL ALGEBRAIC SYSTEMS WITH LAYERED CONSTRAINTS (81 INSTANCES)

Sub-type B1 presents multi-step algebraic problems in which each step introduces a new constraint that depends on the output of the preceding step, forming a directed dependency chain that must be traversed in order. The layered structure is designed to amplify error propagation: an incorrect intermediate value contaminates all downstream quantities, so models that approximate rather than compute exactly will exhibit systematic deviations. This sub-type tests sequential algebraic processing, a form of working-memory-dependent computation, and is invariant under reordering of independent constraint layers, providing a fine-grained probe of whether models correctly identify which constraints are mutually dependent and which can be resolved in any order. It is particularly diagnostic of whether models maintain exact symbolic state across multiple reasoning steps or collapse to heuristic estimates at intermediate stages.

B.2.2. B2. OPTIMIZATION OVER COUPLED ALGEBRAIC EQUATIONS (81 INSTANCES)

Sub-type B2 tasks models with finding an optimal value (maximum, minimum, or extremal configuration) for an objective function that depends on several variables linked by algebraic coupling constraints, requiring both constraint propagation and optimization reasoning in a unified pass. The problems are designed to preclude brute-force enumeration at realistic scales while remaining analytically tractable via substitution, Lagrange-style reasoning, or boundary analysis, thus distinguishing models that understand algebraic structure from those that attempt grid search. This sub-type evaluates whether models can identify active constraints, reduce degrees of freedom, and correctly apply optimality conditions, a composite skill set underlying linear programming, resource allocation, and engineering design problems. Invariance testing here examines whether the optimal value is stable when the same coupling constraints are expressed via different algebraic rearrangements.

B.2.3. B3. PROBABILISTIC REASONING WITH CONDITIONAL DEPENDENCE (62 INSTANCES)

Sub-type B3 requires computing probabilities in scenarios involving explicit conditional dependence structures, chains, forks, or colliders, where naive application of marginalization or Bayes' rule without correct treatment of dependence yields wrong answers. The problems are constructed so that independence assumptions must be verified or refuted from the stated generative process, and where incorrect assumption of independence produces a plausible but systematically biased answer. This sub-type directly probes Bayesian reasoning fidelity, testing whether models can identify d-separation conditions, correctly apply the chain rule of probability, and update beliefs through multi-hop conditional distributions. The invariance dimension assesses whether computed probabilities remain stable under equivalent reformulations of the same joint distribution (e.g., specifying a joint table versus specifying marginals and conditionals).

B.2.4. B4. INTERDEPENDENT EQUATION SYSTEMS WITH STRUCTURAL CONSTRAINTS (88 INSTANCES)

Sub-type B4 presents systems of equations that are not purely linear, they may involve product terms, piecewise definitions, or domain restrictions, embedded within a structural constraint (e.g., integrality, non-negativity, ordering relations) that

eliminates some algebraic solutions and demands careful case disambiguation. The construction tests whether models can simultaneously manage algebraic manipulation and constraint-driven filtering, a combination that frequently fails due to premature discarding of valid cases or retention of constraint-violating solutions. This sub-type evaluates a capacity closely aligned with mixed-integer and combinatorial algebraic reasoning, with direct relevance to scheduling, resource planning, and combinatorial search problems encountered in applied AI. Invariance here is tested by restating structural constraints in logically equivalent but syntactically distinct forms, verifying that the feasible solution set is correctly identified regardless of surface phrasing.

B.2.5. B5. CONSTRAINT-DRIVEN ALGEBRAIC REASONING PROBLEMS (87 INSTANCES)

Sub-type B5 presents algebraic reasoning tasks in which a set of explicit domain constraints (range restrictions, modular conditions, divisibility requirements) is the *primary* driver of the solution rather than a secondary filter on algebraic solutions. Models must reason backward from constraints to infer permissible variable assignments, often without forming an explicit equation system, testing constraint-first rather than equation-first problem-solving strategies. This sub-type probes a reasoning mode distinct from standard algebra, it requires treating constraints as generators of a feasible set and then identifying which algebraic expressions are fixed or bounded over that set. It is particularly relevant to formal verification, program analysis, and type-system reasoning, where constraints govern what is possible before any specific value is computed.

B.3. Relational, Planning, Algorithmic and Semantic Reasoning (Category C)

Category C comprises 408 instances spanning graph-structured, multi-agent, code-level algorithmic, and language-semantic reasoning tasks. These sub-types test whether models can perform structured symbolic search, infer structural equivalences, and process semantic information correctly under surface reframings that leave underlying relational structure intact.

B.3.1. C1. ALGEBRA, PSEUDOCODE, AND ALGORITHMIC REASONING (67 INSTANCES)

Sub-type C1 interleaves algebraic sub-problems with pseudocode execution and algorithmic reasoning, requiring models to trace through a specified procedure, evaluate expressions at intermediate steps, and report the final state or output variable. Problems are constructed so that correct answers depend on precise simulation of the algorithm rather than high-level pattern recognition, with loop-carried dependencies, conditional branches, and accumulator updates that must be tracked exactly. This sub-type is most closely aligned with code understanding and program analysis, and its invariance dimension tests whether algorithmic outputs are stable under semantics-preserving refactorings such as variable renaming, loop unrolling by one step, or rewriting recursive calls iteratively. Failures on this sub-type are interpretable as breakdowns in the model's capacity to simulate a computational process rather than describe it abstractly.

B.3.2. C2. GRAPH SEARCH AND TRAVERSAL REASONING (105 INSTANCES)

Sub-type C2 presents explicit graph structures (adjacency lists, edge sets, or natural-language descriptions of node relationships) and asks models to compute properties that require search or traversal, shortest paths, reachability, connected component membership, topological order, or cycle detection. Graph-structured reasoning is a canonical testbed for relational inference, and this sub-type constitutes the largest in Category C (105 instances), reflecting the breadth and centrality of graph reasoning to planning, knowledge-base querying, and network analysis. The invariance dimension tests whether computed properties are preserved under node/edge relabeling and under equivalent graph representations (e.g., adjacency matrix versus edge list), directly probing whether models perform genuine structural reasoning or exploit incidental surface regularities such as alphabetical node orderings. Correct performance requires implementing a mental traversal procedure rather than pattern-matching on familiar graph shapes.

B.3.3. C3. PATH-BASED SEARCH AND MULTI-AGENT INTERACTION REASONING (96 INSTANCES)

Sub-type C3 extends single-agent graph traversal to scenarios involving multiple agents, resource constraints, or competitive/cooperative interaction dynamics, where the set of feasible paths is restricted not only by graph topology but also by the joint state of all agents. Problems include resource-constrained routing, pursuit-evasion, synchronized multi-agent navigation, and turn-based interaction scenarios that require reasoning about the joint trajectory space rather than any individual agent's path. This sub-type tests a form of combinatorial planning that is central to robotics, game-playing AI, and multi-agent simulation, where the state space is exponential in the number of agents and correct solutions require disciplined

state-space search. The invariance dimension examines whether solutions are stable under equivalent relabeling of agents and isomorphic restructurings of the interaction graph.

B.3.4. C4. GRAPH ISOMORPHISM AND STRUCTURAL MAPPING (104 INSTANCES)

Sub-type C4 presents pairs of graphs (or relational structures) and requires models to determine whether an isomorphism exists, construct an explicit vertex mapping if it does, or identify the largest common substructure. This sub-type directly targets structural relational reasoning at the level of whole-graph equivalence rather than local property computation, a task for which no known polynomial-time algorithm exists in general and which therefore taxes genuine structural search rather than cached pattern recognition. Correct performance requires verifying degree sequence compatibility, exploring candidate mappings, and pruning infeasible partial mappings through constraint propagation, a disciplined combinatorial procedure. The invariance dimension tests whether isomorphism judgments are preserved under vertex relabeling of one graph, probing whether the model's equivalence decision reflects structural content rather than coincidental label matching.

B.3.5. C5. LANGUAGE, SEMANTICS & INFORMATION STRUCTURE-BASED REASONING (36 INSTANCES)

Sub-type C5 presents tasks grounded in linguistic and semantic structure, coreference resolution, presupposition projection, implicature cancellation, information-structural focus/topic decomposition, and discourse-level entailment, where correct answers depend on tracking meaning-relevant linguistic distinctions that are absent from purely syntactic surface forms. This sub-type is the smallest in the benchmark (36 instances) and is designed to test the interface between formal semantic reasoning and natural language understanding, a frontier that requires models to apply linguistically informed inference rules rather than general world-knowledge associations. Problems are constructed to make superficially plausible wrong answers available for models that process sentences holistically without parsing semantic scope, presupposition triggers, or implicature cancellation environments. The invariance dimension tests whether semantic inferences are stable under paraphrase and register variation while remaining sensitive to meaning-altering syntactic changes, providing a fine-grained probe of semantic compositionality.

C. Invariant Type Descriptions

INVARIANTBENCH evaluates model robustness via twelve invariant transformation types organized into three families: Symbolic and Linguistic, Structural and Representational, and Abstraction and Information-Structure, as shown in Table 6. Each transformation constructs a semantically equivalent problem variant; a robust model should produce the same answer on both the original and transformed instances, whereas a fragile model reveals its reliance on surface features that change under the transformation.

Family	Transformation τ	Operation
Symbolic & Linguistic	Variable renaming	Injective renaming of free and bound variables.
	Entity relabeling	Consistent relabeling of named entities and predicates.
	Synonym / paraphrase	Lexical substitution and sentence-level paraphrasing.
	Style / register	Register shifts without altering semantic content.
Structural & Representational	Premise reordering	Permutation of independent clauses under dependency constraints.
	Redundant restatement	Addition of paraphrased clauses without new information.
	Distractor injection / removal	Addition or removal of semantically irrelevant clauses.
	Representation swap	Conversion across equivalent encodings (text, symbolic, tabular).
Abstraction & Information Structure	Equivalent logical reformulation	Sound logical equivalence transformations.
	Equivalent mathematical reformulation	Algebraic and arithmetic equivalences.
	Abstraction level shift	Change in granularity while preserving task semantics.
	Perspective / role reframing	Reassignment of roles or viewpoints without altering relations.

Table 6. The twelve invariance axes in INVARIANTBENCH.

C.1. Symbolic and Linguistic Invariants

Symbolic and linguistic invariants modify the *surface vocabulary* of a problem, the names of entities, the phrasing of conditions, the syntactic register of the question, while leaving the underlying logical, mathematical, or relational content entirely unchanged. These four invariant types collectively contribute 1,200 instances and probe whether model performance is driven by semantic understanding or by lexical and symbolic surface cues.

C.1.1. ENTITY RELABELING (304 INSTANCES)

Entity Relabeling replaces the named participants in a problem (persons, cities, objects, organizations) with semantically neutral alternatives, typically unfamiliar proper nouns or placeholder identifiers, so that any answer derived from world-knowledge associations about the original entities would be disrupted. The transformation preserves the complete logical and relational structure of the problem while eliminating the possibility of bypassing reasoning by recalling facts about the specific entities named. This invariant is especially diagnostic in relational and planning tasks (Category C) where familiar entity names (canonical graph nodes, well-known route networks) may trigger memorized traversal patterns. Correct invariance here provides evidence that a model is processing the relational structure of the problem rather than pattern-completing from training-distribution entity co-occurrences.

C.1.2. STYLE / REGISTER INVARIANCE (285 INSTANCES)

Style and Register Invariance rewrites the problem statement in a different communicative register, formal academic prose to informal colloquial phrasing, passive to active voice, second-person to third-person narration, or telegraphic notation to full prose, without altering any factual content, logical connectives, or quantitative values. The transformation targets the well-documented sensitivity of language models to stylistic surface variation, which can cause identical questions to receive different answers when phrased in unfamiliar registers. This invariant is particularly relevant to deployment robustness: models used in consumer-facing settings must produce consistent answers across the broad stylistic variation of user inputs, and failures on this invariant reveal that internal representations conflate style with substance. Correct invariance demonstrates that the model's reasoning is anchored to propositional content rather than surface fluency patterns.

C.1.3. SYNONYM / SEMANTIC PARAPHRASE INVARIANCE (260 INSTANCES)

Synonym and Semantic Paraphrase Invariance substitutes key content words and phrases with semantically equivalent alternatives, synonyms, definitional expansions, or well-known paraphrases, that preserve truth conditions while changing lexical form. Unlike register variation, which affects global stylistic register, this transformation targets local lexical substitutions that individually preserve meaning but collectively produce a surface form distant from the original, testing whether the model's internal semantic representation is robust to vocabulary diversity. The transformation is constructed to avoid pragmatic shifts: substitutions are selected so that the paraphrase is truth-conditionally equivalent in context, ruling out cases where synonymy fails under specific presuppositions. This invariant most directly probes lexical-semantic grounding, whether concepts are represented in terms of their meanings or their most frequent surface lexicalizations.

C.1.4. VARIABLE RENAMING (351 INSTANCES)

Variable Renaming substitutes all symbolic identifiers in a formal problem, mathematical variable names, logical proposition symbols, set labels, graph node names, with fresh symbols drawn from a different alphabet or naming convention, while maintaining all structural relationships among identifiers identically. This transformation directly tests whether model reasoning over symbolic expressions is anchored to the algebraic/logical structure or to the specific symbols used, a distinction that is definitional for genuine formal reasoning: a sound algebraic solver should be entirely indifferent to whether the unknown is called x, z, α, or `myVar`. Variable Renaming is the most frequently applied invariant type (351 instances), reflecting its fundamental importance across all three task categories and its tight connection to the notion of syntactic invariance in formal systems. Failures, where renaming a variable causes a model to solve a different equation or draw a different logical conclusion, constitute unambiguous evidence of surface-level symbol-matching rather than formal symbolic reasoning.

C.2. Structural and Representational Invariants

Structural and representational invariants modify the *organization and encoding* of problem information, the logical form, mathematical notation, premise order, or representational modality, while preserving semantic content. These four invariant types probe the depth of structural comprehension: a model that has internalized the mathematical or logical structure of a problem should be unaffected by surface restructuring, while a model exploiting positional or notational regularities will exhibit systematic failures.

C.2.1. EQUIVALENT LOGICAL REFORMULATION (408 INSTANCES)

Equivalent Logical Reformulation rewrites the logical structure of a problem using a known equivalence transformation, replacing $P \Rightarrow Q$ with $\neg Q \Rightarrow \neg P$ (contrapositive), distributing quantifiers, applying De Morgan's laws, converting between disjunctive and conjunctive normal forms, or replacing biconditionals with conjunctions of conditionals, producing a problem whose propositional content is provably identical. This is the most frequently applied invariant (408 instances, matching Category C count), reflecting the ubiquity of equivalent logical reformulation as a natural source of surface variation in reasoning tasks. Correct invariance requires that the model recognize the semantic equivalence of the reformulated structure, a competency that is foundational to logical reasoning and formal argumentation. Models that fail this invariant are demonstrably operating on surface syntactic form rather than logical content, a critical failure mode for applications in formal verification, legal reasoning, and automated theorem proving.

C.2.2. EQUIVALENT MATHEMATICAL REFORMULATION (244 INSTANCES)

Equivalent Mathematical Reformulation transforms the algebraic or arithmetic structure of quantitative problem conditions using known mathematical identities, distributing products, factoring sums, applying commutativity and associativity, rewriting fractions as decimals, or converting between parametric and explicit representations, so that all quantities and relationships are preserved but their syntactic form changes. This invariant isolates the algebraic layer of mathematical understanding from the arithmetic execution layer: a model that computes correctly on one form but fails on an equivalent rearrangement has demonstrated shallow algebraic competence grounded in lexical form recognition rather than structural mathematical understanding. The transformation is particularly stringent for Category B tasks, where algebraic coupling and constraint structures are the primary reasoning substrate, and failures directly reveal whether models have internalized mathematical identities as generative principles.

C.2.3. PREMISE REORDERING (321 INSTANCES)

Premise Reordering permutes the order in which problem conditions are stated, placing the goal before the constraints, reversing the order of case definitions, interleaving conditions that were originally grouped by type, without adding or removing any information. This invariant directly tests positional sensitivity: models that construct reasoning chains by attending locally to adjacent sentences or that leverage positional priors ("earlier premises are more important") will exhibit degraded performance under arbitrary reorderings that a logically competent reasoner would be indifferent to. Premise Reordering is the second most common structural invariant (321 instances), reflecting the natural variation in premise presentation order across problem formulations in the wild. Correct invariance on this transformation is a necessary condition for robust multi-step reasoning, since real-world problems rarely present information in the order most convenient for sequential processing.

C.2.4. REPRESENTATION / ENCODING SWAP (227 INSTANCES)

Representation and Encoding Swap translates the entire problem from one representational modality to an equivalent one, graph adjacency list to adjacency matrix, set notation to characteristic-function description, truth table to Boolean formula, or decimal numbers to binary/hexadecimal, while preserving all structural information. This invariant targets the representational layer of understanding: a model that has genuinely internalized the abstract structure being described should be able to process it regardless of the particular encoding, while a model that has learned to process specific surface encodings will fail when the same information is encoded differently. Encoding Swap tests are especially powerful diagnostics for graph reasoning (C2, C3, C4) and propositional logic (A4) tasks, where multiple standard encodings co-exist and competent human solvers routinely translate between them. Failures on this invariant are evidence of format-specific pattern matching rather than abstract structure comprehension.

C.3. Abstraction and Information-Structure Invariants

Abstraction and information-structure invariants modify the *level of detail, perspective, or information density* at which a problem is presented, testing whether models can retrieve the correct answer when surface information is reorganized, augmented with irrelevant material, or reframed from a different viewpoint. These invariants probe the robustness of reasoning to the kind of contextual and framing variation that characterizes real-world language use.

C.3.1. ABSTRACTION LEVEL SHIFT (329 INSTANCES)

Abstraction Level Shift rewrites the problem at a higher or lower level of conceptual granularity, replacing concrete numerical instantiations with parameterized variables (upward abstraction), or replacing abstract structural descriptions with fully concrete worked examples (downward instantiation), while preserving the specific answer required. This transformation tests whether a model's reasoning capability is calibrated to a specific level of abstraction in the training distribution or is genuinely level-agnostic; models overfit to, say, concrete arithmetic problems will fail on their algebraically generalized counterparts. The upward variant is particularly relevant to evaluating algebraic generalization: can a model that correctly solves "if $x = 3$, compute $2x + 1$" also correctly evaluate the general expression $2x + 1$ for a symbolically stated x? Abstraction Level Shift is the third most common invariant (329 instances), reflecting its broad applicability across all task categories.

C.3.2. DISTRACTOR INJECTION / REMOVAL (286 INSTANCES)

Distractor Injection and Removal adds semantically irrelevant but topically plausible information to a problem (injection) or strips away such distractors that were present in the original (removal), testing whether model performance degrades in the presence of irrelevant context or improves spuriously when distractors are removed. Injected distractors are constructed to be relevant in surface form, they mention entities and quantities that appear in the problem, but to be logically irrelevant to the correct solution, so that a model attending to all mentioned quantities will be misled while a focused reasoner will correctly ignore them. This invariant is closely related to the classic comprehension construct of "reading to find relevant information" and is diagnostic of selective attention and relevance filtering, two capabilities critical for robust performance in long-context reasoning. Removal of distractors from the original problem tests the complementary direction: whether the original problem's answer is recovered when the confounding material is excised.

C.3.3. PERSPECTIVE / ROLE REFRAMING (227 INSTANCES)

Perspective and Role Reframing restates the problem from the vantage point of a different participant, switching from a first-person agent to an observer, from a questioner to a responder, from a buyer to a seller in a transaction scenario, while leaving the underlying facts and relational structure unchanged. This invariant tests whether a model's reasoning is anchored to the objective structure of the problem or to the particular role assignment used to introduce it, a failure mode that would manifest as role-specific answer biases without corresponding changes in the correct answer. Perspective Reframing is particularly diagnostic for multi-agent planning tasks (C3) and relational reasoning tasks (C2, C4) where role assignments are explicit and the same graph-structural or logical content is accessed from different positions in the interaction. Correct invariance on this transformation provides evidence for genuinely objective structural reasoning rather than role-anchored heuristic responses.

C.3.4. REDUNDANT RESTATEMENT INVARIANCE (358 INSTANCES)

Redundant Restatement Invariance augments the problem with additional sentences that restate already-given information in different words, paraphrasing a key constraint, re-expressing a given quantity using equivalent arithmetic, or confirming an already-stated relationship, without adding any new information or contradicting existing content. This invariant tests robustness to verbosity: a logically competent model should recognize the restatements as redundant and neither weight them as new evidence nor become confused by the apparent repetition, whereas a model that tracks information by occurrence frequency may spuriously amplify the importance of redundantly stated facts. Redundant Restatement is the most frequent abstraction-layer invariant (358 instances), reflecting both the ease of generating such variants and their diagnostic value for probing information-redundancy sensitivity. Correct invariance on this transformation is a necessary condition for reliable reasoning over real-world texts, which routinely contain paraphrastic repetition for rhetorical and communicative purposes.

D. More Details on Benchmark Construction of INVARIANTBENCH

D.1. Annotator Recruitment

We recruited annotators aged 24–30, all with engineering backgrounds and prior exposure to logic- or reasoning-oriented tasks. Participation was voluntary, and annotators worked independently without external incentives to reduce bias induced by reward-driven labeling. Before annotation began, we conducted structured onboarding sessions covering task objectives, invariance principles, and quality expectations. Ethical considerations and data handling protocols were strictly enforced throughout the process. This ensured that all annotations were produced under consistent and controlled conditions aligned with the benchmark design.

D.2. Annotator Training and Protocol

We developed a structured training pipeline to ensure consistency across the full invariance taxonomy. Annotators were first introduced to the three reasoning families and their associated task structures, with emphasis on how semantic equivalence must be preserved under surface transformations. For each invariance type, we provided curated examples and counter-examples illustrating valid and invalid transformations, including cases where subtle semantic drift could occur.

Training further emphasized the four transformation families (symbolic, structural, abstraction-level, and representation-based invariances) and their fine-grained subtypes, ensuring annotators could reliably distinguish permissible reformulations from those that alter underlying semantics. We also trained annotators to explicitly verify equivalence under the invariance contract, particularly for logical, mathematical, and compositional transformations.

To ensure robustness across visual and textual variability, we incorporated guided exercises on handling ambiguous cases, including multi-step reasoning chains, compositional transformations, and distractor-rich formulations. Annotators practiced constructing both multiple-choice and open-ended variants, with emphasis on preserving answer invariance while varying surface structure.

D.3. Quality Control and Verification

To ensure correctness and invariance integrity, each problem–variant set is validated through a two-stage verification pipeline. The first stage consists of consistency checks that enforce structural equivalence across variants and identify invalid transformations. These checks ensure that problem constraints are preserved, that no unintended assumptions are introduced or removed, and that symbolic or structural modifications do not alter the underlying solution space. Where applicable, deterministic recomputation is used to confirm answer consistency across transformed instances. Any instance that violates these constraints is rejected or returned for regeneration.

In the second stage, expert reviewers with training in logic, mathematics, and formal reasoning independently inspect a stratified sample of the automatically validated data. This manual evaluation focuses on detecting semantic drift not captured by automated procedures, including subtle ambiguity introduced through rephrasing, compositional restructuring, or multi-step transformations. Reviewers also assess whether invariance is preserved under the full transformation contract and whether difficulty levels remain consistent across paired variants. When discrepancies arise, adjudication by a senior expert is used to resolve inconsistencies and enforce uniform acceptance standards.

Only instances that satisfy both automated validation and expert review criteria are retained in the final benchmark. This dual-layer verification process produces a high-confidence dataset of invariantly paired reasoning tasks designed to enable controlled evaluation of representation sensitivity in modern large language models.

Finally, after these iterative calibration rounds, final dataset sample is generated. This iterative process ensured high-fidelity alignment between intended invariance transformations and final benchmark instances.

E. More details on Experiments and Evaluation

E.1. Evaluation Protocol

We evaluate all models using a controlled prompting and automated judging pipeline to ensure consistency and comparability across variants.

Answer Generation. Each model is prompted with a strict output constraint to minimize verbosity and reduce format-induced variance. Specifically, responses are restricted to at most a single word or a short sentence (10–12 words), with no explanation or extraneous text. This enforces alignment with the benchmark's exact-answer evaluation format and avoids partial-credit ambiguity.

Prompt for Answering Model

Answer the following question as concisely as possible, using at most one word or a short sentence (10–12 words). Give only the answer itself. No explanation, no preamble, and no extra punctuation.

Question: `<question>`

System Prompt for Judge Model

You are a strict, impartial answer evaluator.

You will be given:

- A question
- The correct answer
- A model's response

Your job: determine whether the model's response is CORRECT or INCORRECT.

Rules:

- Match to given correct answer, instead of reasoning.
- Be flexible about formatting.
- Be strict about semantic content — a vague or partially correct answer is INCORRECT.
- Ignore extra whitespace, punctuation, capitalisation differences.
- If the model's response contains the correct answer within a sentence, mark it CORRECT.
- If the model says "I don't know" or gives an irrelevant answer, mark it INCORRECT.
- Any ambiguous answer is INCORRECT.

Respond with ONLY a JSON object in this exact format (no markdown fences, no extra text):

```
{"score": 1, "reason": "one sentence (7--10 words) explaining the decision"}
```

or

```
{"score": 0, "reason": "one sentence (7--10 words) explaining the decision"}
```

Base Prompt for Judge Model

Question:

`<question>`

Correct Answer:

`<correct_answer>`

Model Response:

`<model_response>`

Judging Protocol. All responses are evaluated using a separate judge model under a strict binary scoring scheme. The judge compares the model output directly against the ground-truth answer, allowing minor formatting differences (e.g., whitespace, capitalization), but rejecting semantically incomplete or approximate answers. If the correct answer appears within a longer response, it is accepted; otherwise, any ambiguity results in a score of 0. The judge returns a structured JSON object containing a binary score and a brief justification.

Decoding and Inference Settings. We set `temperature` to 0.1 to reduce stochasticity and ensure stable outputs across variants. The maximum generation length is set to 600 tokens for standard models and 20,000 tokens for reasoning-augmented models, ensuring sufficient capacity for intermediate reasoning when required. The judge model is allocated a maximum of 1,200 tokens. All evaluations are conducted via the OPENROUTER API with default parameters otherwise unchanged, with API usage costs ranging from approximately \$0.10 to \$15 per 1M input/output tokens; the full set of experiments, including initial testing and ablations, incurred a total cost of approximately \$550.

Rationale. This evaluation design isolates correctness from verbosity and stylistic variation, ensuring that measured differences reflect reasoning performance rather than output formatting. The strict judge further enforces consistency across models, enabling reliable comparison under invariant transformations.

E.2. Evaluation of LLM-as-a-Judge for Open-ended Evaluation

To evaluate *Gemma-4-26B-A4B-Instruct* as an automatic judge, we selected a balanced set of 240 samples spanning all task categories and sub-types. Each sample consists of a question, ground truth, and model-generated output. Three independent human annotators evaluated each response, assigning binary labels: $\{1 = \text{correct}, 0 = \text{incorrect}\}$.

To obtain a robust human reference signal, we construct a **majority-vote label** for each sample:

$$y_i^{\text{maj}} = \text{mode}\big(y_i^{(1)}, y_i^{(2)}, y_i^{(3)}\big), \tag{3}$$

which serves as the primary proxy for ground-truth judgment in all comparisons with the LLM judge.

We evaluate the LLM judge against human annotations using multiple complementary agreement metrics, computed both globally and stratified by *Task Type* and *Task Sub-Type*. This enables us to assess not only overall alignment with human judgments but also the stability of the judge across different evaluation regimes.

E.2.1. COHEN'S KAPPA

Cohen's kappa (κ) measures chance-corrected agreement between two raters. It is defined as:

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \tag{4}$$

where p_o denotes the observed agreement and p_e is the expected agreement under independence:

$$p_e = \sum_{c \in \{0,1\}} p_c^{(1)} p_c^{(2)}. \tag{5}$$

We compute κ between the LLM judge and each individual annotator, as well as between the LLM and the majority vote label:

$$\kappa_{\text{LLM,maj}}. \tag{6}$$

This formulation allows direct comparison of the LLM judge against both individual human variability and aggregated human consensus.

E.2.2. ACCURACY AGAINST MAJORITY VOTE

We further report raw agreement with the majority label:

$$\text{Accuracy} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\big(y_i^{\text{LLM}} = y_i^{\text{maj}}\big), \tag{7}$$

which captures direct alignment with human consensus but does not correct for chance agreement effects.

E.2.3. FLEISS' KAPPA (HUMAN RELIABILITY)

To quantify inter-annotator reliability, we compute Fleiss' kappa (κ_F), defined as:

$$P_i = \frac{1}{n(n-1)} \sum_j n_{ij}(n_{ij} - 1), \quad \bar{P} = \frac{1}{N} \sum_{i=1}^{N} P_i, \tag{8}$$

$$p_j = \frac{1}{Nn} \sum_i n_{ij}, \quad \bar{P}_e = \sum_j p_j^2, \tag{9}$$

$$\kappa_F = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}. \tag{10}$$

This serves as an empirical *human agreement ceiling*, against which the LLM judge is compared.

E.2.4. STRATIFIED EVALUATION ACROSS TASK STRUCTURE

Beyond aggregate performance, we perform a fine-grained analysis conditioned on *Task Type* and *Task Sub-Type*. For each subset $\mathcal{D}_t$, we compute:

$$\text{Accuracy}_t, \quad \kappa_t^{\text{LLM vs maj}}, \quad \kappa_t^{\text{human}} \tag{11}$$

where κ_t^{human} denotes the mean pairwise Cohen's kappa among human annotators within the subset. This allows us to measure whether the LLM judge achieves *human-level agreement locally within specific reasoning regimes*, rather than only in aggregate.

E.2.5. HUMAN-LEVEL ALIGNMENT CRITERION

We define the LLM judge as achieving human-level performance if:

$$\kappa_{\text{LLM, maj}} \approx \kappa_{\text{human}}, \tag{12}$$

and if its accuracy lies within the empirical confidence interval of human inter-annotator agreement. We additionally report bootstrap confidence intervals over accuracy and conduct permutation tests to assess whether observed agreement significantly exceeds random alignment.

Across all evaluations, results are reported both globally and disaggregated by task structure, providing a comprehensive view of the reliability and calibration of LLM-based automatic evaluation.

Task Type	**N**	**Match**	$\kappa_{\text{LLM-Human}}$	κ_{Human}
All	4800	0.9998	0.9995	0.9656
A. Formal Logic & Discrete Structure	1264	1.0000	1.0000	0.9764
B. Quantitative Reasoning & Optimization	1960	1.0000	1.0000	0.9492
C. Relational, Planning, Algorithmic and Semantic Reasoning	1576	0.9994	0.9986	0.9661

Table 7. **Agreement between LLM-as-a-Judge and Human Judge across Task Types.** We report sample size (N), accuracy against human majority vote, Cohen's kappa between the LLM judge and human majority, and inter-annotator agreement among humans. Across all categories, the LLM judge achieves near-perfect alignment with human consensus, closely matching or exceeding the human inter-annotator agreement ceiling.

E.2.6. INTERPRETATION

Across all task categories, the LLM-as-a-judge demonstrates near-perfect alignment with human majority judgments, with accuracy values effectively saturating at ≈ 1.0 and Cohen's kappa consistently matching or approaching unity. Importantly, this level of agreement is not restricted to any single task regime; it holds uniformly across quantitative reasoning, formal logic, and relational or planning-heavy tasks. When compared against human inter-annotator reliability, the LLM judge achieves parity with, and in some cases marginally exceeds, the empirical human agreement ceiling. This indicates that the judge is not merely replicating surface-level patterns but is closely tracking the underlying decision boundary implicitly shared by human annotators.

However, this apparent saturation also highlights a critical interpretability constraint: when both human agreement and model agreement approach near-perfect levels, Cohen's kappa becomes less sensitive to finer-grained systematic differences. In such regimes, small but meaningful divergences may be masked by the ceiling effect of binary evaluation. Therefore, while the results strongly support the claim that the LLM functions as a human-level evaluator in aggregate, the remaining gap between κ_{LLM} and κ_{human} across sub-regions (especially in more structured reasoning tasks) should be interpreted as indicative of subtle calibration differences rather than fundamental misalignment.

On LLM Judge Circularity. We note that GEMMA-4-26B-A4B-INSTRUCT appears both as an evaluated model and as the LLM judge, which raises a potential concern regarding circularity. In our setup, generation and evaluation are strictly decoupled: the judge model is never exposed to its own generation traces, logits, or intermediate reasoning states, and operates solely on the triplet (question, ground-truth answer, model response). The judging prompt enforces a deterministic, rubric-based comparison against the reference answer rather than open-ended reasoning, effectively reducing the task to structured answer matching. Empirically, the judge demonstrates near-perfect agreement with human annotators ($\kappa = 0.9995$), matching the human agreement ceiling, which indicates that its decisions closely track human consensus rather than model-specific biases. Overall, the combination of strict input isolation and human-level agreement provides strong evidence that the evaluation is not affected by circularity.

E.3. Human Baselines Development

To construct reliable human baselines, we evaluate two distinct proficiency levels: a *Human (Base)* group and a *Human (Expert)* group. The Human (Base) cohort consists of annotators with educational backgrounds ranging from high-school graduates to early undergraduate (first-year bachelor) students, representing non-specialist reasoning ability. The Human (Expert) cohort comprises individuals with advanced STEM training, including Master's and PhD-level students, representing domain-competent expert reasoning.

Each group consists of three annotators, and for every evaluation sample we collect independent responses from all three. Final answers for both Human (Base) and Human (Expert) are determined via majority voting, ensuring robustness against individual variation and annotation noise. This aggregated protocol is used to report human performance across all experiments, providing a stable reference for comparison against model behavior.

All annotators evaluated the full benchmark, including all invariant variants. To mitigate recognition bias, questions were randomly sampled and presented in a shuffled order, with the constraint that semantically equivalent variants of the same seed problem were separated by at least 50 intervening questions. Annotators were not provided with ground-truth answers or feedback at any stage.

We acknowledge that this within-subject design may introduce partial recognition effects, potentially inflating human invariance performance relative to a strict between-subjects protocol. However, enforced spacing and randomization substantially reduce short-term recall, and each instance still requires independent recomputation under varied formulations. Accordingly, the reported human results should be interpreted as an approximate upper bound on invariance under controlled, non-adversarial evaluation conditions.

F. Metric Definitions and Interpretations

We evaluate each model on a benchmark of N *complete samples*. Each sample $i \in \{1, \ldots, N\}$ consists of $K = 4$ question variants:

- $x_{i,\text{base}} \in \{0, 1\}$ — the original (base) question,
- $x_{i,1} \in \{0, 1\}$ — Invariant $\mathcal{I}_1$: Symbolic & Linguistic reformulation,
- $x_{i,2} \in \{0, 1\}$ — Invariant $\mathcal{I}_2$: Structural & Representational reformulation,
- $x_{i,3} \in \{0, 1\}$ — Invariant $\mathcal{I}_3$: Abstraction & Information-Structure reformulation,

where 1 denotes a correct model response and 0 denotes an incorrect one. All four variants are semantically equivalent and logically identical — only their surface representation differs. A robust model should answer all four variants consistently.

We define the *correctness count* for sample i as

$$c_i = x_{i,\text{base}} + x_{i,1} + x_{i,2} + x_{i,3} \ \in \ \{0, 1, 2, 3, 4\}. \tag{13}$$

Only samples for which all $K = 4$ variant scores are available are included in any metric computation. The set of invariant variants is denoted $\mathcal{I} = \{1, 2, 3\}$ and the full variant set $\mathcal{V} = \{\text{base}\} \cup \mathcal{I}$.

F.1. Baseline Accuracy Metrics

Accuracy per Variant. The per-variant accuracy is simply the mean correctness across all N samples:

$$\text{Acc}(v) = \frac{1}{N} \sum_{i=1}^{N} x_{i,v}, \qquad v \in \mathcal{V}. \tag{14}$$

The *mean invariant accuracy* averages the three invariant variants:

$$\overline{\mathcal{I}} = \frac{1}{3} \sum_{j \in \mathcal{I}} \text{Acc}(j). \tag{15}$$

The *invariance degradation gap* Δ measures how much performance drops from the base question to invariant reformulations:

$$\Delta = \text{Acc}(\text{base}) - \overline{\mathcal{I}}. \tag{16}$$

$\Delta > 0$ indicates that the model performs worse on reformulations than on the original question — the primary signal of *invariance fragility*. $\Delta \approx 0$ indicates robust, representation-agnostic reasoning. $\Delta < 0$ (rare) indicates that invariant forms are answered more reliably, potentially due to surface cues introduced by reformulation.

F.2. ICP: Invariance Consistency Profile & IMI

The **Invariance Consistency Profile (ICP)** characterises the full empirical distribution of per-sample correctness counts $c_i \in \{0, \ldots, 4\}$:

$$P(c = k) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}[c_i = k], \qquad k \in \{0, 1, 2, 3, 4\}. \tag{17}$$

This five-point distribution provides a richer picture than any single scalar: $P(c = 4)$ means the model got everything right; $P(c = 0)$ means the model failed all variants.

The **Invariant Mass Index (IMI)** is the mode of the distribution at complete correctness:

$$\text{IMI} = P(c_i = 4) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\left[\sum_{v \in \mathcal{V}} x_{i,v} = 4\right]. \tag{18}$$

Interpretation: Higher IMI is better. IMI is the strictest accuracy measure: it counts only samples where the model is *fully invariant* — correct on the base question *and* on all three reformulations. A model with high mean accuracy but low IMI is inconsistent: it gets some variants right by chance rather than through genuine understanding. IMI is equivalent to the ≥ 4 column in the at-least-k table (Section F.13).

F.3. CIR: Conditional Invariance Reliability

Let $\mathcal{B}^+ = \{i : x_{i,\text{base}} = 1\}$ be the set of samples where the model answers the base question correctly. **CIR** measures the fraction of those samples on which the model *also* answers every invariant reformulation correctly:

$$\text{CIR} = \frac{\left|\{i \in \mathcal{B}^+ : x_{i,1} = 1 \wedge x_{i,2} = 1 \wedge x_{i,3} = 1\}\right|}{|\mathcal{B}^+|}. \tag{19}$$

Interpretation: **Higher CIR is better.** CIR conditions on the model having the base knowledge and asks: given that the model *knows* the answer, does it remain correct when the question is re-expressed? A high CIR (≈ 1) means the model's knowledge is representation-stable. A low CIR means the model's correct responses on the base question are *not* robust — rephrasing collapses its performance even when the underlying reasoning should be sufficient.

F.4. ICR: Invariance Collapse Rate

A sample i is said to experience an *invariance collapse* if the model solves the base question but fails the majority of invariant variants. With $K-1=3$ invariant variants, "majority wrong" is defined as $\sum_{j\in\mathcal{I}} x_{i,j} \leq \lfloor (K-1)/2 \rfloor = 1$, i.e. at most one of the three invariants is answered correctly. ICR is:

$$\mathrm{ICR} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left[x_{i,\text{base}} = 1 \ \wedge \ \sum_{j\in\mathcal{I}} x_{i,j} \leq 1 \right]. \tag{20}$$

Interpretation: **Lower ICR is better.** ICR quantifies the most pathological failure mode: the model *knows* the answer (base correct) but catastrophically fails upon reformulation. This is strong evidence that the model has memorised surface-level patterns associated with the base phrasing rather than internalising the underlying concept.

F.5. PIDI: Pairwise Invariance Dependency Index

PIDI quantifies the conditional co-failure probability between pairs of invariant variants. For each ordered pair $(j,k) \in \mathcal{I}\times\mathcal{I}$ with $j \neq k$:

$$\mathrm{PIDI}(j,k) = P(x_{i,j}=0 \mid x_{i,k}=0) = \frac{|\{i : x_{i,j}=0 \wedge x_{i,k}=0\}|}{|\{i : x_{i,k}=0\}|}. \tag{21}$$

All six directed pairs are reported: $P(\mathcal{I}_1{=}0|\mathcal{I}_2{=}0)$, $P(\mathcal{I}_2{=}0|\mathcal{I}_1{=}0)$, and so on.

Interpretation: **Lower PIDI is better.** If a model fails variant k, PIDI measures how likely it is to also fail variant j. High PIDI across all pairs indicates that failures are *correlated* — the model tends to fail multiple invariant types simultaneously rather than independently. This suggests a systematic reasoning deficit rather than random noise. Values near 1 across all pairs indicate that the model *always* fails together (or succeeds together), pointing to a single shared failure mechanism.

F.6. TCIR: Type-Wise Conditional Invariance Retention

Let $\mathcal{S} = \{i : c_i > 0\}$ be the set of samples where the model answered *at least one* variant correctly — the "solvable" samples. **TCIR** measures the per-invariant mean accuracy conditioned on solvability:

$$\mathrm{TCIR}_j = \frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}} x_{i,j}, \qquad j \in \mathcal{I}. \tag{22}$$

Interpretation: **Higher TCIR is better.** By restricting to solvable samples, TCIR removes items the model cannot answer in any representation (perhaps genuinely hard items), and isolates whether the model *retains* its reasoning when only the surface form changes. A low TCIR_j relative to other invariants indicates that invariant type j is a specific vulnerability — the model can answer the question in some representations but consistently fails in the $\mathcal{I}_j$ representation.

F.7. TCC: Type Collapse Contribution

A *partial collapse* occurs at sample i when $c_i \in \{1,2\}$ — the model gets exactly one or two of the four variants right. Let $\mathcal{C} = \{i : c_i \in \{1,2\}\}$. **TCC** measures how much each invariant type *contributes* to these collapse events:

$$\mathrm{TCC}_j = P(x_{i,j}=0 \mid i \in \mathcal{C}) = \frac{|\{i \in \mathcal{C} : x_{i,j}=0\}|}{|\mathcal{C}|}, \qquad j \in \mathcal{I}. \tag{23}$$

Interpretation: **Lower TCC is better.** TCC identifies which invariant types are the *primary drivers* of partial collapses. A high TCC_j means that whenever the model is in a partial collapse (not fully correct and not fully wrong), it is variant j that tends to be wrong. This pinpoints the weakest representational form.

F.8. RIS: Relative Invariance Sensitivity

Let $\bar{x}$ be the global mean accuracy across all variants and all samples:

$$\bar{x} = \frac{1}{K \cdot N} \sum_{i=1}^{N} \sum_{v \in \mathcal{V}} x_{i,v}. \tag{24}$$

RIS measures each invariant's deviation from this global baseline:

$$\mathrm{RIS}_j = \mathrm{Acc}(j) - \bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_{i,j} - \bar{x}, \qquad j \in \mathcal{I}. \tag{25}$$

Interpretation: RIS is a *signed* measure in percentage points. **Positive** values ($\mathrm{RIS}_j > 0$) mean invariant j is *easier than average* for this model; **negative** values mean it is harder. Values near zero indicate that the model treats this invariant form as equally difficult as the average. RIS is useful for comparing across models: a model with strongly negative RIS_3 is particularly sensitive to the Abstraction & Info-Structure reformulation.

F.9. BSS: Bilateral Symmetry Score

Two types of *asymmetric* failure can occur:

$$F^+ = \left|\{i : x_{i,\text{base}} = 1 \ \wedge \ \exists\, j \in \mathcal{I} : \ x_{i,j} = 0\}\right|, \tag{26}$$

$$F^- = \left|\{i : x_{i,\text{base}} = 0 \ \wedge \ \exists\, j \in \mathcal{I} : \ x_{i,j} = 1\}\right|. \tag{27}$$

F^+ counts *forward failures*: the model answers the original correctly but fails at least one reformulation. F^- counts *backward failures*: the model fails the original but recovers on at least one reformulation. The **Bilateral Symmetry Score** is:

$$\mathrm{BSS} = 1 - \frac{|F^+ - F^-|}{F^+ + F^-} \ \in \ [0, 1], \tag{28}$$

with $\mathrm{BSS} = \texttt{null}$ when $F^+ = F^- = 0$ (no asymmetric failures).

Interpretation: **Higher BSS is better.** $\mathrm{BSS} = 1$ means failures are *perfectly symmetric*: the model fails the base exactly as often as it fails invariants, indicating consistent difficulty rather than directional bias. $\mathrm{BSS} \approx 0$ means nearly all failures are in one direction.

- $F^+ \gg F^-$ (low BSS, high forward rate): the model knows the answer in the base phrasing but cannot maintain correctness under reformulation — the canonical *invariance collapse*.
- $F^- \gg F^+$ (low BSS, high backward rate): the model fails the base but recovers on reformulations — a strong signal of *surface-pattern exploitation*, where the reformulated phrasing inadvertently provides exploitable cues.

$\mathrm{BSS} < 0.4$ is flagged as highly asymmetric; $\mathrm{BSS} \in [0.4, 0.7)$ is moderate; $\mathrm{BSS} \geq 0.7$ is considered symmetric.

F.10. CE: Consistency Entropy

For each sample i, define the per-sample correct fraction $\hat{p}_i = c_i/K$. The *binary entropy* of this fraction is:

$$H(\hat{p}_i) = -\hat{p}_i \log_2 \hat{p}_i \ - \ (1 - \hat{p}_i) \log_2 (1 - \hat{p}_i), \tag{29}$$

with $H(0) = H(1) = 0$ by convention. H reaches its maximum of 1 at $\hat{p}_i = 0.5$ ($c_i = 2$: exactly half the variants correct). **CE** averages this per-sample entropy:

$$\mathrm{CE} = \frac{1}{N} \sum_{i=1}^{N} H\left(\frac{c_i}{K}\right). \tag{30}$$

Interpretation: **Lower CE is better.** CE measures how *inconsistent* a model is at the per-sample level. A sample with $c_i \in \{0, 4\}$ contributes $H = 0$ — the model is fully consistent (either always right or always wrong). A sample with $c_i = 2$ contributes the maximum entropy $H = 1$.

- **Low CE** (< 0.25): the model's responses are structured — it consistently succeeds or consistently fails, suggesting genuine reasoning or genuine ignorance rather than noise.
- **High CE** (≥ 0.55): flagged as heuristic or random-looking, indicating the model frequently gets *some but not all* variants right in a way that appears stochastic rather than systematic.

F.11. IRR: Invariant Recovery Rate

Let $\mathcal{B}^- = \{i : x_{i,\text{base}} = 0\}$ be the set of samples where the model *fails* the base question. IRR conditions on these base-failed samples and asks whether invariant reformulations can *recover*:

$$\text{IRR}_{\text{full}} = P\big(\forall j \in \mathcal{I} : \, x_{i,j} = 1 \;\big|\; i \in \mathcal{B}^-\big)\,, \tag{31}$$

$$beginequation4pt]\text{IRR}_{\text{part}} = P\big(\exists j \in \mathcal{I} : \, x_{i,j} = 1 \;\big|\; i \in \mathcal{B}^-\big)\,, \tag{32}$$

$$beginequation4pt]\text{IRR}_{\text{zero}} = 1 - \text{IRR}_{\text{part}} = P\big(\forall j \in \mathcal{I} : \, x_{i,j} = 0 \;\big|\; i \in \mathcal{B}^-\big)\,. \tag{33}$$

The signed **IRR-delta** compares recovery from base failure against retention from base success (CIR):

$$\text{IRR}_{\Delta} = \text{IRR}_{\text{full}} - (1 - \text{CIR}). \tag{34}$$

Interpretation: A high IRR_{full} means that when the model fails the original question, all three reformulations still elicit correct responses — a strong indicator that the *base question itself* is harder than its invariant forms (possible due to base phrasing complexity or training exposure differences). $\text{IRR}_{\Delta} > 0$ means the model finds the base question harder than its invariants; $\text{IRR}_{\Delta} < 0$ means the invariants are harder. High IRR_{part} with low IRR_{full} indicates that the model can occasionally exploit an invariant phrasing but cannot generalise across all reformulations.

F.12. FG: Fragility Gradient

The three invariant types are designed to increase in reformulation difficulty: $\mathcal{I}_1$ (Symbolic & Linguistic) is the most surface-level change; $\mathcal{I}_3$ (Abstraction & Information-Structure) requires the deepest representational re-encoding. A model with genuine understanding should respect this ordering.

Define the *gradient steps*:

$$g_{12} = \text{Acc}(\mathcal{I}_1) - \text{Acc}(\mathcal{I}_2), \qquad g_{23} = \text{Acc}(\mathcal{I}_2) - \text{Acc}(\mathcal{I}_3). \tag{35}$$

The expected direction is $g_{12} \geq 0$ and $g_{23} \geq 0$.

$$\text{FG}_{\text{score}} = \mathbf{1}[g_{12} \geq 0] + \mathbf{1}[g_{23} \geq 0] \;\in\; \{0, 1, 2\}, \tag{36}$$

$$beginequation4pt]\text{FG}_{\text{mag}} = \frac{g_{12} + g_{23}}{2} \quad \text{(signed mean gradient in pp)}, \tag{37}$$

$$beginequation4pt]\text{FG}_{\text{inv}} = \mathbf{1}[\text{FG}_{\text{score}} = 0]. \tag{38}$$

Interpretation: **Higher FG_{score} is better** (maximum 2). $\text{FG}_{\text{score}} = 2$ indicates the model respects the expected difficulty ordering: it performs best on the easiest invariant type and progressively worse on harder types. $\text{FG}_{\text{inv}} = \text{True}$ (both gradients inverted, score $= 0$) is a red flag: the model performs *better* on harder reformulations. This is inconsistent with genuine understanding and is attributed to *surface-pattern exploitation* — the harder reformulations may inadvertently contain surface cues (e.g. unusual phrasing, more explicit logical structure) that the model latches on to.

A negative FG_{mag} of large magnitude confirms the gradient is inverted, not merely flat.

F.13. At-Least-k Correct Analysis

Let $\mathcal{A}_k = \{i : c_i \geq k\}$ for $k \in \{1, 2, 3, 4\}$. The **at-least-k-correct** proportion is:

$$P(\geq k) = \frac{|\mathcal{A}_k|}{N} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}[c_i \geq k], \qquad k = 1, 2, 3, 4. \tag{39}$$

These four quantities form a nested, monotonically decreasing sequence:

$$P(\geq 1) \;\geq\; P(\geq 2) \;\geq\; P(\geq 3) \;\geq\; P(\geq 4) = \mathrm{IMI}. \tag{40}$$

The complementary *none-correct* proportion is:

$$P(=0) = 1 - P(\geq 1) = \frac{|\{i : c_i = 0\}|}{N}. \tag{41}$$

Relationship to ICP. The at-least-k statistics are the complementary cumulative distribution function (CCDF) of the correctness count distribution defined in the ICP (Section F.2):

$$P(\geq k) = \sum_{m=k}^{4} P(c = m). \tag{42}$$

Interpretation: All four at-least-k values are **higher is better**; the none-correct proportion is **lower is better.**

- $P(\geq 1)$ — the model answered at least one of the four formulations correctly. This is the loosest possible measure of task-relevant knowledge and has a high ceiling even for weak models.
- $P(\geq 2)$ — at least two formulations correct. A threshold of $P(\geq 2) \approx P(\geq 1)$ (near-flat drop) indicates that the model is consistent once it gets any variant right.
- $P(\geq 3)$ — at least three of the four formulations correct. A large gap between $P(\geq 2)$ and $P(\geq 3)$ signals that the model frequently knows the answer in *some* surface forms but not others.
- $P(\geq 4)$ **= IMI** — the model is fully invariant. The gap $P(\geq 3) - P(\geq 4)$ reflects how often the model passes three of four variants but fails exactly one invariant type.
- $P(=0)$ — the fraction of questions the model cannot answer in any representation. A high $P(=0)$ reflects genuine knowledge gaps, not reformulation sensitivity.

Illustrative degradation profile. A model with $P(\geq 1) = 0.92$, $P(\geq 2) = 0.75$, $P(\geq 3) = 0.55$, $P(\geq 4) = 0.41$ shows a steep decline: the 0.17 drop from $P(\geq 1)$ to $P(\geq 2)$, the 0.20 drop to $P(\geq 3)$, and the 0.14 drop to $P(\geq 4)$ indicate that each additional reformulation constraint eliminates a sizable fraction of otherwise-correct samples — a hallmark of surface-level reasoning.

Table 8 provides a consolidated reference for all metrics, their directionality, the question they answer, and their failure interpretation.

F.14. Why these metrics are contamination-aware?

Let s_0 denote the base instance and $s_k = \tau_k(s_0)$ its invariant variants, where τ_k is semantics-preserving (Eq. equation 1). A model that has memorised s_0 can achieve high $\mathrm{Acc}(\text{base})$ without exhibiting invariant reasoning. However, since $s_k \neq s_0$ at the string level, success on s_k requires recognising the equivalence class $[s_0]_{[\![]}$ rather than recalling a specific form. Consequently, metrics that depend on joint correctness across variants—such as Δ, IMI, and CIR—cannot be inflated by memorisation of s_0 alone. In particular, the gap Δ and consistency-based measures provide a lower bound on representation-invariant competence: any improvement on these metrics necessarily implies the ability to generalise across semantics-preserving transformations, rather than reliance on surface-form recall.

G. Additional Experiments and Results

G.1. Correctness Profile

Table 9 reveals a strong and systematic invariance collapse pattern across all model families, where performance drops sharply from solving at least one variant to maintaining correctness across all four, indicating that partial success does not

Table 8. Summary of all invariance evaluation metrics. ↑ = higher is better; ↓ = lower is better.

ID	Metric	Dir.	Question answered	Low value means…
Δ	Invariance Gap	↓	How much does accuracy drop under reformulation?	Small/no degradation
IMI	Invariant Mass Index	↑	What fraction of samples are fully invariant?	Model rarely stays correct across all forms
CIR	Cond. Invariance Reliability	↑	Given base correct, are reformulations also correct?	Correct base knowledge does not transfer
ICR	Invariance Collapse Rate	↓	How often does knowing the base lead to collapse?	Rare catastrophic failure mode
PIDI	Pairwise Inv. Dependency	↓	Are failures correlated across invariant types?	Independent failures (less systematic)
TCIR	Type-Wise CIR	↑	Which invariant type is the weakest, for solvable items?	Specific invariant type is a vulnerability
TCC	Type Collapse Contrib.	↓	Which invariant type drives partial collapses?	That invariant type is rarely the weak link
RIS	Relative Inv. Sensitivity	signed	Is this invariant easier or harder than average?	Harder than the model's global mean
BSS	Bilateral Symmetry Score	↑	Are failures direction-symmetric across base↔inv?	Asymmetric — potential pattern exploitation
CE	Consistency Entropy	↓	How random are per-sample correctness patterns?	Structured, consistent model behaviour
$\mathrm{IRR}_{\mathrm{full}}$	Invariant Recovery Rate	↑	Can invariants recover when base fails?	Model cannot recover through reformulation
IRR_{Δ}	IRR Delta	signed	Is the base harder or easier than invariants?	Invariants are harder than the base
FG	Fragility Gradient	↑	Does performance follow the expected difficulty order?	Gradient violated — possible surface cues
$P(\geq k)$	At-Least-k Correct	↑	How broadly correct is the model across variants?	Model rarely sustains correctness

translate into stable cross-representation reasoning. Small open-source models exhibit extreme fragility, with ≥ 4 correctness often below 2–3% and a large failure mass concentrated in the "none correct" bucket, suggesting near-complete dependence on surface-form matching rather than invariant structure. Larger open-source models partially mitigate this trend, with QWEN3-8B showing notably stronger intermediate consistency (≥ 2 and ≥ 3), but still suffering a steep decay toward full invariance. Among ¿30B models, gains are more pronounced, particularly for QWEN3.6-PLUS and GPT-5.2, which achieve substantially higher ≥ 3 and ≥ 4 rates, indicating emerging but incomplete invariance robustness.

However, even these strongest models remain far below human performance, where experts achieve near-perfect retention at all levels, highlighting a fundamental gap in stability rather than raw accuracy. The human baseline shows an almost flat curve across all k, with expert performance maintaining 98.6% at ≥ 4, demonstrating near-ideal invariance under semantic-preserving transformations. A consistent pattern across all model scales is that improvements in ≥ 1 performance do not proportionally translate into ≥ 4 performance, indicating that correctness is not compositionally stable. The "none correct" column further emphasizes this instability, with smaller models failing entirely on over 60–80% of samples, while even frontier models still fail completely on a non-trivial fraction. Overall, the results suggest that scaling improves partial problem-solving ability but does not guarantee cross-variant coherence, which appears to require a qualitatively different form of representation stability.

Table 9. **Proportion of benchmark samples (%) on which each model answers at least k out of 4 variants correctly (Base $+ \mathcal{I}_1 + \mathcal{I}_2 + \mathcal{I}_3$).** ≥ 4 is equivalent to the Invariant Mass Index (IMI).

	At Least k Correct out of Total 4 (Base + 3 Invariants) (%) (↑)				*None (%) (↓)*
Model	≥ 1	≥ 2	≥ 3	≥ 4	0
Open-Source Models (Small) $\leq$30B					
Llama-3.2-1B-I	15.6	7.0	3.9	1.7	84.4
Ministral-3B-2512	24.8	12.0	5.0	1.2	75.2
Gemma-3-4B-I	24.2	9.7	5.0	1.9	75.8
Qwen3-8B	76.3	66.0	52.8	32.2	23.7
Gemma-4-26B-A4B-I	48.0	32.8	20.7	7.2	52.0
Open-Source Models (Large) ¿30B					
Olmo-3.1-32B-I	32.8	17.8	8.8	3.0	67.2
Aya-Expanse-32B	31.5	16.3	8.3	2.7	68.5
Llama-3.3-70B-I	49.7	29.0	16.2	5.5	50.3
Qwen3.6-Plus	80.0	73.0	65.3	**48.6**	20.0
DeepSeek-V3.1	38.3	21.3	9.8	2.2	61.7
Closed-Source Models					
Command-A	42.6	23.3	13.5	4.3	57.4
o4-mini	35.7	20.2	10.7	3.7	64.3
GPT-5.2	**81.8**	**75.9**	**68.8**	48.5	**18.2**
Gemini-3-Flash	60.3	45.8	33.3	17.0	39.7
Claude Haiku 4.5	77.5	59.2	40.7	17.4	22.5
Human Baseline					
Base	**100.0**	99.5	95.2	65.3	**0.0**
Expert	100.0	**100.0**	**99.9**	**98.6**	**0.0**

G.2. On Negative Invariance Gaps

In Table 1, a subset of models exhibit non-positive average invariance gaps ($\Delta \leq 0$), indicating that performance on invariant variants is, on average, equal to or slightly higher than the base formulation. While this may appear to contradict the broader claim of invariance degradation, such aggregate behavior does not necessarily reflect improved reasoning. In practice, invariance transformations are heterogeneous: some operations (e.g., distractor removal or simplification) can make instances easier, while others (e.g., abstraction shifts or distractor injection) increase difficulty. As a result, per-sample gains and losses can cancel out when averaged, producing a near-zero or negative Δ despite underlying instability.
Crucially, this phenomenon often reflects inconsistency rather than robustness. A model may fail on the base form of one problem but succeed on its invariant variant, while exhibiting the opposite behavior on another instance. These offsetting errors lead to superficially stable averages without indicating true invariance. Therefore, negative or near-zero gaps should not be interpreted as evidence of invariant reasoning, but rather as a symptom of unstructured sensitivity to representation. To address this limitation, we introduce complementary metrics that explicitly capture consistency across variants, such as the Invariant Mass Index (Table 2) and at-least-k correctness profiles (Table 9). These measures evaluate whether a model can simultaneously maintain correctness across multiple equivalent forms of the same problem, rather than averaging

performance across unrelated instances. As shown in our results, models with near-zero Δ often still exhibit low consistency scores, confirming that average accuracy alone is insufficient to characterize invariance behavior.

G.3. On Semantics-Preservation and Perturbation Axes

A potential concern is that not all transformation axes are strictly difficulty-neutral, despite satisfying the formal semantics-preservation contract. In particular, transformations such as distractor injection/removal and abstraction-level shifts can systematically alter problem difficulty while preserving the underlying answer. This raises the distinction between *pure invariance* (difficulty-neutral re-expression) and *controlled perturbation* (difficulty-modulating but semantics-preserving transformations).

In our analysis, we explicitly account for this distinction by grouping transformation axes into (i) pure invariance transformations—such as variable renaming, entity relabeling, premise reordering, logical/mathematical reformulation, encoding swaps, and paraphrastic variation—and (ii) controlled perturbations, including distractor injection/removal and abstraction-level shifts. We report performance both at the aggregate level and through fine-grained, axis-wise breakdowns, enabling direct comparison across these groups. This separation ensures that conclusions about invariance robustness are not conflated with effects driven by systematic difficulty changes.

Importantly, even when restricting evaluation to pure invariance axes, we observe consistent degradation in model performance, confirming that the invariance gap is not an artifact of perturbation-induced difficulty variation. In fact, isolating pure invariance transformations strengthens the central claim: models fail to maintain correctness even when only surface form is altered without any intended change in difficulty. Controlled perturbation axes are retained not as strict invariance tests, but as complementary probes that reveal directional sensitivities and robustness limits under realistic input variation. Together, this unified framework captures both strict invariance behavior and practical robustness under semantically equivalent but cognitively diverse formulations.

G.4. Fragility Gradient and Non-Monotonic Difficulty

Table 3 shows that the Fragility Gradient (FG) rarely attains the ideal monotonic ordering (score = 2), with most models clustering at FG = 1 and even GPT-5.2 at FG = 0, while Human (Expert) also does not achieve perfect monotonicity. This outcome does not contradict the framework but rather highlights an important empirical reality: the assumed ordering $\mathcal{I}_1 < \mathcal{I}_2 < \mathcal{I}_3$ is a coarse, aggregate heuristic rather than a strict per-instance property. In practice, different invariance axes interact with task structure in non-uniform ways, and the relative difficulty of variants can fluctuate depending on the underlying reasoning type. For example, a structural reformulation may be harder than a symbolic renaming in one task but easier in another, breaking global monotonicity even for humans.

Therefore, FG should be interpreted as a diagnostic signal of stability rather than a strict ranking validator. A score of 1 indicates partial ordering consistency, while deviations (FG = 0) reflect heterogeneous sensitivity across invariance axes rather than failure of the metric itself. The fact that even human annotators do not consistently achieve FG = 2 further reinforces that invariance difficulty is not inherently linear, but context-dependent. Retaining FG is still informative, as it captures whether models exhibit coherent degradation trends, but we explicitly avoid over-interpreting it as evidence of a universal difficulty hierarchy.

H. Full Per-Task and Per-Sub-Type Results

H.1. A. Formal Logic and Discrete Structure

Table 10. Core accuracy — Task Type: *A. Formal Logic and Discrete Structure*.

Model	Base	*Accuracy (%)* $\mathcal{I}_1$	$\mathcal{I}_2$	$\mathcal{I}_3$	*Mean Inv.* $\overline{\mathcal{I}}$	*Base*$-\overline{\mathcal{I}}$ Δ
Llama-3.2-1B-I	4.83	4.07	6.11	6.11	5.43	-0.60
Ministral-3B-2512	8.65	9.67	11.20	11.20	10.69	**-2.04**
Gemma-3-4B-I	9.92	10.18	9.67	7.89	9.25	+0.67
Qwen3-8B	48.60	49.62	50.38	48.85	49.62	-1.02
Gemma-4-26B-A4B-I	37.40	32.06	33.59	30.28	31.98	+5.42
Olmo-3.1-32B-I	21.88	16.54	22.39	12.47	17.13	+4.75
Aya-Expanse-32B	15.01	9.92	16.28	15.78	13.99	+1.02
Llama-3.3-70B-I	26.21	21.37	25.19	21.37	22.64	+3.57
Qwen3.6-Plus	64.89	59.03	59.03	57.76	58.61	+6.28
DeepSeek-V3.1	22.39	16.79	23.41	17.81	19.34	+3.05
Command-A	24.94	23.92	23.92	25.19	24.34	+0.60
o4-mini	24.17	15.01	21.12	14.50	16.88	+7.29
GPT-5.2	63.10	61.32	61.07	62.34	61.58	+1.52
Gemini-3-Flash	46.06	40.71	41.98	43.51	42.07	+3.99
Claude Haiku 4.5	50.89	47.58	46.06	49.11	47.58	+3.31
Human (Base)	90.08	90.59	90.59	87.53	89.57	+0.51
Human (Expert)	**99.75**	**100.00**	**98.47**	**99.75**	**99.41**	+0.34

Table 11. Invariance stability — Task Type: *A. Formal Logic and Discrete Structure*.

Model	*Global Stability (%)* IMI↑	CIR↑	ICR↓	*Novel (%)* BSS↑	CE↓	*Pairwise Failure Dep. (%)* $P(\mathcal{I}_1{=}0\|\mathcal{I}_2{=}0)$	$P(\mathcal{I}_1{=}0\|\mathcal{I}_3{=}0)$	$P(\mathcal{I}_2{=}0\|\mathcal{I}_3{=}0)$
Llama-3.2-1B-I	1.02	21.05	2.04	65.22	9.83	97.29	98.10	95.66
Ministral-3B-2512	0.76	8.82	5.34	64.58	21.07	91.69	93.98	91.12
Gemma-3-4B-I	0.76	7.69	6.62	75.00	20.78	91.55	91.44	91.44
Qwen3-8B	20.10	41.36	10.43	93.84	47.06	69.74	73.63	62.69
Gemma-4-26B-A4B-I	9.92	26.53	12.47	81.97	40.32	81.61	78.83	77.01
Olmo-3.1-32B-I	3.05	13.95	13.23	94.29	31.49	90.49	86.92	82.56
Aya-Expanse-32B	1.53	10.17	9.16	79.70	29.04	90.88	93.66	87.61
Llama-3.3-70B-I	4.83	18.45	13.23	**100.00**	36.99	86.05	86.41	81.23
Qwen3.6-Plus	36.90	56.86	8.65	64.20	35.46	72.67	69.28	64.46
DeepSeek-V3.1	2.04	9.09	13.23	96.97	36.70	88.04	87.31	82.66
Command-A	5.09	20.41	9.92	95.12	36.30	83.28	83.33	84.69
o4-mini	4.33	17.89	13.74	87.77	30.47	91.29	89.88	83.93
GPT-5.2	37.15	58.87	6.36	80.00	37.01	65.36	69.59	68.24
Gemini-3-Flash	16.28	35.36	12.98	82.41	44.15	77.19	74.32	75.68
Claude Haiku 4.5	16.54	32.50	12.47	84.12	51.36	66.98	67.50	67.50
Human (Base)	64.63	71.75	2.54	56.12	29.61	8.11	6.12	12.24
Human (Expert)	**98.22**	**98.47**	**0.00**	28.57	**1.49**	0.00	0.00	0.00

Table 12. Novel metrics — Task Type: *A. Formal Logic and Discrete Structure*.

Model	*Bilateral Symmetry* BSS↑	F^+(%)	F^-(%)	*Entropy* CE↓	*Invariant Recovery (%)* IRR_{full} ↑	IRR_{part} ↑	IRR_{Δ}	*Fragility Gradient* FG	$\overline{g}$
Llama-3.2-1B-I	65.22	3.82	7.89	9.83	0.00	8.29	-78.95	1	-1.02
Ministral-3B-2512	64.58	7.89	16.54	21.07	0.84	18.11	-90.34	1	-0.76
Gemma-3-4B-I	75.00	9.16	15.27	20.78	0.00	16.95	-92.31	**2**	+1.15
Qwen3-8B	93.84	28.50	25.19	47.06	8.42	49.01	-50.22	1	+0.38
Gemma-4-26B-A4B-I	81.97	27.48	19.08	40.32	2.85	30.49	-70.62	1	+0.89
Olmo-3.1-32B-I	94.29	18.83	16.79	31.49	0.33	21.50	-85.72	1	+2.03
Aya-Expanse-32B	79.70	13.49	20.36	29.04	0.00	23.95	-89.83	1	-2.93
Llama-3.3-70B-I	**100.00**	21.37	21.37	36.99	2.41	28.97	-79.14	1	+0.00
Qwen3.6-Plus	64.20	27.99	13.23	35.46	3.62	37.68	-39.52	2	+0.63
DeepSeek-V3.1	96.97	20.36	21.63	36.70	1.31	27.87	-89.60	1	-0.51
Command-A	95.12	19.85	21.88	36.30	2.03	29.15	-77.56	1	-0.63
o4-mini	87.77	19.85	15.52	30.47	0.34	20.47	-81.77	1	+0.25
GPT-5.2	80.00	25.95	17.30	37.01	8.97	46.90	-32.16	1	-0.51
Gemini-3-Flash	82.41	29.77	20.87	44.15	6.13	38.68	-58.51	0	-1.40
Claude Haiku 4.5	84.12	34.35	24.94	51.36	7.25	50.78	-60.25	1	-0.76
Human (Base)	56.12	25.45	9.92	29.61	**71.79**	**100.00**	+43.54	2	+1.53
Human (Expert)	28.57	1.53	0.25	**1.49**	0.00	100.00	-1.53	1	+0.12

Table 13. Sensitivity profile — Task Type: *A. Formal Logic and Discrete Structure*.

Model	*Cond. Inv. Retention (%)*			*Collapse Contribution (%)*			*Rel. Sensitivity (pp)*		
	$TCIR_1$	$TCIR_2$	$TCIR_3$	$TCC_1 \downarrow$	$TCC_2 \downarrow$	$TCC_3 \downarrow$	RIS_1	RIS_2	RIS_3
Llama-3.2-1B-I	32.00	48.00	48.00	82.05	53.85	66.67	-1.21	+0.83	+0.83
Ministral-3B-2512	38.38	44.44	44.44	68.67	61.45	63.86	-0.51	+1.02	+1.02
Gemma-3-4B-I	40.40	38.38	31.31	65.12	66.28	75.58	+0.77	+0.26	-1.52
Qwen3-8B	67.24	68.28	66.21	63.41	53.66	56.91	+0.26	+1.02	-0.51
Gemma-4-26B-A4B-I	56.76	59.46	53.60	67.52	64.10	64.96	-1.27	+0.26	-3.05
Olmo-3.1-32B-I	42.76	57.89	32.24	65.81	52.99	79.49	-1.78	+4.07	-5.85
Aya-Expanse-32B	28.06	46.04	44.60	76.72	61.21	64.66	-4.33	+2.03	+1.53
Llama-3.3-70B-I	44.92	52.94	44.92	72.09	61.24	69.77	-2.17	+1.65	-2.17
Qwen3.6-Plus	75.57	75.57	73.94	71.60	59.26	59.26	-1.15	-1.15	-2.42
DeepSeek-V3.1	38.15	53.18	40.46	69.17	57.14	71.43	-3.31	+3.31	-2.29
Command-A	51.09	51.09	53.80	62.18	64.71	63.03	-0.57	-0.57	+0.70
o4-mini	37.82	53.21	36.54	76.32	60.53	78.07	-3.69	+2.42	-4.20
GPT-5.2	76.27	75.95	77.53	61.25	58.75	61.25	-0.64	-0.89	+0.38
Gemini-3-Flash	60.84	62.74	65.02	63.33	65.00	60.83	-2.36	-1.09	+0.44
Claude Haiku 4.5	62.75	60.74	64.77	62.41	63.91	59.40	-0.83	-2.35	+0.70
Human (Base)	90.59	90.59	87.53	38.10	**47.62**	71.43	+0.90	+0.90	-2.16
Human (Expert)	**100.00**	**98.47**	**99.75**	**0.00**	100.00	**0.00**	+0.51	-1.02	+0.26

Table 14. Sub-type accuracy breakdown for $\mathcal{I}_1$ — Task Type: *A. Formal Logic and Discrete Structure*.

Model	*Accuracy (%) by Sub-Type — $\mathcal{I}_1$ (Symbolic & Linguistic)*			
	Entity Relabeling	Style / Register Invariance	Synonym / Semantic Paraphrase...	Variable Renaming
Llama-3.2-1B-I	4.30	4.94	3.53	3.73
Ministral-3B-2512	13.98	6.17	16.47	4.48
Gemma-3-4B-I	11.83	8.64	17.65	5.22
Qwen3-8B	52.69	48.15	62.35	40.30
Gemma-4-26B-A4B-I	32.26	22.22	42.35	31.34
Olmo-3.1-32B-I	18.28	3.70	27.06	16.42
Aya-Expanse-32B	10.75	7.41	12.94	8.96
Llama-3.3-70B-I	16.13	22.22	41.18	11.94
Qwen3.6-Plus	63.44	56.79	69.41	50.75
DeepSeek-V3.1	19.35	9.88	27.06	12.69
Command-A	23.66	19.75	35.29	19.40
o4-mini	17.20	9.88	23.53	11.19
GPT-5.2	63.44	66.67	72.94	49.25
Gemini-3-Flash	37.63	39.51	51.76	36.57
Claude Haiku 4.5	44.09	54.32	51.76	43.28
Human (Base)	90.32	92.59	90.59	89.55
Human (Expert)	**100.00**	**100.00**	**100.00**	**100.00**

Table 15. Sub-type accuracy breakdown for $\mathcal{I}_2$ — Task Type: *A. Formal Logic and Discrete Structure*.

Model	*Accuracy (%) by Sub-Type — $\mathcal{I}_2$ (Structural & Representational)*			
	Equivalent Logical Reformulation	Equivalent Mathematical Reformulation	Premise Reordering	Representation / Encoding Swap
Llama-3.2-1B-I	7.63	5.19	6.31	4.05
Ministral-3B-2512	15.27	18.18	6.31	4.05
Gemma-3-4B-I	18.32	10.39	2.70	4.05
Qwen3-8B	65.65	37.66	44.14	45.95
Gemma-4-26B-A4B-I	54.96	12.99	30.63	21.62
Olmo-3.1-32B-I	33.59	18.18	17.12	14.86
Aya-Expanse-32B	27.48	12.99	9.91	9.46
Llama-3.3-70B-I	39.69	16.88	20.72	14.86
Qwen3.6-Plus	67.94	46.75	61.26	52.70
DeepSeek-V3.1	37.40	15.58	18.92	13.51
Command-A	35.11	16.88	21.62	14.86
o4-mini	29.01	12.99	18.02	20.27
GPT-5.2	69.47	44.16	61.26	63.51
Gemini-3-Flash	64.12	29.87	34.23	27.03
Claude Haiku 4.5	50.38	41.56	47.75	40.54
Human (Base)	90.84	89.61	90.99	90.54
Human (Expert)	**99.24**	**100.00**	**96.40**	**98.65**

Table 16. Sub-type accuracy breakdown for $\mathcal{I}_3$ — Task Type: *A. Formal Logic and Discrete Structure*.

Model	*Accuracy (%) by Sub-Type — $\mathcal{I}_3$ (Abstraction & Info-Structure)*			
	Abstraction Level Shift	Distractor Injection / Removal	Perspective / Role Reframing	Redundant Restatement Invariance
Llama-3.2-1B-I	8.94	5.15	6.73	1.45
Ministral-3B-2512	8.94	16.49	8.65	11.59
Gemma-3-4B-I	7.32	12.37	5.77	5.80
Qwen3-8B	42.28	46.39	52.88	57.97
Gemma-4-26B-A4B-I	30.08	37.11	19.23	37.68
Olmo-3.1-32B-I	13.01	13.40	10.58	13.04
Aya-Expanse-32B	13.82	17.53	16.35	15.94
Llama-3.3-70B-I	19.51	23.71	20.19	23.19
Qwen3.6-Plus	43.90	67.01	54.81	73.91
DeepSeek-V3.1	17.07	22.68	11.54	21.74
Command-A	22.76	36.08	18.27	24.64
o4-mini	15.45	18.56	10.58	13.04
GPT-5.2	56.10	61.86	60.58	76.81
Gemini-3-Flash	39.02	55.67	34.62	47.83
Claude Haiku 4.5	43.90	57.73	40.38	59.42
Human (Base)	85.37	88.66	88.46	88.41
Human (Expert)	**99.19**	**100.00**	**100.00**	**100.00**

H.1.1. A1. MULTI-QUANTIFIER FIRST-ORDER LOGIC

Table 17. Core accuracy — Sub-Type: *A1. Multi-Quantifier First-Order Logic*.

Model	Base	*Accuracy (%)* $\mathcal{I}_1$	$\mathcal{I}_2$	$\mathcal{I}_3$	*Mean Inv.* $\overline{\mathcal{I}}$	*Base*$-\overline{\mathcal{I}}$ Δ
Llama-3.2-1B-I	3.30	3.30	9.89	6.59	6.59	-3.29
Ministral-3B-2512	5.49	9.89	16.48	9.89	12.09	-6.60
Gemma-3-4B-I	13.19	13.19	15.38	7.69	12.09	+1.10
Qwen3-8B	28.57	37.36	54.95	26.37	39.56	**-10.99**
Gemma-4-26B-A4B-I	70.33	54.95	67.03	46.15	56.04	+14.29
Olmo-3.1-32B-I	34.07	25.27	42.86	14.29	27.47	+6.60
Aya-Expanse-32B	25.27	8.79	36.26	18.68	21.24	+4.03
Llama-3.3-70B-I	40.66	30.77	48.35	29.67	36.26	+4.40
Qwen3.6-Plus	48.35	48.35	57.14	35.16	46.88	+1.47
DeepSeek-V3.1	23.08	20.88	40.66	16.48	26.01	-2.93
Command-A	32.97	37.36	43.96	35.16	38.83	-5.86
o4-mini	42.86	24.18	39.56	13.19	25.64	+17.22
GPT-5.2	50.55	62.64	60.44	48.35	57.14	-6.59
Gemini-3-Flash	51.65	49.45	54.95	50.55	51.65	+0.00
Claude Haiku 4.5	51.65	49.45	59.34	53.85	54.21	-2.56
Human (Base)	95.60	86.81	91.21	90.11	89.38	+6.22
Human (Expert)	**100.00**	**100.00**	**97.80**	**98.90**	**98.90**	+1.10

Table 18. Invariance stability — Sub-Type: *A1. Multi-Quantifier First-Order Logic*.

Model	*Global Stability (%)* IMI↑	CIR↑	ICR↓	*Novel (%)* BSS↑	CE↓	*Pairwise Failure Dep. (%)* $P(\mathcal{I}_1{=}0\|\mathcal{I}_2{=}0)$	$P(\mathcal{I}_1{=}0\|\mathcal{I}_3{=}0)$	$P(\mathcal{I}_2{=}0\|\mathcal{I}_3{=}0)$
Llama-3.2-1B-I	1.10	33.33	**0.00**	33.33	10.91	98.78	97.65	92.94
Ministral-3B-2512	1.10	20.00	3.30	30.77	24.01	93.42	92.68	85.37
Gemma-3-4B-I	1.10	8.33	9.89	73.33	28.20	90.91	88.10	84.52
Qwen3-8B	9.89	34.62	10.99	59.65	56.00	80.49	68.66	47.76
Gemma-4-26B-A4B-I	17.58	25.00	23.08	58.82	65.81	66.67	48.98	30.61
Olmo-3.1-32B-I	4.40	12.90	16.48	94.12	48.58	84.62	78.21	61.54
Aya-Expanse-32B	4.40	17.39	17.58	80.85	44.39	93.10	95.95	68.92
Llama-3.3-70B-I	9.89	24.32	18.68	**100.00**	53.66	85.11	78.12	56.25
Qwen3.6-Plus	13.19	27.27	18.68	100.00	62.45	71.79	54.24	42.37
DeepSeek-V3.1	2.20	9.52	15.38	70.37	51.87	81.48	82.89	63.16
Command-A	7.69	23.33	13.19	76.67	58.47	70.59	66.10	57.63
o4-mini	5.49	12.82	26.37	76.36	51.94	85.45	79.75	64.56
GPT-5.2	16.48	32.61	10.99	88.57	67.38	41.67	38.30	42.55
Gemini-3-Flash	10.99	21.28	19.78	94.29	68.21	68.29	48.89	46.67
Claude Haiku 4.5	13.19	25.53	12.09	100.00	66.35	56.76	57.14	40.48
Human (Base)	68.13	71.26	1.10	27.59	26.27	12.50	0.00	11.11
Human (Expert)	**96.70**	**96.70**	0.00	0.00	**2.67**	0.00	0.00	0.00

Table 19. Novel metrics — Sub-Type: *A1. Multi-Quantifier First-Order Logic*.

Model	*Bilateral Symmetry* BSS↑	$F^+(\%)$	$F^-(\%)$	*Entropy* CE↓	*Invariant Recovery (%)* IRR_{full} ↑	IRR_{part} ↑	IRR_{Δ}	*Fragility Gradient* FG	$\overline{g}$
Llama-3.2-1B-I	33.33	2.20	10.99	10.91	0.00	11.36	-66.67	1	-1.64
Ministral-3B-2512	30.77	4.40	24.18	24.01	1.16	25.58	-78.84	1	+0.00
Gemma-3-4B-I	73.33	12.09	20.88	28.20	0.00	24.05	-91.67	1	+2.75
Qwen3-8B	59.65	18.68	43.96	56.00	1.54	61.54	-63.84	1	+5.49
Gemma-4-26B-A4B-I	58.82	52.75	21.98	65.81	11.11	74.07	-63.89	1	+4.40
Olmo-3.1-32B-I	94.12	29.67	26.37	48.58	0.00	40.00	-87.10	1	+5.49
Aya-Expanse-32B	80.85	20.88	30.77	44.39	0.00	41.18	-82.61	1	-4.94
Llama-3.3-70B-I	**100.00**	30.77	30.77	53.66	3.70	51.85	-71.98	1	+0.55
Qwen3.6-Plus	100.00	35.16	35.16	62.45	4.26	68.09	-68.47	1	+6.59
DeepSeek-V3.1	70.37	20.88	38.46	51.87	1.43	50.00	-89.05	1	+2.20
Command-A	76.67	25.27	40.66	58.47	1.64	60.66	-75.03	1	+1.10
o4-mini	76.36	37.36	23.08	51.94	0.00	40.38	-87.18	1	+5.49
GPT-5.2	88.57	34.07	42.86	67.38	6.67	86.67	-60.72	**2**	+7.14
Gemini-3-Flash	94.29	40.66	36.26	68.21	13.64	75.00	-65.08	1	-0.55
Claude Haiku 4.5	100.00	38.46	38.46	66.35	11.36	79.55	-63.11	1	-2.20
Human (Base)	27.59	27.47	4.40	26.27	**50.00**	**100.00**	+21.26	1	-1.65
Human (Expert)	0.00	3.30	0.00	**2.67**	—	—	—	1	+0.55

Table 20. Sensitivity profile — Sub-Type: *A1. Multi-Quantifier First-Order Logic*.

Model	*Cond. Inv. Retention (%)* TCIR_1	TCIR_2	TCIR_3	*Collapse Contribution (%)* TCC_1 ↓	TCC_2 ↓	TCC_3 ↓	*Rel. Sensitivity (pp)* RIS_1	RIS_2	RIS_3
Llama-3.2-1B-I	23.08	69.23	46.15	80.00	40.00	70.00	-2.47	+4.12	+0.82
Ministral-3B-2512	33.33	55.56	33.33	75.00	50.00	70.83	-0.55	+6.04	-0.55
Gemma-3-4B-I	38.71	45.16	22.58	67.86	57.14	82.14	+0.83	+3.02	-4.67
Qwen3-8B	51.52	75.76	36.36	63.27	30.61	75.51	+0.55	+18.14	-10.44
Gemma-4-26B-A4B-I	59.52	72.62	50.00	76.32	50.00	63.16	-4.67	+7.41	-13.47
Olmo-3.1-32B-I	41.82	70.91	23.64	71.79	38.46	89.74	-3.85	+13.74	-14.83
Aya-Expanse-32B	15.69	64.71	33.33	90.91	40.91	77.27	-13.46	+14.01	-3.57
Llama-3.3-70B-I	43.08	67.69	41.54	81.40	44.19	72.09	-6.59	+10.99	-7.69
Qwen3.6-Plus	57.89	68.42	42.11	65.96	48.94	65.96	+1.10	+9.89	-12.09
DeepSeek-V3.1	33.93	66.07	26.79	72.92	37.50	81.25	-4.39	+15.39	-8.79
Command-A	50.75	59.70	47.76	66.67	47.92	60.42	+-0.00	+6.60	-2.20
o4-mini	36.67	60.00	20.00	77.78	51.11	93.33	-5.77	+9.61	-16.76
GPT-5.2	67.06	64.71	51.76	50.00	56.52	63.04	+7.15	+4.95	-7.14
Gemini-3-Flash	56.25	62.50	57.50	60.00	60.00	57.78	-2.20	+3.30	-1.10
Claude Haiku 4.5	54.88	65.85	59.76	70.73	48.78	60.98	-4.12	+5.77	+0.28
Human (Base)	86.81	91.21	90.11	33.33	66.67	66.67	-4.12	+0.28	-0.82
Human (Expert)	**100.00**	**97.80**	**98.90**	**0.00**	**0.00**	**0.00**	+0.82	-1.38	-0.28

H.1.2. A2. SET ALGEBRA + CARDINALITY WITH CONSTRAINTS

Table 21. Core accuracy — Sub-Type: *A2. Set Algebra + Cardinality with Constraints.*

Model	Base	*Accuracy (%)* $\mathcal{I}_1$	$\mathcal{I}_2$	$\mathcal{I}_3$	*Mean Inv.* $\overline{\mathcal{I}}$	*Base*$-\overline{\mathcal{I}}$ Δ
Llama-3.2-1B-I	4.71	3.53	3.53	4.71	3.92	+0.79
Ministral-3B-2512	3.53	5.88	8.24	2.35	5.49	-1.96
Gemma-3-4B-I	4.71	7.06	2.35	2.35	3.92	+0.79
Qwen3-8B	48.24	44.71	45.88	44.71	45.10	+3.14
Gemma-4-26B-A4B-I	9.41	12.94	10.59	14.12	12.55	**-3.14**
Olmo-3.1-32B-I	10.59	11.76	4.71	7.06	7.84	+2.75
Aya-Expanse-32B	4.71	7.06	5.88	7.06	6.67	-1.96
Llama-3.3-70B-I	14.12	8.24	9.41	8.24	8.63	+5.49
Qwen3.6-Plus	48.24	47.06	45.88	47.06	46.67	+1.57
DeepSeek-V3.1	9.41	10.59	5.88	7.06	7.84	+1.57
Command-A	10.59	11.76	8.24	8.24	9.41	+1.18
o4-mini	7.06	5.88	9.41	7.06	7.45	-0.39
GPT-5.2	45.88	47.06	41.18	42.35	43.53	+2.35
Gemini-3-Flash	32.94	27.06	22.35	24.71	24.71	+8.23
Claude Haiku 4.5	29.41	28.24	29.41	38.82	32.16	-2.75
Human (Base)	90.59	88.24	87.06	84.71	86.67	+3.92
Human (Expert)	**98.82**	**100.00**	**97.65**	**100.00**	**99.22**	-0.40

Table 22. Invariance stability — Sub-Type: *A2. Set Algebra + Cardinality with Constraints.*

Model	*Global Stability (%)* IMI↑	CIR↑	ICR↓	*Novel (%)* BSS↑	CE↓	*Pairwise Failure Dep. (%)* $P(\mathcal{I}_1{=}0\|\mathcal{I}_2{=}0)$	$P(\mathcal{I}_1{=}0\|\mathcal{I}_3{=}0)$	$P(\mathcal{I}_2{=}0\|\mathcal{I}_3{=}0)$
Llama-3.2-1B-I	2.35	50.00	1.18	**100.00**	3.82	98.78	100.00	98.77
Ministral-3B-2512	0.00	0.00	**0.00**	54.55	10.50	94.87	95.18	92.77
Gemma-3-4B-I	0.00	0.00	3.53	80.00	9.99	92.77	93.98	97.59
Qwen3-8B	35.29	73.17	4.71	53.33	15.43	93.48	91.49	85.11
Gemma-4-26B-A4B-I	3.53	37.50	3.53	55.56	18.07	92.11	91.78	94.52
Olmo-3.1-32B-I	3.53	33.33	5.88	90.91	11.39	91.36	94.94	98.73
Aya-Expanse-32B	1.18	25.00	1.18	60.00	10.21	95.00	97.47	96.20
Llama-3.3-70B-I	1.18	8.33	8.24	70.59	16.89	96.10	94.87	93.59
Qwen3.6-Plus	41.18	85.37	1.18	50.00	7.64	95.65	95.56	95.56
DeepSeek-V3.1	1.18	12.50	5.88	87.50	16.16	92.50	91.14	96.20
Command-A	1.18	11.11	3.53	93.33	15.20	91.03	91.03	94.87
o4-mini	3.53	50.00	3.53	50.00	11.68	98.70	97.47	93.67
GPT-5.2	36.47	79.49	3.53	66.67	11.90	90.00	89.80	93.88
Gemini-3-Flash	16.47	50.00	12.94	72.73	22.55	87.88	89.06	93.75
Claude Haiku 4.5	12.94	44.00	4.71	90.32	31.36	86.67	86.54	90.38
Human (Base)	57.65	63.64	2.35	44.44	35.69	9.09	7.69	0.00
Human (Expert)	**97.65**	**98.81**	0.00	100.00	**2.13**	0.00	0.00	0.00

Table 23. Novel metrics — Sub-Type: *A2. Set Algebra + Cardinality with Constraints*.

	Bilateral Symmetry			*Entropy*	*Invariant Recovery (%)*			*Fragility Gradient*	
Model	BSS↑	$F^{+}(\%)$	$F^{-}(\%)$	CE↓	IRR_{full} ↑	IRR_{part} ↑	IRR_{Δ}	FG	$\overline{g}$
Llama-3.2-1B-I	**100.00**	2.35	2.35	3.82	0.00	2.47	-50.00	1	-0.59
Ministral-3B-2512	54.55	3.53	9.41	10.50	0.00	9.76	-100.00	1	+1.76
Gemma-3-4B-I	80.00	4.71	7.06	9.99	0.00	7.41	-100.00	**2**	+2.36
Qwen3-8B	53.33	12.94	4.71	15.43	2.27	9.09	-24.56	1	+0.00
Gemma-4-26B-A4B-I	55.56	5.88	15.29	18.07	1.30	16.88	-61.20	1	-0.59
Olmo-3.1-32B-I	90.91	7.06	5.88	11.39	0.00	6.58	-66.67	1	+2.35
Aya-Expanse-32B	60.00	3.53	8.24	10.21	0.00	8.64	-75.00	1	+0.00
Llama-3.3-70B-I	70.59	12.94	7.06	16.89	1.37	8.22	-90.30	1	+0.00
Qwen3.6-Plus	50.00	7.06	2.35	7.64	2.27	4.55	-12.36	1	+0.00
DeepSeek-V3.1	87.50	8.24	10.59	16.16	0.00	11.69	-87.50	1	+1.77
Command-A	93.33	9.41	8.24	15.20	0.00	9.21	-88.89	2	+1.76
o4-mini	50.00	3.53	10.59	11.68	0.00	11.39	-50.00	1	-0.59
GPT-5.2	66.67	9.41	4.71	11.90	2.17	8.70	-18.34	1	+2.35
Gemini-3-Flash	72.73	16.47	9.41	22.55	0.00	14.04	-50.00	1	+1.18
Claude Haiku 4.5	90.32	16.47	20.00	31.36	3.33	28.33	-52.67	0	-5.29
Human (Base)	44.44	32.94	9.41	35.69	**50.00**	**100.00**	+13.64	2	+1.76
Human (Expert)	100.00	1.18	1.18	**2.13**	0.00	100.00	-1.19	1	+0.00

Table 24. Sensitivity profile — Sub-Type: *A2. Set Algebra + Cardinality with Constraints*.

	Cond. Inv. Retention (%)			*Collapse Contribution (%)*			*Rel. Sensitivity (pp)*		
Model	TCIR_1	TCIR_2	TCIR_3	$\text{TCC}_1 \downarrow$	$\text{TCC}_2 \downarrow$	$\text{TCC}_3 \downarrow$	RIS_1	RIS_2	RIS_3
Llama-3.2-1B-I	50.00	50.00	66.67	100.00	66.67	66.67	-0.59	-0.59	+0.59
Ministral-3B-2512	45.45	63.64	18.18	62.50	37.50	100.00	+0.88	+3.24	-2.65
Gemma-3-4B-I	60.00	20.00	20.00	44.44	77.78	88.89	+2.94	-1.77	-1.77
Qwen3-8B	84.44	86.67	84.44	85.71	57.14	42.86	-1.17	+-0.00	-1.17
Gemma-4-26B-A4B-I	52.38	42.86	57.14	66.67	73.33	53.33	+1.18	-1.17	+2.36
Olmo-3.1-32B-I	71.43	28.57	42.86	40.00	90.00	80.00	+3.23	-3.82	-1.47
Aya-Expanse-32B	54.55	45.45	54.55	50.00	62.50	62.50	+0.88	-0.30	+0.88
Llama-3.3-70B-I	38.89	44.44	38.89	83.33	75.00	75.00	-1.76	-0.59	-1.76
Qwen3.6-Plus	93.02	90.70	93.02	100.00	100.00	50.00	+0.00	-1.18	+0.00
DeepSeek-V3.1	52.94	29.41	35.29	57.14	78.57	71.43	+2.35	-2.36	-1.18
Command-A	62.50	43.75	43.75	40.00	70.00	80.00	+2.05	-1.47	-1.47
o4-mini	33.33	53.33	40.00	83.33	58.33	75.00	-1.47	+2.06	-0.29
GPT-5.2	93.02	81.40	83.72	50.00	83.33	83.33	+2.94	-2.94	-1.77
Gemini-3-Flash	63.89	52.78	58.33	63.16	78.95	78.95	+0.30	-4.41	-2.05
Claude Haiku 4.5	57.14	59.52	78.57	68.42	73.68	36.84	-3.23	-2.06	+7.35
Human (Base)	88.24	87.06	84.71	50.00	**33.33**	50.00	+0.59	-0.59	-2.94
Human (Expert)	**100.00**	**97.65**	**100.00**	**0.00**	100.00	**0.00**	+0.88	-1.47	+0.88

H.1.3. A3. CONDITIONAL OVERRIDE TRANSFORMATION SYSTEMS

Table 25. Core accuracy — Sub-Type: *A3. Conditional Override Transformation Systems.*

Model	Base	*Accuracy (%)* $\mathcal{I}_1$	$\mathcal{I}_2$	$\mathcal{I}_3$	*Mean Inv.* $\overline{\mathcal{I}}$	*Base*$-\overline{\mathcal{I}}$ Δ
Llama-3.2-1B-I	0.00	0.00	0.00	1.69	0.56	-0.56
Ministral-3B-2512	0.00	0.00	0.00	1.69	0.56	-0.56
Gemma-3-4B-I	0.00	5.08	1.69	0.00	2.26	-2.26
Qwen3-8B	40.68	81.36	42.37	77.97	67.23	**-26.55**
Gemma-4-26B-A4B-I	3.39	3.39	3.39	8.47	5.08	-1.69
Olmo-3.1-32B-I	1.69	0.00	1.69	3.39	1.69	-0.00
Aya-Expanse-32B	5.08	3.39	1.69	3.39	2.82	+2.26
Llama-3.3-70B-I	5.08	8.47	8.47	6.78	7.91	-2.83
Qwen3.6-Plus	89.83	91.53	66.10	84.75	80.79	+9.04
DeepSeek-V3.1	1.69	1.69	0.00	1.69	1.13	+0.56
Command-A	1.69	10.17	3.39	8.47	7.34	-5.65
o4-mini	0.00	1.69	0.00	3.39	1.69	-1.69
GPT-5.2	91.53	91.53	71.19	88.14	83.62	+7.91
Gemini-3-Flash	8.47	6.78	3.39	10.17	6.78	+1.69
Claude Haiku 4.5	76.27	81.36	61.02	69.49	70.62	+5.65
Human (Base)	91.53	96.61	88.14	84.75	89.83	+1.70
Human (Expert)	**100.00**	**100.00**	**98.31**	**100.00**	**99.44**	+0.56

Table 26. Invariance stability — Sub-Type: *A3. Conditional Override Transformation Systems.*

Model	*Global Stability (%)* IMI↑	CIR↑	ICR↓	*Novel (%)* BSS↑	CE↓	*Pairwise Failure Dep. (%)* $P(\mathcal{I}_1{=}0\|\mathcal{I}_2{=}0)$	$P(\mathcal{I}_1{=}0\|\mathcal{I}_3{=}0)$	$P(\mathcal{I}_2{=}0\|\mathcal{I}_3{=}0)$
Llama-3.2-1B-I	0.00	0.00	**0.00**	0.00	**1.38**	100.00	100.00	100.00
Ministral-3B-2512	0.00	0.00	0.00	0.00	1.38	100.00	100.00	100.00
Gemma-3-4B-I	0.00	0.00	0.00	0.00	5.50	94.83	94.92	98.31
Qwen3-8B	18.64	45.83	3.39	63.41	62.13	32.35	53.85	61.54
Gemma-4-26B-A4B-I	0.00	0.00	0.00	66.67	8.57	96.49	100.00	98.15
Olmo-3.1-32B-I	0.00	0.00	1.69	50.00	5.50	100.00	100.00	98.25
Aya-Expanse-32B	0.00	0.00	3.39	**100.00**	8.25	96.55	98.25	98.25
Llama-3.3-70B-I	0.00	0.00	5.08	42.86	20.21	90.74	92.73	90.91
Qwen3.6-Plus	61.02	67.92	1.69	11.11	25.39	25.00	55.56	66.67
DeepSeek-V3.1	0.00	0.00	1.69	100.00	3.07	98.31	98.28	100.00
Command-A	1.69	**100.00**	0.00	0.00	13.75	91.23	90.74	98.15
o4-mini	0.00	0.00	0.00	0.00	3.07	98.31	100.00	100.00
GPT-5.2	69.49	75.93	1.69	0.00	18.20	29.41	71.43	85.71
Gemini-3-Flash	0.00	0.00	3.39	90.91	15.13	92.98	98.11	96.23
Claude Haiku 4.5	38.98	51.11	5.08	53.33	43.81	30.43	44.44	44.44
Human (Base)	67.80	74.07	6.78	52.63	27.41	0.00	11.11	33.33
Human (Expert)	**98.31**	98.31	0.00	0.00	1.38	0.00	0.00	0.00

Table 27. Novel metrics — Sub-Type: *A3. Conditional Override Transformation Systems*.

Model	*Bilateral Symmetry* BSS↑	$F^+(\%)$	$F^-(\%)$	*Entropy* CE↓	*Invariant Recovery (%)* IRR_{full} ↑	IRR_{part} ↑	IRR_{Δ}	*Fragility Gradient* FG	$\overline{g}$
Llama-3.2-1B-I	0.00	0.00	1.69	**1.38**	0.00	1.69	-100.00	1	-0.84
Ministral-3B-2512	0.00	0.00	1.69	1.38	0.00	1.69	-100.00	1	-0.84
Gemma-3-4B-I	0.00	0.00	6.78	5.50	0.00	6.78	-100.00	**2**	+2.54
Qwen3-8B	63.41	22.03	47.46	62.13	25.71	80.00	-28.46	1	+1.70
Gemma-4-26B-A4B-I	66.67	3.39	6.78	8.57	0.00	7.02	-100.00	1	-2.54
Olmo-3.1-32B-I	50.00	1.69	5.08	5.50	0.00	5.17	-100.00	0	-1.69
Aya-Expanse-32B	**100.00**	5.08	5.08	8.25	0.00	5.36	-100.00	1	+0.00
Llama-3.3-70B-I	42.86	5.08	18.64	20.21	0.00	19.64	-100.00	2	+0.84
Qwen3.6-Plus	11.11	28.81	1.69	25.39	0.00	16.67	-32.08	1	+3.39
DeepSeek-V3.1	100.00	1.69	1.69	3.07	0.00	1.72	-100.00	1	+0.00
Command-A	0.00	0.00	16.95	13.75	0.00	17.24	+0.00	1	+0.85
o4-mini	0.00	0.00	3.39	3.07	0.00	3.39	-100.00	1	-0.85
GPT-5.2	0.00	22.03	0.00	18.20	0.00	0.00	-24.07	1	+1.69
Gemini-3-Flash	90.91	8.47	10.17	15.13	0.00	11.11	-100.00	1	-1.69
Claude Haiku 4.5	53.33	37.29	13.56	43.81	7.14	57.14	-41.75	1	+5.93
Human (Base)	52.63	23.73	8.47	27.41	**100.00**	**100.00**	+74.07	2	+5.93
Human (Expert)	0.00	1.69	0.00	1.38	—	—	—	1	+0.00

Table 28. Sensitivity profile — Sub-Type: *A3. Conditional Override Transformation Systems*.

Model	*Cond. Inv. Retention (%)* TCIR_1	TCIR_2	TCIR_3	*Collapse Contribution (%)* TCC_1 ↓	TCC_2 ↓	TCC_3 ↓	*Rel. Sensitivity (pp)* RIS_1	RIS_2	RIS_3
Llama-3.2-1B-I	0.00	0.00	**100.00**	100.00	100.00	**0.00**	-0.42	-0.42	+1.27
Ministral-3B-2512	0.00	0.00	100.00	100.00	100.00	0.00	-0.42	-0.42	+1.27
Gemma-3-4B-I	75.00	25.00	0.00	25.00	75.00	100.00	+3.39	+-0.00	-1.69
Qwen3-8B	92.31	48.08	88.46	19.05	95.24	9.52	+20.77	-18.22	+17.38
Gemma-4-26B-A4B-I	33.33	33.33	83.33	75.00	75.00	25.00	-1.27	-1.27	+3.81
Olmo-3.1-32B-I	0.00	25.00	50.00	100.00	75.00	50.00	-1.69	+-0.00	+1.70
Aya-Expanse-32B	33.33	16.67	33.33	80.00	80.00	80.00	+0.00	-1.70	+0.00
Llama-3.3-70B-I	35.71	35.71	28.57	64.29	64.29	71.43	+1.27	+1.27	-0.42
Qwen3.6-Plus	**100.00**	72.22	92.59	**0.00**	100.00	50.00	+8.48	-16.95	+1.70
DeepSeek-V3.1	50.00	0.00	50.00	50.00	100.00	50.00	+0.42	-1.27	+0.42
Command-A	54.55	18.18	45.45	50.00	90.00	60.00	+4.24	-2.54	+2.54
o4-mini	50.00	0.00	100.00	50.00	100.00	0.00	+0.42	-1.27	+2.12
GPT-5.2	100.00	77.78	96.30	0.00	100.00	100.00	+5.94	-14.40	+2.55
Gemini-3-Flash	36.36	18.18	54.55	87.50	75.00	62.50	-0.42	-3.81	+2.97
Claude Haiku 4.5	90.57	67.92	77.36	30.00	50.00	70.00	+9.33	-11.01	-2.54
Human (Base)	96.61	88.14	84.75	25.00	75.00	100.00	+6.36	-2.11	-5.50
Human (Expert)	100.00	**98.31**	100.00	0.00	**0.00**	0.00	+0.42	-1.27	+0.42

H.1.4. A4. CONSTRAINT SATISFACTION IN PROPOSITIONAL LOGIC

Table 29. Core accuracy — Sub-Type: *A4. Constraint Satisfaction in Propositional Logic*.

Model	Base	*Accuracy (%)* $\mathcal{I}_1$	$\mathcal{I}_2$	$\mathcal{I}_3$	*Mean Inv.* $\overline{\mathcal{I}}$	*Base*$-\overline{\mathcal{I}}$ Δ
Llama-3.2-1B-I	12.31	9.23	10.77	12.31	10.77	+1.54
Ministral-3B-2512	21.54	18.46	21.54	29.23	23.08	**-1.54**
Gemma-3-4B-I	20.00	13.85	18.46	16.92	16.41	+3.59
Qwen3-8B	53.85	38.46	53.85	46.15	46.15	+7.70
Gemma-4-26B-A4B-I	60.00	50.77	41.54	40.00	44.10	+15.90
Olmo-3.1-32B-I	33.85	23.08	38.46	21.54	27.69	+6.16
Aya-Expanse-32B	20.00	20.00	18.46	21.54	20.00	+0.00
Llama-3.3-70B-I	44.62	38.46	35.38	33.85	35.90	+8.72
Qwen3.6-Plus	75.38	61.54	67.69	64.62	64.62	+10.76
DeepSeek-V3.1	47.69	29.23	36.92	32.31	32.82	+14.87
Command-A	44.62	44.62	27.69	44.62	38.98	+5.64
o4-mini	43.08	27.69	35.38	27.69	30.25	+12.83
GPT-5.2	69.23	55.38	64.62	67.69	62.56	+6.67
Gemini-3-Flash	56.92	53.85	66.15	53.85	57.95	-1.03
Claude Haiku 4.5	64.62	50.77	41.54	46.15	46.15	+18.47
Human (Base)	92.31	92.31	90.77	92.31	91.80	+0.51
Human (Expert)	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	+0.00

Table 30. Invariance stability — Sub-Type: *A4. Constraint Satisfaction in Propositional Logic*.

Model	*Global Stability (%)* IMI↑	CIR↑	ICR↓	*Novel (%)* BSS↑	CE↓	*Pairwise Failure Dep. (%)* $P(\mathcal{I}_1{=}0\|\mathcal{I}_2{=}0)$	$P(\mathcal{I}_1{=}0\|\mathcal{I}_3{=}0)$	$P(\mathcal{I}_2{=}0\|\mathcal{I}_3{=}0)$
Llama-3.2-1B-I	0.00	0.00	7.69	94.12	22.96	91.38	94.74	91.23
Ministral-3B-2512	1.54	7.14	15.38	74.29	46.59	82.35	91.30	80.43
Gemma-3-4B-I	1.54	7.69	13.85	80.00	38.90	86.79	87.04	83.33
Qwen3-8B	7.69	14.29	10.77	66.67	58.49	73.33	68.57	57.14
Gemma-4-26B-A4B-I	13.85	23.08	27.69	66.67	60.81	57.89	66.67	66.67
Olmo-3.1-32B-I	6.15	18.18	23.08	**97.14**	49.20	90.00	82.35	66.67
Aya-Expanse-32B	0.00	0.00	15.38	70.27	48.79	77.36	80.39	80.39
Llama-3.3-70B-I	7.69	17.24	21.54	82.93	54.66	71.43	74.42	72.09
Qwen3.6-Plus	44.62	59.18	10.77	33.33	31.99	80.95	73.91	60.87
DeepSeek-V3.1	1.54	3.23	29.23	77.55	65.80	78.05	75.00	68.18
Command-A	10.77	24.14	23.08	84.21	52.36	65.96	69.44	91.67
o4-mini	7.69	17.86	24.62	82.05	52.16	78.57	80.85	68.09
GPT-5.2	38.46	55.56	7.69	46.15	33.32	82.61	85.71	76.19
Gemini-3-Flash	18.46	32.43	15.38	86.36	60.14	59.09	56.67	43.33
Claude Haiku 4.5	15.38	23.81	26.15	60.87	61.48	60.53	57.14	71.43
Human (Base)	69.23	75.00	1.54	50.00	25.25	16.67	0.00	0.00
Human (Expert)	**100.00**	**100.00**	**0.00**	—	**0.00**	0.00	0.00	0.00

Table 31. Novel metrics — Sub-Type: *A4. Constraint Satisfaction in Propositional Logic*.

	Bilateral Symmetry			*Entropy*	*Invariant Recovery (%)*			*Fragility Gradient*	
Model	BSS↑	$F^{+}(\%)$	$F^{-}(\%)$	CE↓	IRR_{full} ↑	IRR_{part} ↑	IRR_{Δ}	FG	$\overline{g}$
Llama-3.2-1B-I	94.12	12.31	13.85	22.96	0.00	15.79	-100.00	0	-1.54
Ministral-3B-2512	74.29	20.00	33.85	46.59	3.92	43.14	-88.94	0	-5.38
Gemma-3-4B-I	80.00	18.46	27.69	38.90	0.00	34.62	-92.31	1	-1.54
Qwen3-8B	66.67	46.15	23.08	58.49	10.00	50.00	-75.71	1	-3.85
Gemma-4-26B-A4B-I	66.67	46.15	23.08	60.81	7.69	57.69	-69.23	**2**	+5.38
Olmo-3.1-32B-I	**97.14**	27.69	26.15	49.20	0.00	39.53	-81.82	1	+0.77
Aya-Expanse-32B	70.27	20.00	36.92	48.79	0.00	46.15	-100.00	1	-0.77
Llama-3.3-70B-I	82.93	36.92	26.15	54.66	8.33	47.22	-74.43	2	+2.31
Qwen3.6-Plus	33.33	30.77	6.15	31.99	6.25	25.00	-34.57	1	-1.54
DeepSeek-V3.1	77.55	46.15	29.23	65.80	5.88	55.88	-90.89	1	-1.54
Command-A	84.21	33.85	24.62	52.36	11.11	44.44	-64.75	1	+0.00
o4-mini	82.05	35.38	24.62	52.16	2.70	43.24	-79.44	1	+0.00
GPT-5.2	46.15	30.77	9.23	33.32	20.00	30.00	-24.44	0	-6.16
Gemini-3-Flash	86.36	38.46	29.23	60.14	17.86	67.86	-49.71	1	+0.00
Claude Haiku 4.5	60.87	49.23	21.54	61.48	4.35	60.87	-71.84	1	+2.31
Human (Base)	50.00	23.08	7.69	25.25	**100.00**	**100.00**	+75.00	1	+0.00
Human (Expert)	—	0.00	0.00	**0.00**	—	—	—	2	+0.00

Table 32. Sensitivity profile — Sub-Type: *A4. Constraint Satisfaction in Propositional Logic*.

	Cond. Inv. Retention (%)			*Collapse Contribution (%)*			*Rel. Sensitivity (pp)*		
Model	TCIR_1	TCIR_2	TCIR_3	TCC_1 ↓	TCC_2 ↓	TCC_3 ↓	RIS_1	RIS_2	RIS_3
Llama-3.2-1B-I	35.29	41.18	47.06	78.57	50.00	64.29	-1.92	-0.38	+1.16
Ministral-3B-2512	33.33	38.89	52.78	76.67	66.67	56.67	-4.23	-1.15	+6.54
Gemma-3-4B-I	29.03	38.71	35.48	77.78	66.67	70.37	-3.46	+1.15	-0.39
Qwen3-8B	50.00	70.00	60.00	89.47	47.37	57.89	-9.62	+5.77	-1.93
Gemma-4-26B-A4B-I	61.11	50.00	48.15	64.52	67.74	74.19	+2.69	-6.54	-8.08
Olmo-3.1-32B-I	38.46	64.10	35.90	71.88	43.75	71.88	-6.15	+9.23	-7.69
Aya-Expanse-32B	35.14	32.43	37.84	67.65	70.59	64.71	+0.00	-1.54	+1.54
Llama-3.3-70B-I	54.35	50.00	47.83	64.29	71.43	71.43	+0.38	-2.70	-4.23
Qwen3.6-Plus	75.47	83.02	79.25	100.00	50.00	50.00	-5.77	+0.38	-2.69
DeepSeek-V3.1	38.00	48.00	42.00	72.22	66.67	69.44	-7.31	+0.38	-4.23
Command-A	64.44	40.00	64.44	48.15	85.19	59.26	+4.24	-12.69	+4.24
o4-mini	40.91	52.27	40.91	77.42	61.29	74.19	-5.77	+1.92	-5.77
GPT-5.2	70.59	82.35	86.27	100.00	71.43	57.14	-8.85	+0.39	+3.46
Gemini-3-Flash	62.50	76.79	62.50	62.50	41.67	62.50	-3.84	+8.46	-3.84
Claude Haiku 4.5	58.93	48.21	53.57	63.33	76.67	70.00	+0.00	-9.23	-4.62
Human (Base)	92.31	90.77	92.31	100.00	100.00	**0.00**	+0.39	-1.15	+0.39
Human (Expert)	**100.00**	**100.00**	**100.00**	**0.00**	**0.00**	0.00	+0.00	+0.00	+0.00

H.1.5. A5. INDUCTIVE RULE + FORMAL DEDUCTION

Table 33. Core accuracy — Sub-Type: *A5. Inductive Rule + Formal Deduction*.

Model	Base	*Accuracy (%)* $\mathcal{I}_1$	$\mathcal{I}_2$	$\mathcal{I}_3$	*Mean Inv.* $\overline{\mathcal{I}}$	*Base*$-\overline{\mathcal{I}}$ Δ
Llama-3.2-1B-I	6.38	6.38	6.38	6.38	6.38	+0.00
Ministral-3B-2512	19.15	14.89	12.77	21.28	16.31	+2.84
Gemma-3-4B-I	17.02	17.02	17.02	17.02	17.02	+0.00
Qwen3-8B	68.09	42.55	42.55	48.94	44.68	+23.41
Gemma-4-26B-A4B-I	61.70	48.94	61.70	57.45	56.03	+5.67
Olmo-3.1-32B-I	40.43	29.79	34.04	23.40	29.08	+11.35
Aya-Expanse-32B	27.66	17.02	23.40	40.43	26.95	+0.71
Llama-3.3-70B-I	34.04	23.40	31.91	38.30	31.20	+2.84
Qwen3.6-Plus	74.47	55.32	65.96	72.34	64.54	+9.93
DeepSeek-V3.1	44.68	23.40	46.81	40.43	36.88	+7.80
Command-A	48.94	25.53	46.81	46.81	39.72	+9.22
o4-mini	36.17	19.15	31.91	38.30	29.79	+6.38
GPT-5.2	59.57	44.68	65.96	70.21	60.28	-0.71
Gemini-3-Flash	70.21	51.06	57.45	72.34	60.28	+9.93
Claude Haiku 4.5	46.81	34.04	53.19	51.06	46.10	+0.71
Human (Base)	78.72	93.62	95.74	93.62	94.33	**-15.61**
Human (Expert)	**100.00**	**100.00**	**97.87**	**100.00**	**99.29**	+0.71

Table 34. Invariance stability — Sub-Type: *A5. Inductive Rule + Formal Deduction*.

Model	*Global Stability (%)* IMI↑	CIR↑	ICR↓	*Novel (%)* BSS↑	CE↓	*Pairwise Failure Dep. (%)* $P(\mathcal{I}_1{=}0\|\mathcal{I}_2{=}0)$	$P(\mathcal{I}_1{=}0\|\mathcal{I}_3{=}0)$	$P(\mathcal{I}_2{=}0\|\mathcal{I}_3{=}0)$
Llama-3.2-1B-I	0.00	0.00	4.26	60.00	17.26	93.18	95.45	93.18
Ministral-3B-2512	0.00	0.00	12.77	**100.00**	34.28	82.93	86.49	91.89
Gemma-3-4B-I	0.00	0.00	10.64	80.00	36.93	84.62	84.62	84.62
Qwen3-8B	12.77	18.75	29.79	42.42	61.78	62.96	75.00	66.67
Gemma-4-26B-A4B-I	21.28	34.48	12.77	88.24	61.10	50.00	60.00	50.00
Olmo-3.1-32B-I	2.13	5.26	27.66	80.00	57.41	74.19	66.67	75.00
Aya-Expanse-32B	0.00	0.00	12.77	89.66	53.27	80.56	82.14	89.29
Llama-3.3-70B-I	4.26	12.50	14.89	100.00	51.54	75.00	82.76	79.31
Qwen3.6-Plus	27.66	37.14	8.51	62.50	57.65	56.25	30.77	38.46
DeepSeek-V3.1	8.51	19.05	17.02	82.76	54.07	84.00	75.00	71.43
Command-A	6.38	13.04	12.77	78.79	58.57	80.00	80.00	64.00
o4-mini	6.38	17.65	14.89	88.00	46.37	84.38	86.21	79.31
GPT-5.2	19.15	32.14	8.51	84.85	59.37	56.25	78.57	57.14
Gemini-3-Flash	21.28	30.30	10.64	64.71	61.10	50.00	38.46	61.54
Claude Haiku 4.5	10.64	22.73	12.77	86.67	55.40	77.27	78.26	65.22
Human (Base)	63.83	81.08	**0.00**	82.35	29.75	0.00	0.00	0.00
Human (Expert)	**97.87**	**97.87**	0.00	0.00	**1.73**	0.00	0.00	0.00

Table 35. Novel metrics — Sub-Type: *A5. Inductive Rule + Formal Deduction.*

Model	*Bilateral Symmetry* BSS↑	$F^+(\%)$	$F^-(\%)$	*Entropy* CE↓	*Invariant Recovery (%)* IRR_{full} ↑	IRR_{part} ↑	IRR_{Δ}	*Fragility Gradient* FG	$\overline{g}$
Llama-3.2-1B-I	60.00	6.38	14.89	17.26	0.00	15.91	-100.00	**2**	+0.00
Ministral-3B-2512	**100.00**	19.15	19.15	34.28	0.00	23.68	-100.00	1	-3.19
Gemma-3-4B-I	80.00	17.02	25.53	36.93	0.00	30.77	-100.00	2	+0.00
Qwen3-8B	42.42	55.32	14.89	61.78	6.67	46.67	-74.58	1	-3.19
Gemma-4-26B-A4B-I	88.24	40.43	31.91	61.10	5.56	83.33	-59.96	1	-4.25
Olmo-3.1-32B-I	80.00	38.30	25.53	57.41	3.57	42.86	-91.17	1	+3.19
Aya-Expanse-32B	89.66	27.66	34.04	53.27	0.00	47.06	-100.00	0	-11.71
Llama-3.3-70B-I	100.00	29.79	29.79	51.54	3.23	45.16	-84.27	0	-7.45
Qwen3.6-Plus	62.50	46.81	21.28	57.65	0.00	83.33	-62.86	0	-8.51
DeepSeek-V3.1	82.76	36.17	25.53	54.07	0.00	46.15	-80.95	1	-8.52
Command-A	78.79	42.55	27.66	58.57	4.17	54.17	-82.79	1	-10.64
o4-mini	88.00	29.79	23.40	46.37	0.00	36.67	-82.35	0	-9.58
GPT-5.2	84.85	40.43	29.79	59.37	21.05	73.68	-46.81	0	-12.76
Gemini-3-Flash	64.71	48.94	23.40	61.10	7.14	78.57	-62.56	0	-10.64
Claude Haiku 4.5	86.67	36.17	27.66	55.40	12.00	52.00	-65.27	1	-8.51
Human (Base)	82.35	14.89	21.28	29.75	**90.00**	**100.00**	+71.08	1	+0.00
Human (Expert)	0.00	2.13	0.00	**1.73**	—	—	—	1	+0.00

Table 36. Sensitivity profile — Sub-Type: *A5. Inductive Rule + Formal Deduction.*

Model	*Cond. Inv. Retention (%)* TCIR_1	TCIR_2	TCIR_3	*Collapse Contribution (%)* TCC_1 ↓	TCC_2 ↓	TCC_3 ↓	*Rel. Sensitivity (pp)* RIS_1	RIS_2	RIS_3
Llama-3.2-1B-I	30.00	30.00	30.00	77.78	66.67	77.78	+-0.00	+-0.00	+-0.00
Ministral-3B-2512	38.89	33.33	55.56	60.00	73.33	53.33	-2.13	-4.25	+4.26
Gemma-3-4B-I	40.00	40.00	40.00	58.82	70.59	64.71	+-0.00	+-0.00	+-0.00
Qwen3-8B	51.28	51.28	58.97	75.00	70.00	65.00	-7.98	-7.98	-1.59
Gemma-4-26B-A4B-I	52.27	65.91	61.36	65.00	60.00	75.00	-8.51	+4.25	+0.00
Olmo-3.1-32B-I	45.16	51.61	35.48	54.17	62.50	79.17	-2.12	+2.13	-8.51
Aya-Expanse-32B	27.59	37.93	65.52	68.18	81.82	40.91	-10.11	-3.73	+13.30
Llama-3.3-70B-I	36.67	50.00	60.00	75.00	60.00	60.00	-8.51	+-0.00	+6.39
Qwen3.6-Plus	57.78	68.89	75.56	71.43	78.57	35.71	-11.70	-1.06	+5.32
DeepSeek-V3.1	33.33	66.67	57.58	70.00	55.00	65.00	-15.43	+7.98	+1.60
Command-A	33.33	61.11	61.11	83.33	66.67	61.11	-16.49	+4.79	+4.79
o4-mini	32.14	53.57	64.29	77.78	66.67	50.00	-12.23	+0.53	+6.92
GPT-5.2	50.00	73.81	78.57	71.43	57.14	57.14	-15.43	+5.85	+10.10
Gemini-3-Flash	54.55	61.36	77.27	60.00	86.67	46.67	-11.71	-5.32	+9.57
Claude Haiku 4.5	45.71	71.43	68.57	68.75	56.25	56.25	-12.24	+6.91	+4.78
Human (Base)	93.62	95.74	93.62	**0.00**	**0.00**	100.00	+3.19	+5.31	+3.19
Human (Expert)	**100.00**	**97.87**	**100.00**	0.00	0.00	**0.00**	+0.53	-1.60	+0.53

H.1.6. A6. UNIT/REPRESENTATION INVARIANCE IN PHYSICS-STYLE WORD PROBLEMS

Table 37. Core accuracy — Sub-Type: *A6. Unit/Representation Invariance in Physics-Style Word Problems*.

Model	Base	*Accuracy (%)* $\mathcal{I}_1$	$\mathcal{I}_2$	$\mathcal{I}_3$	*Mean Inv.* $\overline{\mathcal{I}}$	*Base*$-\overline{\mathcal{I}}$ Δ
Llama-3.2-1B-I	2.17	2.17	4.35	4.35	3.62	-1.45
Ministral-3B-2512	6.52	10.87	4.35	6.52	7.25	-0.73
Gemma-3-4B-I	4.35	4.35	2.17	6.52	4.35	+0.00
Qwen3-8B	71.74	65.22	63.04	67.39	65.22	+6.52
Gemma-4-26B-A4B-I	10.87	15.22	8.70	15.22	13.05	**-2.18**
Olmo-3.1-32B-I	8.70	6.52	6.52	6.52	6.52	+2.18
Aya-Expanse-32B	6.52	4.35	4.35	8.70	5.80	+0.72
Llama-3.3-70B-I	13.04	17.39	8.70	13.04	13.04	-0.00
Qwen3.6-Plus	71.74	60.87	58.70	63.04	60.87	+10.87
DeepSeek-V3.1	13.04	15.22	8.70	17.39	13.77	-0.73
Command-A	13.04	6.52	10.87	8.70	8.70	+4.34
o4-mini	10.87	8.70	2.17	2.17	4.35	+6.52
GPT-5.2	78.26	71.74	76.09	78.26	75.36	+2.90
Gemini-3-Flash	67.39	63.04	52.17	63.04	59.42	+7.97
Claude Haiku 4.5	41.30	45.65	30.43	34.78	36.95	+4.35
Human (Base)	84.78	89.13	93.48	78.26	86.96	-2.18
Human (Expert)	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	+0.00

Table 38. Invariance stability — Sub-Type: *A6. Unit/Representation Invariance in Physics-Style Word Problems*.

Model	*Global Stability (%)* IMI↑	CIR↑	ICR↓	*Novel (%)* BSS↑	CE↓	*Pairwise Failure Dep. (%)* $P(\mathcal{I}_1{=}0\|\mathcal{I}_2{=}0)$	$P(\mathcal{I}_1{=}0\|\mathcal{I}_3{=}0)$	$P(\mathcal{I}_2{=}0\|\mathcal{I}_3{=}0)$
Llama-3.2-1B-I	2.17	**100.00**	**0.00**	0.00	3.53	100.00	100.00	97.73
Ministral-3B-2512	2.17	33.33	4.35	80.00	10.46	90.91	95.35	97.67
Gemma-3-4B-I	2.17	50.00	0.00	**100.00**	3.53	97.78	100.00	100.00
Qwen3-8B	39.13	54.55	8.70	50.00	37.32	58.82	66.67	66.67
Gemma-4-26B-A4B-I	2.17	20.00	2.17	66.67	21.57	88.10	84.62	97.44
Olmo-3.1-32B-I	0.00	0.00	6.52	88.89	16.69	95.35	93.02	95.35
Aya-Expanse-32B	2.17	33.33	2.17	100.00	7.46	97.73	100.00	97.62
Llama-3.3-70B-I	4.35	33.33	8.70	66.67	22.80	85.71	90.00	95.00
Qwen3.6-Plus	43.48	60.61	8.70	37.50	28.63	73.68	82.35	82.35
DeepSeek-V3.1	0.00	0.00	10.87	80.00	28.92	85.71	94.74	92.11
Command-A	2.17	16.67	6.52	75.00	14.93	95.12	97.62	92.86
o4-mini	2.17	20.00	8.70	66.67	10.99	93.33	93.33	100.00
GPT-5.2	54.35	69.44	4.35	62.50	29.86	63.64	70.00	50.00
Gemini-3-Flash	39.13	58.06	10.87	55.56	33.80	63.64	70.59	88.24
Claude Haiku 4.5	8.70	21.05	17.39	84.62	49.96	68.75	66.67	76.67
Human (Base)	60.87	71.79	4.35	77.78	33.80	0.00	10.00	20.00
Human (Expert)	**100.00**	100.00	0.00	—	**0.00**	0.00	0.00	0.00

Table 39. Novel metrics — Sub-Type: *A6. Unit/Representation Invariance in Physics-Style Word Problems*.

	Bilateral Symmetry			*Entropy*	*Invariant Recovery (%)*			*Fragility Gradient*	
Model	BSS↑	$F^{+}(\%)$	$F^{-}(\%)$	CE↓	IRR_{full} ↑	IRR_{part} ↑	IRR_{Δ}	FG	$\overline{g}$
Llama-3.2-1B-I	0.00	0.00	4.35	3.53	0.00	4.44	+0.00	1	-1.09
Ministral-3B-2512	80.00	4.35	6.52	10.46	0.00	6.98	-66.67	1	+2.17
Gemma-3-4B-I	**100.00**	2.17	2.17	3.53	0.00	2.27	-50.00	1	-1.08
Qwen3-8B	50.00	32.61	10.87	37.32	15.38	38.46	-30.07	1	-1.08
Gemma-4-26B-A4B-I	66.67	8.70	17.39	21.57	0.00	19.51	-80.00	1	+0.00
Olmo-3.1-32B-I	88.89	8.70	10.87	16.69	0.00	11.90	-100.00	**2**	+0.00
Aya-Expanse-32B	100.00	4.35	4.35	7.46	0.00	4.65	-66.67	1	-2.17
Llama-3.3-70B-I	66.67	8.70	17.39	22.80	0.00	20.00	-66.67	1	+2.18
Qwen3.6-Plus	37.50	28.26	6.52	28.63	7.69	23.08	-31.70	1	-1.09
DeepSeek-V3.1	80.00	13.04	19.57	28.92	2.50	22.50	-97.50	1	-1.09
Command-A	75.00	10.87	6.52	14.93	0.00	7.50	-83.33	1	-1.09
o4-mini	66.67	8.70	4.35	10.99	0.00	4.88	-80.00	2	+3.26
GPT-5.2	62.50	23.91	10.87	29.86	10.00	50.00	-20.56	0	-3.26
Gemini-3-Flash	55.56	28.26	10.87	33.80	6.67	33.33	-35.27	1	+0.00
Claude Haiku 4.5	84.62	32.61	23.91	49.96	7.41	40.74	-71.54	1	+5.44
Human (Base)	77.78	23.91	15.22	33.80	**42.86**	**100.00**	+14.65	1	+5.44
Human (Expert)	—	0.00	0.00	**0.00**	—	—	—	2	+0.00

Table 40. Sensitivity profile — Sub-Type: *A6. Unit/Representation Invariance in Physics-Style Word Problems*.

	Cond. Inv. Retention (%)			*Collapse Contribution (%)*			*Rel. Sensitivity (pp)*		
Model	TCIR_1	TCIR_2	TCIR_3	$\text{TCC}_1 \downarrow$	$\text{TCC}_2 \downarrow$	$\text{TCC}_3 \downarrow$	RIS_1	RIS_2	RIS_3
Llama-3.2-1B-I	33.33	66.67	66.67	100.00	50.00	50.00	-1.09	+1.09	+1.09
Ministral-3B-2512	83.33	33.33	50.00	20.00	80.00	60.00	+3.80	-2.72	-0.55
Gemma-3-4B-I	66.67	33.33	**100.00**	100.00	100.00	**0.00**	+0.00	-2.18	+2.17
Qwen3-8B	78.95	76.32	81.58	71.43	57.14	57.14	-1.63	-3.81	+0.54
Gemma-4-26B-A4B-I	53.85	30.77	53.85	44.44	100.00	55.56	+2.72	-3.80	+2.72
Olmo-3.1-32B-I	33.33	33.33	33.33	62.50	75.00	75.00	-0.55	-0.55	-0.55
Aya-Expanse-32B	40.00	40.00	80.00	100.00	66.67	33.33	-1.63	-1.63	+2.72
Llama-3.3-70B-I	57.14	28.57	42.86	50.00	83.33	66.67	+4.35	-4.34	+-0.00
Qwen3.6-Plus	77.78	75.00	80.56	83.33	83.33	83.33	-2.72	-4.89	-0.55
DeepSeek-V3.1	46.67	26.67	53.33	61.54	76.92	53.85	+1.63	-4.89	+3.80
Command-A	33.33	55.56	44.44	83.33	50.00	83.33	-3.26	+1.09	-1.08
o4-mini	57.14	14.29	14.29	50.00	100.00	100.00	+2.72	-3.81	-3.81
GPT-5.2	80.49	85.37	87.80	100.00	33.33	33.33	-4.35	+0.00	+2.17
Gemini-3-Flash	80.56	66.67	80.56	66.67	77.78	55.56	+1.63	-9.24	+1.63
Claude Haiku 4.5	70.00	46.67	53.33	47.06	82.35	58.82	+7.61	-7.61	-3.26
Human (Base)	89.13	93.48	78.26	33.33	33.33	83.33	+2.72	+7.07	-8.15
Human (Expert)	**100.00**	**100.00**	100.00	**0.00**	**0.00**	0.00	+0.00	+0.00	+0.00

H.2. B. Quantitative Reasoning and Optimization

Table 41. Core accuracy — Task Type: *B. Quantitative Reasoning and Optimization*.

Model	Base	*Accuracy (%)* $\mathcal{I}_1$	$\mathcal{I}_2$	$\mathcal{I}_3$	*Mean Inv.* $\overline{\mathcal{I}}$	*Base*$-\overline{\mathcal{I}}$ Δ
Llama-3.2-1B-I	4.26	4.51	5.01	2.01	3.84	+0.42
Ministral-3B-2512	3.51	3.51	4.01	2.01	3.18	+0.33
Gemma-3-4B-I	4.26	3.76	4.51	3.26	3.84	+0.42
Qwen3-8B	66.92	75.44	68.17	70.93	71.51	**-4.59**
Gemma-4-26B-A4B-I	10.28	9.77	10.28	8.27	9.44	+0.84
Olmo-3.1-32B-I	5.76	5.26	5.51	4.76	5.18	+0.58
Aya-Expanse-32B	5.51	3.26	6.77	4.51	4.85	+0.66
Llama-3.3-70B-I	18.05	22.56	21.55	18.05	20.72	-2.67
Qwen3.6-Plus	82.71	79.45	76.44	80.20	78.70	+4.01
DeepSeek-V3.1	7.77	7.27	8.52	5.76	7.18	+0.59
Command-A	7.02	13.78	7.52	9.77	10.36	-3.34
o4-mini	8.52	6.27	8.02	4.76	6.35	+2.17
GPT-5.2	81.70	80.20	80.20	79.70	80.03	+1.67
Gemini-3-Flash	29.07	29.07	28.32	30.08	29.16	-0.09
Claude Haiku 4.5	58.90	60.65	55.89	55.64	57.39	+1.51
Human (Base)	91.48	89.72	88.47	89.97	89.39	+2.09
Human (Expert)	**100.00**	**100.00**	**99.75**	**99.25**	**99.67**	+0.33

Table 42. Invariance stability — Task Type: *B. Quantitative Reasoning and Optimization*.

Model	*Global Stability (%)* IMI↑	CIR↑	ICR↓	*Novel (%)* BSS↑	CE↓	*Pairwise Failure Dep. (%)* $P(\mathcal{I}_1{=}0\|\mathcal{I}_2{=}0)$	$P(\mathcal{I}_1{=}0\|\mathcal{I}_3{=}0)$	$P(\mathcal{I}_2{=}0\|\mathcal{I}_3{=}0)$
Llama-3.2-1B-I	1.00	23.53	1.50	**100.00**	5.62	98.15	96.42	96.42
Ministral-3B-2512	0.50	14.29	2.26	77.42	6.63	97.65	97.19	96.93
Gemma-3-4B-I	1.25	29.41	2.26	75.00	6.74	98.16	97.41	96.63
Qwen3-8B	46.37	69.29	3.26	94.87	33.75	60.63	56.03	59.48
Gemma-4-26B-A4B-I	2.51	24.39	5.76	84.93	15.93	94.41	93.99	92.62
Olmo-3.1-32B-I	1.25	21.74	2.76	80.00	9.43	96.55	96.58	96.32
Aya-Expanse-32B	1.25	22.73	3.51	80.95	9.01	99.19	98.16	95.01
Llama-3.3-70B-I	3.76	20.83	6.52	76.51	31.95	82.75	85.02	85.02
Qwen3.6-Plus	66.42	80.30	2.26	47.06	17.90	68.09	67.09	70.89
DeepSeek-V3.1	0.25	3.23	5.76	78.95	16.45	94.52	93.35	92.82
Command-A	1.00	14.29	4.01	47.52	21.34	88.35	88.61	93.89
o4-mini	1.75	20.59	4.51	89.80	10.58	96.73	96.58	94.74
GPT-5.2	66.42	81.29	1.50	56.47	17.80	69.62	69.14	65.43
Gemini-3-Flash	13.78	47.41	6.27	93.85	27.99	86.71	88.17	85.30
Claude Haiku 4.5	26.07	44.26	12.28	88.03	50.70	58.52	57.63	58.19
Human (Base)	63.91	69.86	2.26	47.22	29.89	13.04	7.50	10.00
Human (Expert)	**99.00**	**99.00**	**0.00**	0.00	**0.81**	0.00	0.00	0.00

Table 43. Novel metrics — Task Type: *B. Quantitative Reasoning and Optimization*.

Model	*Bilateral Symmetry* BSS↑	F^+(%)	F^-(%)	*Entropy* CE↓	*Invariant Recovery (%)* IRR_{full} ↑	IRR_{part} ↑	IRR_{Δ}	*Fragility Gradient* FG	$\overline{g}$
Llama-3.2-1B-I	**100.00**	3.26	3.26	5.62	0.00	3.40	-76.47	1	+1.25
Ministral-3B-2512	77.42	3.01	4.76	6.63	0.00	4.94	-85.71	1	+0.75
Gemma-3-4B-I	75.00	3.01	5.01	6.74	0.00	5.24	-70.59	1	+0.25
Qwen3-8B	94.87	20.55	18.55	33.75	17.42	56.06	-13.29	1	+2.26
Gemma-4-26B-A4B-I	84.93	7.77	10.53	15.93	0.28	11.73	-75.33	1	+0.75
Olmo-3.1-32B-I	80.00	4.51	6.77	9.43	0.00	7.18	-78.26	1	+0.25
Aya-Expanse-32B	80.95	4.26	6.27	9.01	0.27	6.63	-77.00	1	-0.63
Llama-3.3-70B-I	76.51	14.29	23.06	31.95	2.14	28.13	-77.03	**2**	+2.26
Qwen3.6-Plus	47.06	16.29	5.01	17.90	1.45	28.99	-18.25	1	-0.38
DeepSeek-V3.1	78.95	7.52	11.53	16.45	0.27	12.50	-96.50	1	+0.75
Command-A	47.52	6.02	19.30	21.34	0.27	20.75	-85.44	1	+2.01
o4-mini	89.80	6.77	5.51	10.58	0.27	6.03	-79.14	1	+0.75
GPT-5.2	56.47	15.29	6.02	17.80	12.33	32.88	-6.38	2	+0.25
Gemini-3-Flash	93.85	15.29	17.29	27.99	1.77	24.38	-50.82	1	-0.51
Claude Haiku 4.5	88.03	32.83	25.81	50.70	12.80	62.80	-42.94	2	+2.51
Human (Base)	47.22	27.57	8.52	29.89	**82.35**	**100.00**	+52.21	1	-0.12
Human (Expert)	0.00	1.00	0.00	**0.81**	—	—	—	2	+0.37

Table 44. Sensitivity profile — Task Type: *B. Quantitative Reasoning and Optimization*.

Model	*Cond. Inv. Retention (%)*			*Collapse Contribution (%)*			*Rel. Sensitivity (pp)*		
	$TCIR_1$	$TCIR_2$	$TCIR_3$	$TCC_1 \downarrow$	$TCC_2 \downarrow$	$TCC_3 \downarrow$	RIS_1	RIS_2	RIS_3
Llama-3.2-1B-I	60.00	66.67	26.67	63.16	52.63	78.95	+0.56	+1.06	-1.94
Ministral-3B-2512	42.42	48.48	24.24	67.86	57.14	82.14	+0.25	+0.75	-1.25
Gemma-3-4B-I	40.54	48.65	35.14	75.86	65.52	72.41	-0.19	+0.56	-0.69
Qwen3-8B	88.27	79.77	82.99	37.50	81.25	34.38	+5.08	-2.19	+0.57
Gemma-4-26B-A4B-I	46.99	49.40	39.76	67.19	60.94	71.88	+0.12	+0.63	-1.38
Olmo-3.1-32B-I	42.00	44.00	38.00	71.05	68.42	73.68	-0.07	+0.18	-0.57
Aya-Expanse-32B	27.66	57.45	38.30	89.47	52.63	68.42	-1.75	+1.76	-0.50
Llama-3.3-70B-I	54.88	52.44	43.90	61.26	59.46	71.17	+2.51	+1.50	-2.00
Qwen3.6-Plus	90.57	87.14	91.43	75.00	78.57	32.14	-0.25	-3.26	+0.50
DeepSeek-V3.1	37.66	44.16	29.87	69.12	63.24	70.59	-0.06	+1.19	-1.57
Command-A	52.38	28.57	37.14	53.26	79.35	66.30	+4.26	-2.00	+0.25
o4-mini	44.64	57.14	33.93	74.36	56.41	82.05	-0.62	+1.13	-2.13
GPT-5.2	91.43	91.43	90.86	76.19	47.62	52.38	-0.25	-0.25	-0.75
Gemini-3-Flash	62.70	61.08	64.86	69.66	64.04	57.30	-0.07	-0.82	+0.94
Claude Haiku 4.5	71.60	65.98	65.68	59.54	64.12	63.36	+2.88	-1.88	-2.13
Human (Base)	89.72	88.47	89.97	46.67	80.00	46.67	-0.19	-1.44	+0.06
Human (Expert)	**100.00**	**99.75**	**99.25**	**0.00**	**0.00**	**0.00**	+0.25	+0.00	-0.50

Table 45. Sub-type accuracy breakdown for $\mathcal{I}_1$ — Task Type: *B. Quantitative Reasoning and Optimization*.

Model	*Accuracy (%) by Sub-Type — $\mathcal{I}_1$ (Symbolic & Linguistic)*			
	Entity Relabeling	Style / Register Invariance	Synonym / Semantic Paraphrase...	Variable Renaming
Llama-3.2-1B-I	1.96	5.41	0.00	9.92
Ministral-3B-2512	0.98	1.35	2.94	7.44
Gemma-3-4B-I	0.98	1.35	0.98	9.92
Qwen3-8B	72.55	74.32	77.45	76.86
Gemma-4-26B-A4B-I	3.92	5.41	12.75	14.88
Olmo-3.1-32B-I	1.96	6.76	0.98	10.74
Aya-Expanse-32B	0.98	4.05	1.96	5.79
Llama-3.3-70B-I	17.65	24.32	21.57	26.45
Qwen3.6-Plus	73.53	86.49	82.35	77.69
DeepSeek-V3.1	6.86	13.51	6.86	4.13
Command-A	15.69	9.46	8.82	19.01
o4-mini	6.86	2.70	4.90	9.09
GPT-5.2	71.57	86.49	88.24	76.86
Gemini-3-Flash	24.51	25.68	31.37	33.06
Claude Haiku 4.5	49.02	60.81	68.63	63.64
Human (Base)	86.27	93.24	89.22	90.91
Human (Expert)	**100.00**	**100.00**	**100.00**	**100.00**

Table 46. Sub-type accuracy breakdown for $\mathcal{I}_2$ — Task Type: *B. Quantitative Reasoning and Optimization*.

Model	*Accuracy (%) by Sub-Type — $\mathcal{I}_2$ (Structural & Representational)*			
	Equivalent Logical Reformulation	Equivalent Mathematical Reformulation	Premise Reordering	Representation / Encoding Swap
Llama-3.2-1B-I	0.00	12.61	2.22	2.06
Ministral-3B-2512	0.00	10.08	0.00	4.12
Gemma-3-4B-I	0.00	13.45	0.74	1.03
Qwen3-8B	77.08	61.34	60.00	83.51
Gemma-4-26B-A4B-I	6.25	15.97	7.41	9.28
Olmo-3.1-32B-I	0.00	11.76	0.74	7.22
Aya-Expanse-32B	4.17	13.45	2.22	6.19
Llama-3.3-70B-I	6.25	23.53	23.70	23.71
Qwen3.6-Plus	77.08	68.07	74.07	89.69
DeepSeek-V3.1	8.33	14.29	5.19	6.19
Command-A	2.08	11.76	5.19	8.25
o4-mini	4.17	15.13	3.70	7.22
GPT-5.2	91.67	69.75	77.78	90.72
Gemini-3-Flash	45.83	28.57	22.96	26.80
Claude Haiku 4.5	47.92	59.66	51.11	61.86
Human (Base)	87.50	87.39	88.89	89.69
Human (Expert)	**100.00**	**100.00**	**99.26**	**100.00**

Table 47. Sub-type accuracy breakdown for $\mathcal{I}_3$ — Task Type: *B. Quantitative Reasoning and Optimization*.

Model	*Accuracy (%) by Sub-Type — $\mathcal{I}_3$ (Abstraction & Info-Structure)*			
	Abstraction Level Shift	Distractor Injection / Removal	Perspective / Role Reframing	Redundant Restatement Invariance
Llama-3.2-1B-I	1.45	4.85	0.00	1.25
Ministral-3B-2512	2.90	1.94	1.28	1.25
Gemma-3-4B-I	4.35	3.88	1.28	2.50
Qwen3-8B	59.42	67.96	80.77	85.00
Gemma-4-26B-A4B-I	12.32	5.83	3.85	8.75
Olmo-3.1-32B-I	8.70	2.91	2.56	2.50
Aya-Expanse-32B	7.25	3.88	2.56	2.50
Llama-3.3-70B-I	20.29	15.53	17.95	17.50
Qwen3.6-Plus	69.57	75.73	92.31	92.50
DeepSeek-V3.1	6.52	8.74	1.28	5.00
Command-A	9.42	11.65	8.97	8.75
o4-mini	7.97	2.91	2.56	3.75
GPT-5.2	63.77	82.52	92.31	91.25
Gemini-3-Flash	28.26	31.07	28.21	33.75
Claude Haiku 4.5	49.28	54.37	58.97	65.00
Human (Base)	92.03	87.38	84.62	95.00
Human (Expert)	**99.28**	**99.03**	**100.00**	**98.75**

H.2.1. B1. SEQUENTIAL ALGEBRAIC SYSTEMS WITH LAYERED CONSTRAINTS

Table 48. Core accuracy — Sub-Type: *B1. Sequential Algebraic Systems with Layered Constraints*.

Model	Base	*Accuracy (%)* $\mathcal{I}_1$	$\mathcal{I}_2$	$\mathcal{I}_3$	*Mean Inv.* $\overline{\mathcal{I}}$	*Base*$-\overline{\mathcal{I}}$ Δ
Llama-3.2-1B-I	7.41	7.41	4.94	2.47	4.94	+2.47
Ministral-3B-2512	2.47	1.23	6.17	6.17	4.52	-2.05
Gemma-3-4B-I	4.94	3.70	6.17	1.23	3.70	+1.24
Qwen3-8B	65.43	74.07	67.90	65.43	69.13	-3.70
Gemma-4-26B-A4B-I	11.11	9.88	14.81	4.94	9.88	+1.23
Olmo-3.1-32B-I	8.64	6.17	8.64	3.70	6.17	+2.47
Aya-Expanse-32B	3.70	3.70	8.64	8.64	6.99	-3.29
Llama-3.3-70B-I	22.22	27.16	24.69	19.75	23.87	-1.65
Qwen3.6-Plus	87.65	80.25	80.25	82.72	81.07	+6.58
DeepSeek-V3.1	8.64	4.94	7.41	9.88	7.41	+1.23
Command-A	3.70	9.88	7.41	11.11	9.47	**-5.77**
o4-mini	7.41	4.94	11.11	3.70	6.58	+0.83
GPT-5.2	86.42	81.48	81.48	82.72	81.89	+4.53
Gemini-3-Flash	28.40	24.69	32.10	28.40	28.40	+0.00
Claude Haiku 4.5	58.02	56.79	60.49	53.09	56.79	+1.23
Human (Base)	95.06	92.59	91.36	87.65	90.53	+4.53
Human (Expert)	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	+0.00

Table 49. Invariance stability — Sub-Type: *B1. Sequential Algebraic Systems with Layered Constraints*.

Model	*Global Stability (%)* IMI↑	CIR↑	ICR↓	*Novel (%)* BSS↑	CE↓	*Pairwise Failure Dep. (%)* $P(\mathcal{I}_1{=}0\|\mathcal{I}_2{=}0)$	$P(\mathcal{I}_1{=}0\|\mathcal{I}_3{=}0)$	$P(\mathcal{I}_2{=}0\|\mathcal{I}_3{=}0)$
Llama-3.2-1B-I	2.47	33.33	2.47	0.00	4.47	97.40	94.94	97.47
Ministral-3B-2512	1.23	50.00	1.23	25.00	8.25	100.00	100.00	96.05
Gemma-3-4B-I	0.00	0.00	4.94	80.00	10.71	96.05	96.25	93.75
Qwen3-8B	40.74	62.26	3.70	88.89	38.15	65.38	50.00	50.00
Gemma-4-26B-A4B-I	1.23	11.11	3.70	93.33	15.96	97.10	90.91	88.31
Olmo-3.1-32B-I	1.23	14.29	3.70	90.91	11.25	97.30	96.15	92.31
Aya-Expanse-32B	1.23	33.33	1.23	36.36	11.72	100.00	97.30	94.59
Llama-3.3-70B-I	7.41	33.33	8.64	80.00	32.38	81.97	81.54	84.62
Qwen3.6-Plus	66.67	76.06	2.47	21.05	19.73	62.50	64.29	57.14
DeepSeek-V3.1	0.00	0.00	6.17	87.50	17.19	97.33	95.89	94.52
Command-A	1.23	33.33	1.23	25.00	16.96	92.00	94.44	94.44
o4-mini	2.47	33.33	4.94	80.00	10.48	98.61	98.72	92.31
GPT-5.2	66.67	77.14	1.23	31.58	19.50	60.00	64.29	57.14
Gemini-3-Flash	14.81	52.17	3.70	**95.65**	24.20	90.91	94.83	81.03
Claude Haiku 4.5	24.69	42.55	12.35	94.12	53.41	65.62	63.16	47.37
Human (Base)	69.14	72.73	1.23	32.00	25.04	14.29	10.00	10.00
Human (Expert)	**100.00**	**100.00**	**0.00**	—	**0.00**	0.00	0.00	0.00

Table 50. Novel metrics — Sub-Type: *B1. Sequential Algebraic Systems with Layered Constraints*.

Model	*Bilateral Symmetry* BSS↑	$F^+(\%)$	$F^-(\%)$	*Entropy* CE↓	*Invariant Recovery (%)* IRR_{full} ↑	IRR_{part} ↑	IRR_{Δ}	*Fragility Gradient* FG	$\overline{g}$
Llama-3.2-1B-I	0.00	4.94	0.00	4.47	0.00	0.00	-66.67	**2**	+2.47
Ministral-3B-2512	25.00	1.23	8.64	8.25	0.00	8.86	-50.00	1	-2.47
Gemma-3-4B-I	80.00	4.94	7.41	10.71	0.00	7.79	-100.00	1	+1.23
Qwen3-8B	88.89	24.69	19.75	38.15	17.86	57.14	-19.88	2	+4.32
Gemma-4-26B-A4B-I	93.33	9.88	8.64	15.96	0.00	9.72	-88.89	1	+2.47
Olmo-3.1-32B-I	90.91	7.41	6.17	11.25	0.00	6.76	-85.71	1	+1.23
Aya-Expanse-32B	36.36	2.47	11.11	11.72	0.00	11.54	-66.67	1	-2.47
Llama-3.3-70B-I	80.00	14.81	22.22	32.38	1.59	28.57	-65.08	2	+3.70
Qwen3.6-Plus	21.05	20.99	2.47	19.73	0.00	20.00	-23.94	1	-1.23
DeepSeek-V3.1	87.50	8.64	11.11	17.19	0.00	12.16	-100.00	0	-2.47
Command-A	25.00	2.47	17.28	16.96	0.00	17.95	-66.67	1	-0.61
o4-mini	80.00	4.94	7.41	10.48	1.33	8.00	-65.34	1	+0.62
GPT-5.2	31.58	19.75	3.70	19.50	9.09	27.27	-13.77	1	-0.62
Gemini-3-Flash	**95.65**	13.58	14.81	24.20	0.00	20.69	-47.83	1	-1.86
Claude Haiku 4.5	94.12	33.33	29.63	53.41	14.71	70.59	-42.74	1	+1.85
Human (Base)	32.00	25.93	4.94	25.04	**100.00**	**100.00**	+72.73	2	+2.47
Human (Expert)	—	0.00	0.00	**0.00**	—	—	—	2	+0.00

Table 51. Sensitivity profile — Sub-Type: *B1. Sequential Algebraic Systems with Layered Constraints*.

Model	*Cond. Inv. Retention (%)* TCIR_1	TCIR_2	TCIR_3	*Collapse Contribution (%)* TCC_1 ↓	TCC_2 ↓	TCC_3 ↓	*Rel. Sensitivity (pp)* RIS_1	RIS_2	RIS_3
Llama-3.2-1B-I	**100.00**	66.67	33.33	**0.00**	100.00	100.00	+1.85	-0.62	-3.09
Ministral-3B-2512	11.11	55.56	55.56	100.00	50.00	50.00	-2.78	+2.16	+2.16
Gemma-3-4B-I	30.00	50.00	10.00	70.00	50.00	90.00	-0.31	+2.16	-2.78
Qwen3-8B	86.96	79.71	76.81	42.86	78.57	35.71	+5.86	-0.31	-2.78
Gemma-4-26B-A4B-I	50.00	75.00	25.00	70.00	40.00	80.00	-0.31	+4.62	-5.25
Olmo-3.1-32B-I	41.67	58.33	25.00	87.50	50.00	87.50	-0.62	+1.85	-3.09
Aya-Expanse-32B	25.00	58.33	58.33	90.00	50.00	40.00	-2.47	+2.47	+2.47
Llama-3.3-70B-I	61.11	55.56	44.44	54.17	62.50	70.83	+3.70	+1.23	-3.71
Qwen3.6-Plus	89.04	89.04	91.78	100.00	50.00	25.00	-2.47	-2.47	+0.00
DeepSeek-V3.1	25.00	37.50	50.00	85.71	71.43	42.86	-2.78	-0.31	+2.16
Command-A	47.06	35.29	52.94	60.00	73.33	46.67	+1.86	-0.61	+3.09
o4-mini	33.33	75.00	25.00	88.89	33.33	100.00	-1.85	+4.32	-3.09
GPT-5.2	90.41	90.41	91.78	100.00	33.33	33.33	-1.54	-1.54	-0.30
Gemini-3-Flash	57.14	74.29	65.71	86.67	33.33	66.67	-3.71	+3.70	+0.00
Claude Haiku 4.5	64.79	69.01	60.56	72.41	58.62	68.97	-0.31	+3.39	-4.01
Human (Base)	92.59	91.36	87.65	100.00	100.00	100.00	+0.92	-0.31	-4.02
Human (Expert)	100.00	**100.00**	**100.00**	0.00	**0.00**	**0.00**	+0.00	+0.00	+0.00

H.2.2. B2. OPTIMIZATION OVER COUPLED ALGEBRAIC EQUATIONS

Table 52. Core accuracy — Sub-Type: *B2. Optimization over Coupled Algebraic Equations*.

Model	Base	*Accuracy (%)* $\mathcal{I}_1$	$\mathcal{I}_2$	$\mathcal{I}_3$	*Mean Inv.* $\overline{\mathcal{I}}$	*Base*$-\overline{\mathcal{I}}$ Δ
Llama-3.2-1B-I	4.94	6.17	4.94	0.00	3.70	+1.24
Ministral-3B-2512	2.47	4.94	3.70	2.47	3.70	-1.23
Gemma-3-4B-I	6.17	8.64	3.70	1.23	4.52	+1.65
Qwen3-8B	69.14	79.01	70.37	75.31	74.90	-5.76
Gemma-4-26B-A4B-I	13.58	12.35	8.64	12.35	11.11	+2.47
Olmo-3.1-32B-I	4.94	8.64	2.47	6.17	5.76	-0.82
Aya-Expanse-32B	6.17	1.23	3.70	1.23	2.05	+4.12
Llama-3.3-70B-I	19.75	35.80	25.93	23.46	28.40	**-8.65**
Qwen3.6-Plus	83.95	83.95	76.54	85.19	81.89	+2.06
DeepSeek-V3.1	6.17	7.41	7.41	4.94	6.59	-0.42
Command-A	6.17	19.75	7.41	13.58	13.58	-7.41
o4-mini	12.35	6.17	6.17	7.41	6.58	+5.77
GPT-5.2	81.48	81.48	75.31	76.54	77.78	+3.70
Gemini-3-Flash	28.40	23.46	23.46	32.10	26.34	+2.06
Claude Haiku 4.5	61.73	66.67	51.85	58.02	58.85	+2.88
Human (Base)	91.36	88.89	86.42	93.83	89.71	+1.65
Human (Expert)	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	+0.00

Table 53. Invariance stability — Sub-Type: *B2. Optimization over Coupled Algebraic Equations*.

Model	*Global Stability (%)* IMI↑	CIR↑	ICR↓	*Novel (%)* BSS↑	CE↓	*Pairwise Failure Dep. (%)* $P(\mathcal{I}_1{=}0\|\mathcal{I}_2{=}0)$	$P(\mathcal{I}_1{=}0\|\mathcal{I}_3{=}0)$	$P(\mathcal{I}_2{=}0\|\mathcal{I}_3{=}0)$
Llama-3.2-1B-I	0.00	0.00	1.23	66.67	6.24	97.40	93.83	95.06
Ministral-3B-2512	0.00	0.00	1.23	57.14	7.48	96.15	96.20	97.47
Gemma-3-4B-I	1.23	20.00	2.47	85.71	7.24	94.87	92.50	97.50
Qwen3-8B	51.85	75.00	1.23	**100.00**	29.91	58.33	60.00	60.00
Gemma-4-26B-A4B-I	4.94	36.36	6.17	92.31	14.19	91.89	95.77	97.18
Olmo-3.1-32B-I	1.23	25.00	3.70	60.00	10.95	92.41	93.42	100.00
Aya-Expanse-32B	0.00	0.00	6.17	75.00	8.48	98.72	98.75	96.25
Llama-3.3-70B-I	3.70	18.75	4.94	65.00	41.69	71.67	70.97	83.87
Qwen3.6-Plus	74.07	88.24	**0.00**	54.55	11.02	68.42	83.33	83.33
DeepSeek-V3.1	0.00	0.00	6.17	71.43	15.19	94.67	93.51	94.81
Command-A	1.23	20.00	3.70	30.77	26.97	82.67	81.43	95.71
o4-mini	1.23	10.00	8.64	61.54	14.19	96.05	94.67	97.33
GPT-5.2	66.67	81.82	1.23	28.57	14.26	70.00	73.68	68.42
Gemini-3-Flash	16.05	56.52	6.17	95.24	21.97	91.94	96.36	92.73
Claude Haiku 4.5	27.16	44.00	9.88	75.56	47.87	51.28	50.00	70.59
Human (Base)	66.67	72.97	2.47	51.85	28.21	18.18	0.00	0.00
Human (Expert)	**100.00**	**100.00**	0.00	—	**0.00**	0.00	0.00	0.00

Table 54. Novel metrics — Sub-Type: *B2. Optimization over Coupled Algebraic Equations*.

	Bilateral Symmetry			*Entropy*	*Invariant Recovery (%)*			*Fragility Gradient*	
Model	BSS↑	$F^{+}(\%)$	$F^{-}(\%)$	CE↓	IRR_{full} ↑	IRR_{part} ↑	IRR_{Δ}	FG	$\overline{g}$
Llama-3.2-1B-I	66.67	4.94	2.47	6.24	0.00	2.60	-100.00	**2**	+3.08
Ministral-3B-2512	57.14	2.47	6.17	7.48	0.00	6.33	-100.00	2	+1.23
Gemma-3-4B-I	85.71	4.94	3.70	7.24	0.00	3.95	-80.00	2	+3.70
Qwen3-8B	**100.00**	17.28	17.28	29.91	16.00	56.00	-9.00	1	+1.85
Gemma-4-26B-A4B-I	92.31	8.64	7.41	14.19	0.00	8.57	-63.64	1	+0.00
Olmo-3.1-32B-I	60.00	3.70	8.64	10.95	0.00	9.09	-75.00	1	+1.23
Aya-Expanse-32B	75.00	6.17	3.70	8.48	0.00	3.95	-100.00	1	+0.00
Llama-3.3-70B-I	65.00	16.05	33.33	41.69	6.15	41.54	-75.10	2	+6.17
Qwen3.6-Plus	54.55	9.88	3.70	11.02	0.00	23.08	-11.76	1	-0.62
DeepSeek-V3.1	71.43	6.17	11.11	15.19	1.32	11.84	-98.68	2	+1.23
Command-A	30.77	4.94	27.16	26.97	1.32	28.95	-78.68	1	+3.08
o4-mini	61.54	11.11	4.94	14.19	0.00	5.63	-90.00	1	-0.62
GPT-5.2	28.57	14.81	2.47	14.26	6.67	13.33	-11.51	1	+2.47
Gemini-3-Flash	95.24	12.35	13.58	21.97	0.00	18.97	-43.48	1	-4.32
Claude Haiku 4.5	75.56	34.57	20.99	47.87	12.90	54.84	-43.10	1	+4.32
Human (Base)	51.85	24.69	8.64	28.21	**57.14**	**100.00**	+30.11	1	-2.47
Human (Expert)	—	0.00	0.00	**0.00**	—	—	—	2	+0.00

Table 55. Sensitivity profile — Sub-Type: *B2. Optimization over Coupled Algebraic Equations*.

	Cond. Inv. Retention (%)			*Collapse Contribution (%)*			*Rel. Sensitivity (pp)*		
Model	TCIR_1	TCIR_2	TCIR_3	$\text{TCC}_1 \downarrow$	$\text{TCC}_2 \downarrow$	$\text{TCC}_3 \downarrow$	RIS_1	RIS_2	RIS_3
Llama-3.2-1B-I	83.33	66.67	0.00	33.33	66.67	100.00	+2.16	+0.93	-4.01
Ministral-3B-2512	57.14	42.86	28.57	50.00	50.00	83.33	+1.54	+0.30	-0.93
Gemma-3-4B-I	87.50	37.50	12.50	20.00	100.00	100.00	+3.70	-1.24	-3.71
Qwen3-8B	91.43	81.43	87.14	27.27	72.73	36.36	+5.55	-3.09	+1.85
Gemma-4-26B-A4B-I	58.82	41.18	58.82	63.64	72.73	63.64	+0.62	-3.09	+0.62
Olmo-3.1-32B-I	63.64	18.18	45.45	40.00	90.00	60.00	+3.08	-3.09	+0.61
Aya-Expanse-32B	12.50	37.50	12.50	87.50	62.50	87.50	-1.86	+0.61	-1.86
Llama-3.3-70B-I	67.44	48.84	44.19	48.15	70.37	70.37	+9.57	-0.30	-2.77
Qwen3.6-Plus	95.77	87.32	97.18	100.00	100.00	**0.00**	+1.54	-5.87	+2.78
DeepSeek-V3.1	42.86	42.86	28.57	61.54	61.54	76.92	+0.93	+0.93	-1.54
Command-A	59.26	22.22	40.74	45.83	87.50	62.50	+8.02	-4.32	+1.85
o4-mini	35.71	35.71	42.86	72.73	81.82	63.64	-1.85	-1.85	-0.61
GPT-5.2	97.06	89.71	91.18	100.00	50.00	50.00	+2.78	-3.39	-2.16
Gemini-3-Flash	55.88	55.88	76.47	81.25	75.00	50.00	-3.39	-3.39	+5.25
Claude Haiku 4.5	80.60	62.69	70.15	38.10	80.95	61.90	+7.10	-7.72	-1.55
Human (Base)	88.89	86.42	93.83	40.00	80.00	20.00	-1.23	-3.70	+3.71
Human (Expert)	**100.00**	**100.00**	**100.00**	**0.00**	**0.00**	0.00	+0.00	+0.00	+0.00

H.2.3. B3. PROBABILISTIC REASONING WITH CONDITIONAL DEPENDENCE

Table 56. Core accuracy — Sub-Type: *B3. Probabilistic Reasoning with Conditional Dependence*.

Model	Base	*Accuracy (%)* $\mathcal{I}_1$	$\mathcal{I}_2$	$\mathcal{I}_3$	*Mean Inv.* $\overline{\mathcal{I}}$	*Base*$-\overline{\mathcal{I}}$ Δ
Llama-3.2-1B-I	0.00	0.00	4.84	3.23	2.69	-2.69
Ministral-3B-2512	3.23	3.23	4.84	1.61	3.23	+0.00
Gemma-3-4B-I	3.23	0.00	1.61	4.84	2.15	+1.08
Qwen3-8B	77.42	74.19	80.65	82.26	79.03	-1.61
Gemma-4-26B-A4B-I	12.90	6.45	11.29	9.68	9.14	+3.76
Olmo-3.1-32B-I	6.45	1.61	4.84	3.23	3.23	+3.22
Aya-Expanse-32B	0.00	1.61	6.45	4.84	4.30	-4.30
Llama-3.3-70B-I	11.29	11.29	12.90	4.84	9.68	+1.61
Qwen3.6-Plus	75.81	66.13	77.42	83.87	75.81	+0.00
DeepSeek-V3.1	14.52	12.90	12.90	9.68	11.83	+2.69
Command-A	8.06	8.06	8.06	8.06	8.06	+0.00
o4-mini	9.68	4.84	3.23	4.84	4.30	+5.38
GPT-5.2	80.65	79.03	87.10	87.10	84.41	-3.76
Gemini-3-Flash	50.00	53.23	33.87	41.94	43.01	+6.99
Claude Haiku 4.5	46.77	56.45	61.29	62.90	60.21	**-13.44**
Human (Base)	83.87	91.94	91.94	91.94	91.94	-8.07
Human (Expert)	**100.00**	**100.00**	**98.39**	**98.39**	**98.93**	+1.07

Table 57. Invariance stability — Sub-Type: *B3. Probabilistic Reasoning with Conditional Dependence*.

Model	*Global Stability (%)* IMI↑	CIR↑	ICR↓	*Novel (%)* BSS↑	CE↓	*Pairwise Failure Dep. (%)* $P(\mathcal{I}_1{=}0\|\mathcal{I}_2{=}0)$	$P(\mathcal{I}_1{=}0\|\mathcal{I}_3{=}0)$	$P(\mathcal{I}_2{=}0\|\mathcal{I}_3{=}0)$
Llama-3.2-1B-I	0.00	0.00	**0.00**	0.00	5.54	100.00	100.00	96.67
Ministral-3B-2512	1.61	50.00	1.61	66.67	4.23	98.31	98.36	96.72
Gemma-3-4B-I	0.00	0.00	3.23	66.67	7.85	100.00	100.00	98.31
Qwen3-8B	58.06	75.00	3.23	80.00	27.69	66.67	63.64	63.64
Gemma-4-26B-A4B-I	3.23	25.00	9.68	85.71	19.23	98.18	96.43	91.07
Olmo-3.1-32B-I	1.61	25.00	4.84	**100.00**	7.85	100.00	100.00	96.67
Aya-Expanse-32B	0.00	0.00	0.00	0.00	7.85	100.00	100.00	94.92
Llama-3.3-70B-I	1.61	14.29	9.68	66.67	24.47	88.89	89.83	88.14
Qwen3.6-Plus	64.52	85.11	3.23	100.00	19.54	92.86	100.00	90.00
DeepSeek-V3.1	1.61	11.11	11.29	80.00	27.69	88.89	87.50	89.29
Command-A	1.61	20.00	6.45	61.54	17.92	96.49	92.98	92.98
o4-mini	1.61	16.67	8.06	57.14	10.07	96.67	98.31	100.00
GPT-5.2	69.35	86.00	4.84	93.33	21.45	62.50	75.00	75.00
Gemini-3-Flash	16.13	32.26	17.74	80.00	50.06	58.54	66.67	75.00
Claude Haiku 4.5	24.19	51.72	9.68	75.68	52.68	62.50	65.22	60.87
Human (Base)	64.52	76.92	3.23	90.91	29.70	40.00	0.00	0.00
Human (Expert)	**96.77**	**96.77**	0.00	0.00	**2.62**	0.00	0.00	0.00

Table 58. Novel metrics — Sub-Type: *B3. Probabilistic Reasoning with Conditional Dependence*.

Model	*Bilateral Symmetry*			*Entropy*	*Invariant Recovery (%)*			*Fragility Gradient*	
	BSS↑	F^{+}(%)	F^{-}(%)	CE↓	IRR_{full} ↑	IRR_{part} ↑	IRR_{Δ}	FG	$\overline{g}$
Llama-3.2-1B-I	0.00	0.00	6.45	5.54	0.00	6.45	-100.00	1	-1.61
Ministral-3B-2512	66.67	1.61	3.23	4.23	0.00	3.33	-50.00	1	+0.81
Gemma-3-4B-I	66.67	3.23	6.45	7.85	0.00	6.67	-100.00	0	-2.42
Qwen3-8B	80.00	19.35	12.90	27.69	21.43	57.14	-3.57	0	-4.04
Gemma-4-26B-A4B-I	85.71	9.68	12.90	19.23	0.00	14.81	-75.00	1	-1.61
Olmo-3.1-32B-I	**100.00**	4.84	4.84	7.85	0.00	5.17	-75.00	1	-0.81
Aya-Expanse-32B	0.00	0.00	9.68	7.85	1.61	9.68	-98.39	1	-1.61
Llama-3.3-70B-I	66.67	9.68	19.35	24.47	0.00	21.82	-85.71	1	+3.23
Qwen3.6-Plus	100.00	11.29	11.29	19.54	0.00	46.67	-14.89	0	-8.87
DeepSeek-V3.1	80.00	12.90	19.35	27.69	0.00	22.64	-88.89	**2**	+1.61
Command-A	61.54	6.45	14.52	17.92	0.00	15.79	-80.00	2	+0.00
o4-mini	57.14	8.06	3.23	10.07	0.00	3.57	-83.33	1	+0.00
GPT-5.2	93.33	11.29	12.90	21.45	25.00	66.67	+11.00	1	-4.03
Gemini-3-Flash	80.00	33.87	22.58	50.06	3.23	45.16	-64.51	1	+5.65
Claude Haiku 4.5	75.68	22.58	37.10	52.68	21.21	69.70	-27.07	0	-3.23
Human (Base)	90.91	19.35	16.13	29.70	**90.00**	**100.00**	+66.92	2	+0.00
Human (Expert)	0.00	3.23	0.00	**2.62**	—	—	—	2	+0.80

Table 59. Sensitivity profile — Sub-Type: *B3. Probabilistic Reasoning with Conditional Dependence*.

Model	*Cond. Inv. Retention (%)*			*Collapse Contribution (%)*			*Rel. Sensitivity (pp)*		
	TCIR_1	TCIR_2	TCIR_3	TCC_1 ↓	TCC_2 ↓	TCC_3 ↓	RIS_1	RIS_2	RIS_3
Llama-3.2-1B-I	0.00	75.00	50.00	100.00	25.00	50.00	-2.02	+2.82	+1.21
Ministral-3B-2512	50.00	75.00	25.00	66.67	33.33	100.00	+0.00	+1.61	-1.62
Gemma-3-4B-I	0.00	16.67	50.00	100.00	83.33	50.00	-2.42	-0.81	+2.42
Qwen3-8B	82.14	89.29	91.07	57.14	71.43	28.57	-4.44	+2.02	+3.63
Gemma-4-26B-A4B-I	25.00	43.75	37.50	85.71	64.29	71.43	-3.63	+1.21	-0.40
Olmo-3.1-32B-I	14.29	42.86	28.57	100.00	66.67	83.33	-2.42	+0.81	-0.80
Aya-Expanse-32B	16.67	66.67	50.00	100.00	40.00	60.00	-1.62	+3.22	+1.61
Llama-3.3-70B-I	36.84	42.11	15.79	66.67	61.11	88.89	+1.21	+2.82	-5.24
Qwen3.6-Plus	75.93	88.89	96.30	100.00	55.56	22.22	-9.68	+1.61	+8.06
DeepSeek-V3.1	38.10	38.10	28.57	68.42	68.42	73.68	+0.40	+0.40	-2.82
Command-A	35.71	35.71	35.71	69.23	69.23	69.23	+-0.00	+-0.00	+-0.00
o4-mini	37.50	25.00	37.50	71.43	85.71	71.43	-0.81	-2.42	-0.81
GPT-5.2	84.48	93.10	93.10	75.00	37.50	50.00	-4.44	+3.63	+3.63
Gemini-3-Flash	73.33	46.67	57.78	50.00	75.00	62.50	+8.47	-10.89	-2.82
Claude Haiku 4.5	67.31	73.08	75.00	59.09	50.00	54.55	-0.40	+4.44	+6.05
Human (Base)	91.94	91.94	91.94	66.67	100.00	**0.00**	+2.02	+2.02	+2.02
Human (Expert)	**100.00**	**98.39**	**98.39**	**0.00**	**0.00**	0.00	+0.81	-0.80	-0.80

H.2.4. B4. INTERDEPENDENT EQUATION SYSTEMS WITH STRUCTURAL CONSTRAINTS

Table 60. Core accuracy — Sub-Type: *B4. Interdependent Equation Systems with Structural Constraints*.

Model	Base	*Accuracy (%)* $\mathcal{I}_1$	$\mathcal{I}_2$	$\mathcal{I}_3$	*Mean Inv.* $\overline{\mathcal{I}}$	*Base*$-\overline{\mathcal{I}}$ Δ
Llama-3.2-1B-I	3.41	3.41	7.95	3.41	4.92	-1.51
Ministral-3B-2512	5.68	5.68	2.27	0.00	2.65	+3.03
Gemma-3-4B-I	3.41	3.41	6.82	4.55	4.93	-1.52
Qwen3-8B	61.36	69.32	60.23	61.36	63.64	-2.28
Gemma-4-26B-A4B-I	10.23	14.77	7.95	10.23	10.98	-0.75
Olmo-3.1-32B-I	3.41	3.41	5.68	7.95	5.68	-2.27
Aya-Expanse-32B	9.09	3.41	10.23	3.41	5.68	+3.41
Llama-3.3-70B-I	11.36	13.64	14.77	20.45	16.29	**-4.93**
Qwen3.6-Plus	79.55	77.27	70.45	73.86	73.86	+5.69
DeepSeek-V3.1	5.68	9.09	11.36	3.41	7.95	-2.27
Command-A	9.09	13.64	6.82	5.68	8.71	+0.38
o4-mini	7.95	6.82	11.36	5.68	7.95	-0.00
GPT-5.2	78.41	78.41	73.86	75.00	75.76	+2.65
Gemini-3-Flash	23.86	25.00	27.27	31.82	28.03	-4.17
Claude Haiku 4.5	60.23	64.77	47.73	50.00	54.17	+6.06
Human (Base)	94.32	88.64	85.23	87.50	87.12	+7.20
Human (Expert)	**100.00**	**100.00**	**100.00**	**98.86**	**99.62**	+0.38

Table 61. Invariance stability — Sub-Type: *B4. Interdependent Equation Systems with Structural Constraints*.

Model	*Global Stability (%)* IMI↑	CIR↑	ICR↓	*Novel (%)* BSS↑	CE↓	*Pairwise Failure Dep. (%)* $P(\mathcal{I}_1{=}0\|\mathcal{I}_2{=}0)$	$P(\mathcal{I}_1{=}0\|\mathcal{I}_3{=}0)$	$P(\mathcal{I}_2{=}0\|\mathcal{I}_3{=}0)$
Llama-3.2-1B-I	1.14	33.33	1.14	50.00	7.80	98.77	97.65	94.12
Ministral-3B-2512	0.00	0.00	4.55	88.89	8.51	95.35	94.32	97.73
Gemma-3-4B-I	2.27	66.67	**0.00**	33.33	5.53	100.00	98.81	95.24
Qwen3-8B	38.64	62.96	5.68	88.89	35.33	68.57	58.82	64.71
Gemma-4-26B-A4B-I	2.27	22.22	6.82	73.68	19.45	88.89	89.87	93.67
Olmo-3.1-32B-I	0.00	0.00	0.00	50.00	11.06	96.39	97.53	96.30
Aya-Expanse-32B	2.27	25.00	5.68	90.91	10.57	100.00	98.82	91.76
Llama-3.3-70B-I	0.00	0.00	4.55	66.67	30.02	85.33	92.86	88.57
Qwen3.6-Plus	61.36	77.14	3.41	22.22	17.67	69.23	69.57	73.91
DeepSeek-V3.1	0.00	0.00	2.27	55.56	17.02	93.59	91.76	88.24
Command-A	0.00	0.00	5.68	72.73	20.93	86.59	89.16	92.77
o4-mini	1.14	14.29	2.27	**92.31**	12.63	94.87	97.59	90.36
GPT-5.2	65.91	84.06	0.00	30.77	11.98	78.26	77.27	77.27
Gemini-3-Flash	13.64	57.14	2.27	78.26	22.49	93.75	91.67	90.00
Claude Haiku 4.5	23.86	39.62	12.50	76.92	51.37	50.00	52.27	59.09
Human (Base)	61.36	65.06	3.41	29.41	31.99	7.69	18.18	18.18
Human (Expert)	**98.86**	**98.86**	0.00	0.00	**0.92**	0.00	0.00	0.00

Table 62. Novel metrics — Sub-Type: *B4. Interdependent Equation Systems with Structural Constraints*.

Model	*Bilateral Symmetry* BSS↑	$F^+(\%)$	$F^-(\%)$	*Entropy* CE↓	*Invariant Recovery (%)* IRR_{full} ↑	IRR_{part} ↑	IRR_{Δ}	*Fragility Gradient* FG	$\overline{g}$
Llama-3.2-1B-I	50.00	2.27	6.82	7.80	0.00	7.06	-66.67	1	+0.00
Ministral-3B-2512	88.89	5.68	4.55	8.51	0.00	4.82	-100.00	**2**	+2.84
Gemma-3-4B-I	33.33	1.14	5.68	5.53	0.00	5.88	-33.33	1	-0.57
Qwen3-8B	88.89	22.73	18.18	35.33	14.71	47.06	-22.33	1	+3.98
Gemma-4-26B-A4B-I	73.68	7.95	13.64	19.45	0.00	15.19	-77.78	1	+2.27
Olmo-3.1-32B-I	50.00	3.41	10.23	11.06	0.00	10.59	-100.00	0	-2.27
Aya-Expanse-32B	90.91	6.82	5.68	10.57	0.00	6.25	-75.00	1	+0.00
Llama-3.3-70B-I	66.67	11.36	22.73	30.02	0.00	25.64	-100.00	0	-3.41
Qwen3.6-Plus	22.22	18.18	2.27	17.67	0.00	11.11	-22.86	1	+1.70
DeepSeek-V3.1	55.56	5.68	14.77	17.02	0.00	15.66	-100.00	1	+2.84
Command-A	72.73	9.09	15.91	20.93	0.00	17.50	-100.00	2	+3.98
o4-mini	**92.31**	6.82	7.95	12.63	0.00	8.64	-85.71	1	+0.57
GPT-5.2	30.77	12.50	2.27	11.98	5.26	10.53	-10.68	1	+1.70
Gemini-3-Flash	78.26	10.23	15.91	22.49	2.99	20.90	-39.87	0	-3.41
Claude Haiku 4.5	76.92	36.36	22.73	51.37	2.86	57.14	-57.52	1	+7.38
Human (Base)	29.41	32.95	5.68	31.99	**80.00**	**100.00**	+45.06	1	+0.57
Human (Expert)	0.00	1.14	0.00	**0.92**	—	—	—	2	+0.57

Table 63. Sensitivity profile — Sub-Type: *B4. Interdependent Equation Systems with Structural Constraints*.

Model	*Cond. Inv. Retention (%)* TCIR_1	TCIR_2	TCIR_3	*Collapse Contribution (%)* TCC_1 ↓	TCC_2 ↓	TCC_3 ↓	*Rel. Sensitivity (pp)* RIS_1	RIS_2	RIS_3
Llama-3.2-1B-I	33.33	77.78	33.33	85.71	28.57	71.43	-1.14	+3.40	-1.14
Ministral-3B-2512	55.56	22.22	0.00	50.00	87.50	100.00	+2.27	-1.14	-3.41
Gemma-3-4B-I	37.50	75.00	50.00	100.00	40.00	60.00	-1.14	+2.27	+0.00
Qwen3-8B	87.14	75.71	77.14	43.75	87.50	37.50	+6.25	-2.84	-1.71
Gemma-4-26B-A4B-I	61.90	33.33	42.86	44.44	72.22	66.67	+3.97	-2.85	-0.57
Olmo-3.1-32B-I	25.00	41.67	58.33	77.78	66.67	55.56	-1.70	+0.57	+2.84
Aya-Expanse-32B	23.08	69.23	23.08	100.00	40.00	90.00	-3.12	+3.70	-3.12
Llama-3.3-70B-I	40.00	43.33	60.00	70.83	54.17	45.83	-1.42	-0.29	+5.39
Qwen3.6-Plus	94.44	86.11	90.28	40.00	100.00	20.00	+1.99	-4.83	-1.42
DeepSeek-V3.1	44.44	55.56	16.67	66.67	53.33	80.00	+1.70	+3.97	-3.98
Command-A	54.55	27.27	22.73	52.63	73.68	84.21	+4.83	-1.99	-3.13
o4-mini	42.86	71.43	35.71	77.78	22.22	88.89	-1.13	+3.41	-2.27
GPT-5.2	97.18	91.55	92.96	100.00	100.00	**0.00**	+1.99	-2.56	-1.42
Gemini-3-Flash	62.86	68.57	80.00	71.43	71.43	28.57	-1.99	+0.28	+4.83
Claude Haiku 4.5	78.08	57.53	60.27	50.00	70.00	63.33	+9.09	-7.95	-5.68
Human (Base)	88.64	85.23	87.50	50.00	75.00	75.00	-0.28	-3.69	-1.42
Human (Expert)	**100.00**	**100.00**	**98.86**	**0.00**	**0.00**	0.00	+0.28	+0.28	-0.86

H.2.5. B5. CONSTRAINT-DRIVEN ALGEBRAIC REASONING PROBLEMS

Table 64. Core accuracy — Sub-Type: *B5. Constraint-Driven Algebraic Reasoning Problems*.

Model	Base	*Accuracy (%)* $\mathcal{I}_1$	$\mathcal{I}_2$	$\mathcal{I}_3$	*Mean Inv.* $\overline{\mathcal{I}}$	*Base*$-\overline{\mathcal{I}}$ Δ
Llama-3.2-1B-I	4.60	4.60	2.30	1.15	2.68	+1.92
Ministral-3B-2512	3.45	2.30	3.45	0.00	1.92	+1.53
Gemma-3-4B-I	3.45	2.30	3.45	4.60	3.45	+0.00
Qwen3-8B	64.37	80.46	65.52	73.56	73.18	**-8.81**
Gemma-4-26B-A4B-I	4.60	4.60	9.20	4.60	6.13	-1.53
Olmo-3.1-32B-I	5.75	5.75	5.75	2.30	4.60	+1.15
Aya-Expanse-32B	6.90	5.75	4.60	4.60	4.98	+1.92
Llama-3.3-70B-I	24.14	22.99	27.59	18.39	22.99	+1.15
Qwen3.6-Plus	85.06	86.21	78.16	77.01	80.46	+4.60
DeepSeek-V3.1	5.75	3.45	4.60	2.30	3.45	+2.30
Command-A	8.05	16.09	8.05	10.34	11.49	-3.44
o4-mini	5.75	8.05	6.90	2.30	5.75	+0.00
GPT-5.2	81.61	80.46	85.06	79.31	81.61	+0.00
Gemini-3-Flash	20.69	25.29	26.44	19.54	23.76	-3.07
Claude Haiku 4.5	64.37	57.47	59.77	56.32	57.85	+6.52
Human (Base)	90.80	87.36	88.51	89.66	88.51	+2.29
Human (Expert)	**100.00**	**100.00**	**100.00**	**98.85**	**99.62**	+0.38

Table 65. Invariance stability — Sub-Type: *B5. Constraint-Driven Algebraic Reasoning Problems*.

Model	*Global Stability (%)* IMI↑	CIR↑	ICR↓	*Novel (%)* BSS↑	CE↓	*Pairwise Failure Dep. (%)* $P(\mathcal{I}_1{=}0\|\mathcal{I}_2{=}0)$	$P(\mathcal{I}_1{=}0\|\mathcal{I}_3{=}0)$	$P(\mathcal{I}_2{=}0\|\mathcal{I}_3{=}0)$
Llama-3.2-1B-I	1.15	25.00	2.30	50.00	3.95	97.65	96.51	98.84
Ministral-3B-2512	0.00	0.00	2.30	50.00	4.16	98.81	97.70	96.55
Gemma-3-4B-I	2.30	66.67	1.15	66.67	3.01	100.00	100.00	98.80
Qwen3-8B	45.98	71.43	2.30	88.89	35.96	46.67	52.17	60.87
Gemma-4-26B-A4B-I	1.15	25.00	3.45	50.00	11.62	97.47	97.59	92.77
Olmo-3.1-32B-I	2.30	40.00	2.30	**100.00**	5.81	97.56	96.47	96.47
Aya-Expanse-32B	2.30	33.33	3.45	66.67	6.25	97.59	96.39	97.59
Llama-3.3-70B-I	5.75	23.81	5.75	96.77	29.78	85.71	88.73	80.28
Qwen3.6-Plus	65.52	77.03	2.30	52.17	21.66	52.63	40.00	60.00
DeepSeek-V3.1	0.00	0.00	4.60	75.00	8.33	96.39	96.47	96.47
Command-A	1.15	14.29	3.45	50.00	23.03	86.25	85.90	93.59
o4-mini	2.30	40.00	**0.00**	100.00	5.60	97.53	94.12	95.29
GPT-5.2	64.37	78.87	1.15	75.00	22.81	69.23	55.56	50.00
Gemini-3-Flash	9.20	44.44	4.60	71.43	26.98	89.06	84.29	84.29
Claude Haiku 4.5	29.89	46.43	16.09	77.55	48.73	68.57	60.53	55.26
Human (Base)	58.62	64.56	1.15	44.44	34.00	0.00	0.00	11.11
Human (Expert)	**98.85**	**98.85**	0.00	0.00	**0.93**	0.00	0.00	0.00

Table 66. Novel metrics — Sub-Type: *B5. Constraint-Driven Algebraic Reasoning Problems*.

	Bilateral Symmetry			*Entropy*	*Invariant Recovery (%)*			*Fragility Gradient*	
Model	BSS↑	$F^{+}(\%)$	$F^{-}(\%)$	CE↓	IRR_{full} ↑	IRR_{part} ↑	IRR_{Δ}	FG	$\overline{g}$
Llama-3.2-1B-I	50.00	3.45	1.15	3.95	0.00	1.20	-75.00	**2**	+1.73
Ministral-3B-2512	50.00	3.45	1.15	4.16	0.00	1.19	-100.00	1	+1.15
Gemma-3-4B-I	66.67	1.15	2.30	3.01	0.00	2.38	-33.33	0	-1.15
Qwen3-8B	88.89	18.39	22.99	35.96	19.35	64.52	-9.22	1	+3.45
Gemma-4-26B-A4B-I	50.00	3.45	10.34	11.62	1.20	10.84	-73.80	1	+0.00
Olmo-3.1-32B-I	**100.00**	3.45	3.45	5.81	0.00	3.66	-60.00	2	+1.73
Aya-Expanse-32B	66.67	4.60	2.30	6.25	0.00	2.47	-66.67	2	+0.57
Llama-3.3-70B-I	96.77	18.39	17.24	29.78	3.03	22.73	-73.16	1	+2.30
Qwen3.6-Plus	52.17	19.54	6.90	21.66	7.69	46.15	-15.28	2	+4.60
DeepSeek-V3.1	75.00	5.75	3.45	8.33	0.00	3.66	-100.00	1	+0.57
Command-A	50.00	6.90	20.69	23.03	0.00	22.50	-85.71	1	+2.87
o4-mini	100.00	3.45	3.45	5.60	0.00	3.66	-60.00	2	+2.87
GPT-5.2	75.00	17.24	10.34	22.81	18.75	56.25	-2.38	1	+0.58
Gemini-3-Flash	71.43	11.49	20.69	26.98	2.90	26.09	-52.66	1	+2.88
Claude Haiku 4.5	77.55	34.48	21.84	48.73	12.90	61.29	-40.67	1	+0.58
Human (Base)	44.44	32.18	9.20	34.00	**87.50**	**100.00**	+52.06	0	-1.15
Human (Expert)	0.00	1.15	0.00	**0.93**	—	—	—	2	+0.57

Table 67. Sensitivity profile — Sub-Type: *B5. Constraint-Driven Algebraic Reasoning Problems*.

	Cond. Inv. Retention (%)			*Collapse Contribution (%)*			*Rel. Sensitivity (pp)*		
Model	TCIR_1	TCIR_2	TCIR_3	$\text{TCC}_1 \downarrow$	$\text{TCC}_2 \downarrow$	$\text{TCC}_3 \downarrow$	RIS_1	RIS_2	RIS_3
Llama-3.2-1B-I	80.00	40.00	20.00	33.33	100.00	100.00	+1.44	-0.86	-2.01
Ministral-3B-2512	50.00	75.00	0.00	66.67	33.33	100.00	+0.00	+1.15	-2.30
Gemma-3-4B-I	40.00	60.00	80.00	100.00	66.67	33.33	-1.15	+0.00	+1.15
Qwen3-8B	92.11	75.00	84.21	25.00	87.50	31.25	+9.48	-5.46	+2.58
Gemma-4-26B-A4B-I	30.77	61.54	30.77	81.82	45.45	81.82	-1.15	+3.45	-1.15
Olmo-3.1-32B-I	62.50	62.50	25.00	60.00	60.00	100.00	+0.86	+0.86	-2.59
Aya-Expanse-32B	62.50	50.00	50.00	60.00	80.00	60.00	+0.29	-0.86	-0.86
Llama-3.3-70B-I	55.56	66.67	44.44	72.22	44.44	88.89	-0.29	+4.31	-4.89
Qwen3.6-Plus	93.75	85.00	83.75	42.86	100.00	71.43	+4.60	-3.45	-4.60
DeepSeek-V3.1	37.50	50.00	25.00	57.14	57.14	85.71	-0.57	+0.58	-1.72
Command-A	56.00	28.00	36.00	47.62	85.71	66.67	+5.46	-2.58	-0.29
o4-mini	87.50	75.00	25.00	33.33	66.67	100.00	+2.30	+1.15	-3.45
GPT-5.2	87.50	92.50	86.25	57.14	57.14	71.43	-1.15	+3.45	-2.30
Gemini-3-Flash	61.11	63.89	47.22	70.00	60.00	70.00	+2.30	+3.45	-3.45
Claude Haiku 4.5	66.67	69.33	65.33	72.41	62.07	65.52	-2.01	+0.29	-3.16
Human (Base)	87.36	88.51	89.66	**0.00**	50.00	100.00	-1.72	-0.57	+0.58
Human (Expert)	**100.00**	**100.00**	**98.85**	0.00	**0.00**	**0.00**	+0.29	+0.29	-0.86

H.3. C. Relational, Planning, Algorithmic and Semantic Reasoning

Table 68. Core accuracy — Task Type: *C. Relational, Planning, Algorithmic and Semantic Reasoning*.

Model	Base	*Accuracy (%)* $\mathcal{I}_1$	$\mathcal{I}_2$	$\mathcal{I}_3$	*Mean Inv.* $\overline{\mathcal{I}}$	*Base*$-\overline{\mathcal{I}}$ Δ
Llama-3.2-1B-I	11.52	9.31	11.76	14.46	11.84	-0.32
Ministral-3B-2512	20.10	15.93	19.85	18.38	18.05	+2.05
Gemma-3-4B-I	15.44	16.67	18.63	17.40	17.57	-2.13
Qwen3-8B	49.75	48.53	54.66	50.00	51.06	-1.31
Gemma-4-26B-A4B-I	39.95	37.75	36.52	39.22	37.83	+2.12
Olmo-3.1-32B-I	24.75	19.36	23.28	24.51	22.38	+2.37
Aya-Expanse-32B	26.23	22.55	27.21	22.55	24.10	+2.13
Llama-3.3-70B-I	30.88	28.68	35.78	30.88	31.78	-0.90
Qwen3.6-Plus	61.76	58.58	60.54	60.54	59.89	+1.87
DeepSeek-V3.1	27.70	24.26	28.43	24.02	25.57	+2.13
Command-A	30.39	24.02	31.86	28.43	28.10	+2.29
o4-mini	31.62	24.02	26.72	25.25	25.33	+6.29
GPT-5.2	64.46	60.29	64.71	65.69	63.56	+0.90
Gemini-3-Flash	50.49	43.87	42.89	43.14	43.30	+7.19
Claude Haiku 4.5	36.52	40.69	44.61	38.48	41.26	**-4.74**
Human (Base)	88.48	90.93	91.42	90.93	91.09	-2.61
Human (Expert)	**100.00**	**100.00**	**99.51**	**99.02**	**99.51**	+0.49

Table 69. Invariance stability — Task Type: *C. Relational, Planning, Algorithmic and Semantic Reasoning*.

Model	*Global Stability (%)* IMI↑	CIR↑	ICR↓	*Novel (%)* BSS↑	CE↓	*Pairwise Failure Dep. (%)* $P(\mathcal{I}_1{=}0\|\mathcal{I}_2{=}0)$	$P(\mathcal{I}_1{=}0\|\mathcal{I}_3{=}0)$	$P(\mathcal{I}_2{=}0\|\mathcal{I}_3{=}0)$
Llama-3.2-1B-I	2.94	25.53	5.64	73.68	19.95	94.72	95.42	93.12
Ministral-3B-2512	2.21	10.98	11.52	93.59	33.38	89.30	88.29	84.68
Gemma-3-4B-I	3.68	23.81	6.62	69.06	29.07	88.86	89.32	86.35
Qwen3-8B	29.90	60.10	7.35	99.39	34.40	83.24	78.92	73.04
Gemma-4-26B-A4B-I	9.31	23.31	12.99	92.70	49.52	72.97	76.21	73.79
Olmo-3.1-32B-I	4.66	18.81	12.25	95.35	36.47	87.54	87.34	84.42
Aya-Expanse-32B	5.15	19.63	11.52	99.42	36.45	86.20	85.13	80.70
Llama-3.3-70B-I	7.84	25.40	12.25	88.26	45.59	80.15	81.21	74.47
Qwen3.6-Plus	42.40	68.65	6.13	78.46	27.56	79.50	77.02	77.02
DeepSeek-V3.1	4.17	15.04	14.22	**99.48**	41.29	85.62	81.94	80.32
Command-A	6.86	22.58	11.52	98.97	40.89	84.53	85.62	78.08
o4-mini	4.90	15.50	16.67	88.78	41.98	83.28	83.93	80.98
GPT-5.2	41.91	65.02	6.13	72.22	30.25	77.78	80.71	74.29
Gemini-3-Flash	20.83	41.26	13.48	73.30	40.43	75.11	75.86	78.45
Claude Haiku 4.5	9.80	26.85	13.97	85.83	54.58	71.24	72.91	65.34
Human (Base)	67.40	76.18	2.45	70.68	27.32	11.43	13.51	10.81
Human (Expert)	**98.53**	**98.53**	**0.00**	0.00	**1.19**	0.00	0.00	0.00

Table 70. Novel metrics — Task Type: *C. Relational, Planning, Algorithmic and Semantic Reasoning*.

Model	*Bilateral Symmetry* BSS↑	$F^+(\%)$	$F^-(\%)$	*Entropy* CE↓	*Invariant Recovery (%)* IRR_{full} ↑	IRR_{part} ↑	IRR_{Δ}	*Fragility Gradient* FG	$\overline{g}$
Llama-3.2-1B-I	73.68	8.58	14.71	19.95	0.28	16.62	-74.19	0	-2.58
Ministral-3B-2512	93.59	17.89	20.34	33.38	1.23	25.46	-87.79	1	-1.23
Gemma-3-4B-I	69.06	11.76	22.30	29.07	0.87	26.38	-75.32	1	-0.36
Qwen3-8B	99.39	19.85	20.10	34.40	7.80	40.00	-32.10	1	-0.73
Gemma-4-26B-A4B-I	92.70	30.64	26.47	49.52	5.71	44.08	-70.98	1	-0.73
Olmo-3.1-32B-I	95.35	20.10	22.06	36.47	2.28	29.32	-78.91	0	-2.57
Aya-Expanse-32B	99.42	21.08	20.83	36.45	2.66	28.24	-77.71	1	+0.00
Llama-3.3-70B-I	88.26	23.04	29.17	45.59	2.84	42.20	-71.76	1	-1.10
Qwen3.6-Plus	78.46	19.36	12.50	27.56	5.77	32.69	-25.58	1	-0.98
DeepSeek-V3.1	**99.48**	23.53	23.77	41.29	4.41	32.88	-80.55	1	+0.12
Command-A	98.97	23.53	24.02	40.89	2.46	34.51	-74.96	1	-2.21
o4-mini	88.78	26.72	21.32	41.98	2.87	31.18	-81.63	1	-0.62
GPT-5.2	72.22	22.55	12.75	30.25	15.17	35.86	-19.81	0	-2.70
Gemini-3-Flash	73.30	29.66	17.16	40.43	4.95	34.65	-53.79	1	+0.36
Claude Haiku 4.5	85.83	26.72	35.54	54.58	9.27	55.98	-63.88	1	+1.11
Human (Base)	70.68	21.08	11.52	27.32	**76.60**	**100.00**	+52.78	1	+0.00
Human (Expert)	0.00	1.47	0.00	**1.19**	—	—	—	**2**	+0.49

Table 71. Sensitivity profile — Task Type: *C. Relational, Planning, Algorithmic and Semantic Reasoning*.

Model	*Cond. Inv. Retention (%)*			*Collapse Contribution (%)*			*Rel. Sensitivity (pp)*		
	$TCIR_1$	$TCIR_2$	$TCIR_3$	$TCC_1 \downarrow$	$TCC_2 \downarrow$	$TCC_3 \downarrow$	RIS_1	RIS_2	RIS_3
Llama-3.2-1B-I	35.51	44.86	55.14	80.49	64.63	54.88	-2.45	+-0.00	+2.70
Ministral-3B-2512	39.39	49.09	45.45	69.84	59.52	67.46	-2.64	+1.28	-0.19
Gemma-3-4B-I	44.16	49.35	46.10	70.43	57.39	68.70	-0.36	+1.60	+0.37
Qwen3-8B	69.47	78.25	71.58	70.83	52.08	63.54	-2.21	+3.92	-0.74
Gemma-4-26B-A4B-I	56.83	54.98	59.04	65.99	62.59	60.54	-0.61	-1.84	+0.86
Olmo-3.1-32B-I	41.36	49.74	52.36	75.94	63.91	60.90	-3.62	+0.30	+1.53
Aya-Expanse-32B	47.92	57.81	47.92	70.16	54.84	70.16	-2.08	+2.58	-2.08
Llama-3.3-70B-I	47.76	59.59	51.43	68.94	52.80	65.84	-2.88	+4.22	-0.68
Qwen3.6-Plus	78.88	81.52	81.52	65.67	59.70	56.72	-1.78	+0.18	+0.18
DeepSeek-V3.1	47.14	55.24	46.67	66.90	61.27	68.31	-1.84	+2.33	-2.08
Command-A	44.14	58.56	52.25	76.09	55.07	66.67	-4.66	+3.18	-0.25
o4-mini	45.37	50.46	47.69	70.07	65.31	66.67	-2.88	-0.18	-1.65
GPT-5.2	78.10	83.81	85.08	80.00	52.73	49.09	-3.50	+0.92	+1.90
Gemini-3-Flash	64.86	63.41	63.77	63.48	68.70	69.57	-1.23	-2.21	-1.96
Claude Haiku 4.5	56.46	61.90	53.40	62.36	51.12	69.10	+0.62	+4.54	-1.59
Human (Base)	90.93	91.42	90.93	57.14	57.14	42.86	+0.49	+0.98	+0.49
Human (Expert)	**100.00**	**99.51**	**99.02**	**0.00**	**0.00**	**0.00**	+0.37	-0.12	-0.61

Table 72. Sub-type accuracy breakdown for $\mathcal{I}_1$ — Task Type: *C. Relational, Planning, Algorithmic and Semantic Reasoning*.

Model	*Accuracy (%) by Sub-Type — $\mathcal{I}_1$ (Symbolic & Linguistic)*			
	Entity Relabeling	Style / Register Invariance	Synonym / Semantic Paraphrase...	Variable Renaming
Llama-3.2-1B-I	4.59	9.23	16.44	9.38
Ministral-3B-2512	11.93	19.23	21.92	11.46
Gemma-3-4B-I	13.76	21.54	19.18	11.46
Qwen3-8B	66.06	40.00	53.42	36.46
Gemma-4-26B-A4B-I	38.53	40.77	42.47	29.17
Olmo-3.1-32B-I	15.60	20.00	26.03	17.71
Aya-Expanse-32B	14.68	24.62	35.62	18.75
Llama-3.3-70B-I	26.61	24.62	38.36	29.17
Qwen3.6-Plus	67.89	54.62	64.38	48.96
DeepSeek-V3.1	14.68	27.69	35.62	21.88
Command-A	21.10	19.23	38.36	22.92
o4-mini	13.76	23.85	32.88	29.17
GPT-5.2	70.64	60.00	60.27	48.96
Gemini-3-Flash	36.70	43.08	58.90	41.67
Claude Haiku 4.5	41.28	38.46	49.32	36.46
Human (Base)	95.41	89.23	90.41	88.54
Human (Expert)	**100.00**	**100.00**	**100.00**	**100.00**

Table 73. Sub-type accuracy breakdown for $\mathcal{I}_2$ — Task Type: *C. Relational, Planning, Algorithmic and Semantic Reasoning*.

Model	*Accuracy (%) by Sub-Type — $\mathcal{I}_2$ (Structural & Representational)*			
	Equivalent Logical Reformulation	Equivalent Mathematical Reformulation	Premise Reordering	Representation / Encoding Swap
Llama-3.2-1B-I	7.42	25.00	13.33	16.07
Ministral-3B-2512	14.85	33.33	20.00	28.57
Gemma-3-4B-I	13.97	29.17	24.00	21.43
Qwen3-8B	48.03	62.50	56.00	73.21
Gemma-4-26B-A4B-I	28.38	54.17	41.33	48.21
Olmo-3.1-32B-I	14.85	47.92	25.33	33.93
Aya-Expanse-32B	21.40	47.92	28.00	32.14
Llama-3.3-70B-I	29.69	58.33	40.00	35.71
Qwen3.6-Plus	55.46	64.58	61.33	76.79
DeepSeek-V3.1	19.65	50.00	36.00	35.71
Command-A	20.09	50.00	37.33	57.14
o4-mini	20.52	45.83	29.33	32.14
GPT-5.2	62.01	66.67	61.33	78.57
Gemini-3-Flash	27.07	60.42	52.00	80.36
Claude Haiku 4.5	41.48	62.50	41.33	46.43
Human (Base)	89.08	91.67	93.33	**98.21**
Human (Expert)	**99.56**	**100.00**	**100.00**	98.21

Table 74. Sub-type accuracy breakdown for $\mathcal{I}_3$ — Task Type: *C. Relational, Planning, Algorithmic and Semantic Reasoning*.

Model	*Accuracy (%) by Sub-Type — $\mathcal{I}_3$ (Abstraction & Info-Structure)*			
	Abstraction Level Shift	Distractor Injection / Removal	Perspective / Role Reframing	Redundant Restatement Invariance
Llama-3.2-1B-I	16.18	16.28	17.78	12.44
Ministral-3B-2512	27.94	19.77	17.78	14.83
Gemma-3-4B-I	35.29	19.77	11.11	11.96
Qwen3-8B	64.71	51.16	53.33	44.02
Gemma-4-26B-A4B-I	55.88	39.53	44.44	32.54
Olmo-3.1-32B-I	35.29	25.58	26.67	20.10
Aya-Expanse-32B	38.24	19.77	26.67	17.70
Llama-3.3-70B-I	47.06	25.58	46.67	24.40
Qwen3.6-Plus	72.06	66.28	66.67	53.11
DeepSeek-V3.1	38.24	30.23	33.33	14.83
Command-A	47.06	32.56	40.00	18.18
o4-mini	42.65	29.07	20.00	19.14
GPT-5.2	69.12	59.30	75.56	65.07
Gemini-3-Flash	64.71	52.33	64.44	27.75
Claude Haiku 4.5	61.76	45.35	44.44	26.79
Human (Base)	92.65	86.05	86.67	93.30
Human (Expert)	**100.00**	**98.84**	**97.78**	**99.04**

H.3.1. C1. ALGEBRA, PSEUDOCODE, AND ALGORITHMIC REASONING

Table 75. Core accuracy — Sub-Type: *C1. Algebra, Pseudocode, and Algorithmic Reasoning*.

Model	Base	*Accuracy (%)* $\mathcal{I}_1$	$\mathcal{I}_2$	$\mathcal{I}_3$	*Mean Inv.* $\overline{\mathcal{I}}$	*Base*$-\overline{\mathcal{I}}$ Δ
Llama-3.2-1B-I	7.46	8.96	11.94	16.42	12.44	-4.98
Ministral-3B-2512	16.42	17.91	23.88	28.36	23.38	-6.96
Gemma-3-4B-I	10.45	11.94	28.36	22.39	20.90	-10.45
Qwen3-8B	52.24	55.22	70.15	67.16	64.18	-11.94
Gemma-4-26B-A4B-I	47.76	44.78	55.22	55.22	51.74	-3.98
Olmo-3.1-32B-I	29.85	25.37	41.79	37.31	34.82	-4.97
Aya-Expanse-32B	25.37	22.39	41.79	32.84	32.34	-6.97
Llama-3.3-70B-I	34.33	34.33	49.25	46.27	43.28	-8.95
Qwen3.6-Plus	68.66	65.67	73.13	82.09	73.63	-4.97
DeepSeek-V3.1	29.85	40.30	52.24	40.30	44.28	**-14.43**
Command-A	37.31	37.31	47.76	43.28	42.78	-5.47
o4-mini	37.31	23.88	37.31	34.33	31.84	+5.47
GPT-5.2	67.16	67.16	71.64	76.12	71.64	-4.48
Gemini-3-Flash	62.69	58.21	62.69	68.66	63.19	-0.50
Claude Haiku 4.5	55.22	53.73	52.24	58.21	54.73	+0.49
Human (Base)	85.07	85.07	88.06	86.57	86.57	-1.50
Human (Expert)	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	+0.00

Table 76. Invariance stability — Sub-Type: *C1. Algebra, Pseudocode, and Algorithmic Reasoning*.

Model	*Global Stability (%)* IMI↑	CIR↑	ICR↓	*Novel (%)* BSS↑	CE↓	*Pairwise Failure Dep. (%)* $P(\mathcal{I}_1{=}0\|\mathcal{I}_2{=}0)$	$P(\mathcal{I}_1{=}0\|\mathcal{I}_3{=}0)$	$P(\mathcal{I}_2{=}0\|\mathcal{I}_3{=}0)$
Llama-3.2-1B-I	0.00	0.00	5.97	52.63	24.98	94.92	94.64	91.07
Ministral-3B-2512	0.00	0.00	13.43	61.11	47.53	96.08	81.25	75.00
Gemma-3-4B-I	2.99	28.57	2.99	34.48	36.24	95.83	92.31	73.08
Qwen3-8B	28.36	54.29	10.45	76.19	54.80	75.00	68.18	40.91
Gemma-4-26B-A4B-I	11.94	25.00	14.93	**100.00**	62.63	66.67	66.67	60.00
Olmo-3.1-32B-I	7.46	25.00	16.42	78.95	49.96	92.31	85.71	69.05
Aya-Expanse-32B	4.48	17.65	8.96	75.68	47.90	87.18	82.22	73.33
Llama-3.3-70B-I	4.48	13.04	19.40	81.63	65.25	79.41	69.44	66.67
Qwen3.6-Plus	43.28	63.04	5.97	96.97	42.49	61.11	66.67	58.33
DeepSeek-V3.1	2.99	10.00	13.43	70.59	66.26	87.50	60.00	52.50
Command-A	13.43	36.00	10.45	88.89	47.25	80.00	78.95	71.05
o4-mini	4.48	12.00	22.39	95.24	55.08	85.71	81.82	72.73
GPT-5.2	34.33	51.11	5.97	84.21	47.98	52.63	68.75	43.75
Gemini-3-Flash	29.85	47.62	14.93	92.68	53.59	68.00	52.38	57.14
Claude Haiku 4.5	16.42	29.73	20.90	96.00	64.77	53.12	67.86	67.86
Human (Base)	55.22	64.91	4.48	66.67	37.73	25.00	11.11	11.11
Human (Expert)	**100.00**	**100.00**	**0.00**	—	**0.00**	0.00	0.00	0.00

Table 77. Novel metrics — Sub-Type: *C1. Algebra, Pseudocode, and Algorithmic Reasoning*.

Model	*Bilateral Symmetry*			*Entropy*	*Invariant Recovery (%)*			*Fragility Gradient*	
	BSS↑	$F^{+}(\%)$	$F^{-}(\%)$	CE↓	IRR_{full} ↑	IRR_{part} ↑	IRR_{Δ}	FG	$\overline{g}$
Llama-3.2-1B-I	52.63	7.46	20.90	24.98	1.61	22.58	-98.39	0	-3.73
Ministral-3B-2512	61.11	16.42	37.31	47.53	3.57	44.64	-96.43	0	-5.23
Gemma-3-4B-I	34.48	7.46	35.82	36.24	1.67	40.00	-69.76	1	-5.23
Qwen3-8B	76.19	23.88	38.81	54.80	21.88	81.25	-23.83	1	-5.97
Gemma-4-26B-A4B-I	**100.00**	35.82	35.82	62.63	17.14	68.57	-57.86	1	-5.22
Olmo-3.1-32B-I	78.95	22.39	34.33	49.96	10.64	48.94	-64.36	1	-5.97
Aya-Expanse-32B	75.68	20.90	34.33	47.90	6.00	46.00	-76.35	1	-5.23
Llama-3.3-70B-I	81.63	29.85	43.28	65.25	15.91	65.91	-71.05	1	-5.97
Qwen3.6-Plus	96.97	25.37	23.88	42.49	23.81	76.19	-13.15	0	-8.21
DeepSeek-V3.1	70.59	26.87	49.25	66.26	17.02	70.21	-72.98	1	+0.00
Command-A	88.89	23.88	29.85	47.25	9.52	47.62	-54.48	1	-2.98
o4-mini	95.24	32.84	29.85	55.08	7.14	47.62	-80.86	1	-5.23
GPT-5.2	84.21	32.84	23.88	47.98	36.36	72.73	-12.53	0	-4.48
Gemini-3-Flash	92.68	32.84	28.36	53.59	20.00	76.00	-32.38	0	-5.23
Claude Haiku 4.5	96.00	38.81	35.82	64.77	23.33	80.00	-46.94	1	-2.24
Human (Base)	66.67	29.85	14.93	37.73	**70.00**	**100.00**	+34.91	1	-0.75
Human (Expert)	—	0.00	0.00	**0.00**	—	—	—	**2**	+0.00

Table 78. Sensitivity profile — Sub-Type: *C1. Algebra, Pseudocode, and Algorithmic Reasoning*.

Model	*Cond. Inv. Retention (%)*			*Collapse Contribution (%)*			*Rel. Sensitivity (pp)*		
	TCIR_1	TCIR_2	TCIR_3	$\text{TCC}_1 \downarrow$	$\text{TCC}_2 \downarrow$	$\text{TCC}_3 \downarrow$	RIS_1	RIS_2	RIS_3
Llama-3.2-1B-I	31.58	42.11	57.89	76.47	64.71	41.18	-2.23	+0.75	+5.23
Ministral-3B-2512	33.33	44.44	52.78	75.00	59.38	50.00	-3.73	+2.24	+6.72
Gemma-3-4B-I	25.81	61.29	48.39	84.00	44.00	64.00	-6.34	+10.08	+4.11
Qwen3-8B	60.66	77.05	73.77	76.92	46.15	50.00	-5.97	+8.96	+5.97
Gemma-4-26B-A4B-I	53.57	66.07	66.07	71.43	57.14	50.00	-5.97	+4.47	+4.47
Olmo-3.1-32B-I	39.53	65.12	58.14	82.76	51.72	55.17	-8.21	+8.21	+3.73
Aya-Expanse-32B	37.50	70.00	55.00	76.92	46.15	57.69	-8.21	+11.19	+2.24
Llama-3.3-70B-I	44.23	63.46	59.62	68.57	54.29	54.29	-6.71	+8.21	+5.23
Qwen3.6-Plus	70.97	79.03	88.71	80.00	60.00	26.67	-6.72	+0.74	+9.70
DeepSeek-V3.1	50.94	66.04	50.94	67.65	52.94	58.82	-0.37	+11.57	-0.37
Command-A	55.56	71.11	64.44	65.22	47.83	60.87	-4.11	+6.34	+1.86
o4-mini	35.56	55.56	51.11	78.12	59.38	62.50	-9.33	+4.10	+1.12
GPT-5.2	73.77	78.69	83.61	91.67	41.67	41.67	-3.36	+1.12	+5.60
Gemini-3-Flash	63.93	68.85	75.41	66.67	75.00	41.67	-4.85	-0.37	+5.60
Claude Haiku 4.5	59.02	57.38	63.93	70.97	58.06	67.74	-1.12	-2.61	+3.36
Human (Base)	85.07	88.06	86.57	66.67	66.67	33.33	-1.12	+1.87	+0.38
Human (Expert)	**100.00**	**100.00**	**100.00**	**0.00**	**0.00**	**0.00**	+0.00	+0.00	+0.00

H.3.2. C2. GRAPH SEARCH AND TRAVERSAL REASONING

Table 79. Core accuracy — Sub-Type: *C2. Graph Search and Traversal Reasoning*.

Model	Base	*Accuracy (%)* $\mathcal{I}_1$	$\mathcal{I}_2$	$\mathcal{I}_3$	*Mean Inv.* $\overline{\mathcal{I}}$	*Base*$-\overline{\mathcal{I}}$ Δ
Llama-3.2-1B-I	12.38	10.48	16.19	16.19	14.29	-1.91
Ministral-3B-2512	13.33	14.29	18.10	12.38	14.92	-1.59
Gemma-3-4B-I	8.57	19.05	17.14	13.33	16.51	-7.94
Qwen3-8B	39.05	36.19	44.76	40.00	40.32	-1.27
Gemma-4-26B-A4B-I	26.67	31.43	24.76	32.38	29.52	-2.85
Olmo-3.1-32B-I	15.24	13.33	17.14	18.10	16.19	-0.95
Aya-Expanse-32B	18.10	20.00	20.00	17.14	19.05	-0.95
Llama-3.3-70B-I	20.95	20.00	27.62	21.90	23.17	-2.22
Qwen3.6-Plus	48.57	50.48	47.62	48.57	48.89	-0.32
DeepSeek-V3.1	20.00	19.05	21.90	15.24	18.73	+1.27
Command-A	17.14	17.14	22.86	19.05	19.68	-2.54
o4-mini	18.10	20.95	22.86	17.14	20.32	-2.22
GPT-5.2	55.24	55.24	55.24	55.24	55.24	+0.00
Gemini-3-Flash	32.38	29.52	23.81	25.71	26.35	+6.03
Claude Haiku 4.5	21.90	34.29	42.86	28.57	35.24	**-13.34**
Human (Base)	90.48	91.43	87.62	92.38	90.48	+0.00
Human (Expert)	**100.00**	**100.00**	**100.00**	**99.05**	**99.68**	+0.32

Table 80. Invariance stability — Sub-Type: *C2. Graph Search and Traversal Reasoning*.

Model	*Global Stability (%)* IMI↑	CIR↑	ICR↓	*Novel (%)* BSS↑	CE↓	*Pairwise Failure Dep. (%)* $P(\mathcal{I}_1{=}0\|\mathcal{I}_2{=}0)$	$P(\mathcal{I}_1{=}0\|\mathcal{I}_3{=}0)$	$P(\mathcal{I}_2{=}0\|\mathcal{I}_3{=}0)$
Llama-3.2-1B-I	7.62	61.54	2.86	52.63	15.22	97.73	97.73	90.91
Ministral-3B-2512	2.86	21.43	4.76	75.86	23.49	89.53	92.39	85.87
Gemma-3-4B-I	4.76	55.56	0.95	30.77	20.99	86.21	89.01	91.21
Qwen3-8B	28.57	73.17	6.67	73.33	24.98	89.66	92.06	80.95
Gemma-4-26B-A4B-I	5.71	21.43	10.48	88.00	42.05	77.22	84.51	84.51
Olmo-3.1-32B-I	3.81	25.00	5.71	77.42	25.39	93.10	91.86	89.53
Aya-Expanse-32B	0.95	5.26	8.57	87.80	33.84	86.90	85.06	81.61
Llama-3.3-70B-I	5.71	27.27	7.62	74.42	35.38	84.21	91.46	76.83
Qwen3.6-Plus	38.10	78.43	3.81	88.00	20.39	83.64	85.19	85.19
DeepSeek-V3.1	6.67	33.33	9.52	87.50	26.16	89.02	89.89	85.39
Command-A	2.86	16.67	9.52	69.77	35.20	86.42	89.41	81.18
o4-mini	4.76	26.32	8.57	66.67	33.71	82.72	87.36	85.06
GPT-5.2	41.90	75.86	4.76	88.00	20.57	82.98	89.36	85.11
Gemini-3-Flash	12.38	38.24	14.29	**92.31**	32.29	82.50	84.62	89.74
Claude Haiku 4.5	6.67	30.43	6.67	53.33	49.59	81.67	76.00	61.33
Human (Base)	67.62	74.74	1.90	58.82	26.99	15.38	12.50	12.50
Human (Expert)	**99.05**	**99.05**	**0.00**	0.00	**0.77**	0.00	0.00	0.00

Table 81. Novel metrics — Sub-Type: *C2. Graph Search and Traversal Reasoning*.

Model	*Bilateral Symmetry* BSS↑	F^+(%)	F^-(%)	*Entropy* CE↓	*Invariant Recovery (%)* IRR_{full} ↑	IRR_{part} ↑	IRR_{Δ}	*Fragility Gradient* FG	$\overline{g}$
Llama-3.2-1B-I	52.63	4.76	13.33	15.22	0.00	15.22	-38.46	1	-2.85
Ministral-3B-2512	75.86	10.48	17.14	23.49	1.10	19.78	-77.47	1	+0.95
Gemma-3-4B-I	30.77	3.81	20.95	20.99	2.08	22.92	-42.36	**2**	+2.86
Qwen3-8B	73.33	10.48	18.10	24.98	0.00	29.69	-26.83	1	-1.90
Gemma-4-26B-A4B-I	88.00	20.95	26.67	42.05	3.90	36.36	-74.67	1	-0.48
Olmo-3.1-32B-I	77.42	11.43	18.10	25.39	0.00	21.35	-75.00	0	-2.38
Aya-Expanse-32B	87.80	17.14	21.90	33.84	2.33	26.74	-92.41	2	+1.43
Llama-3.3-70B-I	74.42	15.24	25.71	35.38	0.00	32.53	-72.73	1	-0.95
Qwen3.6-Plus	88.00	10.48	13.33	20.39	0.00	25.93	-21.57	1	+0.95
DeepSeek-V3.1	87.50	13.33	17.14	26.16	2.38	21.43	-64.29	1	+1.91
Command-A	69.77	14.29	26.67	35.20	2.30	32.18	-81.03	1	-0.95
o4-mini	66.67	13.33	26.67	33.71	2.33	32.56	-71.35	1	+1.91
GPT-5.2	88.00	13.33	10.48	20.57	6.38	23.40	-17.76	2	+0.00
Gemini-3-Flash	**92.31**	20.00	17.14	32.29	1.41	25.35	-60.35	1	+1.90
Claude Haiku 4.5	53.33	15.24	41.90	49.59	6.10	53.66	-63.47	1	+2.86
Human (Base)	58.82	22.86	9.52	26.99	**70.00**	**100.00**	+44.74	1	-0.48
Human (Expert)	0.00	0.95	0.00	**0.77**	—	—	—	2	+0.47

Table 82. Sensitivity profile — Sub-Type: *C2. Graph Search and Traversal Reasoning*.

Model	*Cond. Inv. Retention (%)* TCIR_1	TCIR_2	TCIR_3	*Collapse Contribution (%)* TCC_1 ↓	TCC_2 ↓	TCC_3 ↓	*Rel. Sensitivity (pp)* RIS_1	RIS_2	RIS_3
Llama-3.2-1B-I	40.74	62.96	62.96	88.24	52.94	58.82	-3.33	+2.38	+2.38
Ministral-3B-2512	46.88	59.38	40.62	68.18	45.45	81.82	-0.23	+3.58	-2.14
Gemma-3-4B-I	64.52	58.06	45.16	52.38	47.62	80.95	+4.53	+2.62	-1.19
Qwen3-8B	63.33	78.33	70.00	76.92	42.31	69.23	-3.81	+4.76	+0.00
Gemma-4-26B-A4B-I	58.93	46.43	60.71	52.78	63.89	61.11	+2.62	-4.05	+3.57
Olmo-3.1-32B-I	40.00	51.43	54.29	76.00	64.00	52.00	-2.62	+1.19	+2.15
Aya-Expanse-32B	50.00	50.00	42.86	66.67	53.33	70.00	+1.19	+1.19	-1.67
Llama-3.3-70B-I	42.86	59.18	46.94	74.29	45.71	68.57	-2.62	+5.00	-0.72
Qwen3.6-Plus	81.54	76.92	78.46	55.56	61.11	72.22	+1.67	-1.19	-0.24
DeepSeek-V3.1	51.28	58.97	41.03	69.23	57.69	80.77	+0.00	+2.85	-3.81
Command-A	39.13	52.17	43.48	77.78	52.78	66.67	-1.91	+3.81	+0.00
o4-mini	46.81	51.06	38.30	68.57	57.14	80.00	+1.19	+3.10	-2.62
GPT-5.2	84.06	84.06	84.06	61.54	53.85	69.23	+0.00	+0.00	+0.00
Gemini-3-Flash	59.62	48.08	51.92	62.50	75.00	71.88	+1.66	-4.05	-2.15
Claude Haiku 4.5	53.73	67.16	44.78	63.04	43.48	69.57	+2.39	+10.96	-3.33
Human (Base)	91.43	87.62	92.38	60.00	60.00	40.00	+0.95	-2.86	+1.90
Human (Expert)	**100.00**	**100.00**	**99.05**	**0.00**	**0.00**	**0.00**	+0.24	+0.24	-0.71

H.3.3. C3. PATH-BASED SEARCH AND MULTI-AGENT INTERACTION REASONING

Table 83. Core accuracy — Sub-Type: *C3. Path-Based Search and Multi-Agent Interaction Reasoning*.

Model	Base	*Accuracy (%)* $\mathcal{I}_1$	$\mathcal{I}_2$	$\mathcal{I}_3$	*Mean Inv.* $\overline{\mathcal{I}}$	*Base*$-\overline{\mathcal{I}}$ Δ
Llama-3.2-1B-I	15.62	8.33	10.42	17.71	12.15	+3.47
Ministral-3B-2512	20.83	11.46	17.71	19.79	16.32	+4.51
Gemma-3-4B-I	23.96	12.50	17.71	20.83	17.01	+6.95
Qwen3-8B	65.62	58.33	69.79	61.46	63.19	+2.43
Gemma-4-26B-A4B-I	42.71	35.42	41.67	39.58	38.89	+3.82
Olmo-3.1-32B-I	31.25	14.58	23.96	34.38	24.31	+6.94
Aya-Expanse-32B	35.42	27.08	30.21	32.29	29.86	+5.56
Llama-3.3-70B-I	32.29	26.04	38.54	34.38	32.99	-0.70
Qwen3.6-Plus	76.04	63.54	72.92	70.83	69.10	+6.94
DeepSeek-V3.1	32.29	18.75	32.29	28.12	26.39	+5.90
Command-A	40.62	22.92	37.50	35.42	31.95	+8.67
o4-mini	36.46	25.00	29.17	29.17	27.78	+8.68
GPT-5.2	66.67	61.46	71.88	71.88	68.41	-1.74
Gemini-3-Flash	66.67	55.21	66.67	59.38	60.42	+6.25
Claude Haiku 4.5	36.46	36.46	45.83	48.96	43.75	**-7.29**
Human (Base)	91.58	92.63	96.84	91.58	93.68	-2.10
Human (Expert)	**100.00**	**100.00**	**97.89**	**98.95**	**98.95**	+1.05

Table 84. Invariance stability — Sub-Type: *C3. Path-Based Search and Multi-Agent Interaction Reasoning*.

Model	*Global Stability (%)* IMI↑	CIR↑	ICR↓	*Novel (%)* BSS↑	CE↓	*Pairwise Failure Dep. (%)* $P(\mathcal{I}_1{=}0\|\mathcal{I}_2{=}0)$	$P(\mathcal{I}_1{=}0\|\mathcal{I}_3{=}0)$	$P(\mathcal{I}_2{=}0\|\mathcal{I}_3{=}0)$
Llama-3.2-1B-I	1.04	6.67	10.42	93.33	26.93	93.02	94.94	96.20
Ministral-3B-2512	2.08	10.00	10.42	90.91	29.85	92.41	89.61	89.61
Gemma-3-4B-I	4.17	17.39	13.54	91.43	31.35	89.87	94.74	90.79
Qwen3-8B	36.46	55.56	6.25	69.77	37.91	82.76	62.16	64.86
Gemma-4-26B-A4B-I	12.50	29.27	11.46	96.43	48.70	75.00	79.31	72.41
Olmo-3.1-32B-I	5.21	16.67	16.67	86.36	39.94	90.41	93.65	88.89
Aya-Expanse-32B	12.50	35.29	12.50	84.21	34.08	85.07	87.69	87.69
Llama-3.3-70B-I	7.29	22.58	11.46	90.57	48.13	79.66	77.78	79.37
Qwen3.6-Plus	53.12	69.86	8.33	37.04	24.98	84.62	75.00	71.43
DeepSeek-V3.1	6.25	19.35	17.71	**97.96**	44.95	84.62	84.06	78.26
Command-A	7.29	17.95	14.58	76.92	46.89	80.00	85.48	79.03
o4-mini	5.21	14.29	17.71	72.34	43.65	82.35	83.82	79.41
GPT-5.2	41.67	62.50	4.17	73.68	33.88	77.78	70.37	74.07
Gemini-3-Flash	33.33	50.00	8.33	63.83	41.88	65.62	66.67	58.97
Claude Haiku 4.5	10.42	28.57	14.58	81.97	56.66	73.08	79.59	67.35
Human (Base)	75.79	82.76	1.05	69.57	20.24	0.00	12.50	0.00
Human (Expert)	**96.84**	**96.84**	**0.00**	0.00	**2.56**	0.00	0.00	0.00

Table 85. Novel metrics — Sub-Type: *C3. Path-Based Search and Multi-Agent Interaction Reasoning*.

	Bilateral Symmetry			*Entropy*	*Invariant Recovery (%)*			*Fragility Gradient*	
Model	BSS↑	F^{+}(%)	F^{-}(%)	CE↓	IRR_{full} ↑	IRR_{part} ↑	IRR_{Δ}	FG	$\overline{g}$
Llama-3.2-1B-I	93.33	14.58	16.67	26.93	0.00	19.75	-93.33	0	-4.69
Ministral-3B-2512	90.91	18.75	15.62	29.85	0.00	19.74	-90.00	0	-4.16
Gemma-3-4B-I	91.43	19.79	16.67	31.35	0.00	21.92	-82.61	0	-4.16
Qwen3-8B	69.77	29.17	15.62	37.91	15.15	45.45	-29.29	**1**	-1.56
Gemma-4-26B-A4B-I	96.43	30.21	28.12	48.70	5.45	49.09	-65.28	1	-2.08
Olmo-3.1-32B-I	86.36	26.04	19.79	39.94	3.03	28.79	-80.30	0	-9.90
Aya-Expanse-32B	84.21	22.92	16.67	34.08	3.23	25.81	-61.48	0	-2.61
Llama-3.3-70B-I	90.57	25.00	30.21	48.13	1.54	44.62	-75.88	1	-4.17
Qwen3.6-Plus	37.04	22.92	5.21	24.98	4.35	21.74	-25.79	1	-3.64
DeepSeek-V3.1	**97.96**	26.04	25.00	44.95	0.00	36.92	-80.65	1	-4.68
Command-A	76.92	33.33	20.83	46.89	0.00	35.09	-82.05	1	-6.25
o4-mini	72.34	31.25	17.71	43.65	1.64	27.87	-84.07	1	-2.09
GPT-5.2	73.68	25.00	14.58	33.88	21.88	43.75	-15.62	1	-5.21
Gemini-3-Flash	63.83	33.33	15.62	41.88	6.25	46.88	-43.75	1	-2.08
Claude Haiku 4.5	81.97	26.04	37.50	56.66	9.84	59.02	-61.59	0	-6.25
Human (Base)	69.57	15.79	8.42	20.24	**75.00**	**100.00**	+57.76	1	+0.53
Human (Expert)	0.00	3.16	0.00	**2.56**	—	—	—	1	+0.53

Table 86. Sensitivity profile — Sub-Type: *C3. Path-Based Search and Multi-Agent Interaction Reasoning*.

	Cond. Inv. Retention (%)			*Collapse Contribution (%)*			*Rel. Sensitivity (pp)*		
Model	TCIR_1	TCIR_2	TCIR_3	TCC_1 ↓	TCC_2 ↓	TCC_3 ↓	RIS_1	RIS_2	RIS_3
Llama-3.2-1B-I	25.81	32.26	54.84	80.77	73.08	53.85	-4.69	-2.60	+4.69
Ministral-3B-2512	31.43	48.57	54.29	76.00	68.00	56.00	-5.99	+0.26	+2.34
Gemma-3-4B-I	30.77	43.59	51.28	82.76	65.52	65.52	-6.25	-1.04	+2.08
Qwen3-8B	71.79	85.90	75.64	62.50	56.25	68.75	-5.47	+5.99	-2.34
Gemma-4-26B-A4B-I	50.00	58.82	55.88	77.14	62.86	71.43	-4.42	+1.83	-0.26
Olmo-3.1-32B-I	28.57	46.94	67.35	87.88	69.70	48.48	-11.46	-2.08	+8.34
Aya-Expanse-32B	52.00	58.00	62.00	69.23	69.23	69.23	-4.17	-1.04	+1.04
Llama-3.3-70B-I	41.67	61.67	55.00	64.10	56.41	64.10	-6.77	+5.73	+1.57
Qwen3.6-Plus	78.21	89.74	87.18	75.00	50.00	50.00	-7.29	+2.09	+-0.00
DeepSeek-V3.1	32.73	56.36	49.09	73.17	56.10	68.29	-9.11	+4.43	+0.26
Command-A	37.29	61.02	57.63	73.53	52.94	70.59	-11.19	+3.39	+1.31
o4-mini	46.15	53.85	53.85	63.64	69.70	57.58	-4.95	-0.78	-0.78
GPT-5.2	75.64	88.46	88.46	72.73	54.55	27.27	-6.51	+3.91	+3.91
Gemini-3-Flash	67.09	81.01	72.15	66.67	42.86	76.19	-6.77	+4.69	-2.60
Claude Haiku 4.5	49.30	61.97	66.20	63.64	54.55	54.55	-5.47	+3.90	+7.03
Human (Base)	92.63	96.84	91.58	66.67	33.33	33.33	-0.53	+3.68	-1.58
Human (Expert)	**100.00**	**97.89**	**98.95**	**0.00**	**0.00**	**0.00**	+0.79	-1.32	-0.26

H.3.4. C4. GRAPH ISOMORPHISM AND STRUCTURAL MAPPING

Table 87. Core accuracy — Sub-Type: *C4. Graph Isomorphism and Structural Mapping*.

Model	Base	*Accuracy (%)* $\mathcal{I}_1$	$\mathcal{I}_2$	$\mathcal{I}_3$	*Mean Inv.* $\overline{\mathcal{I}}$	*Base*$-\overline{\mathcal{I}}$ Δ
Llama-3.2-1B-I	7.69	7.69	6.73	6.73	7.05	+0.64
Ministral-3B-2512	21.15	20.19	18.27	17.31	18.59	+2.56
Gemma-3-4B-I	10.58	16.35	11.54	13.46	13.78	-3.20
Qwen3-8B	44.23	49.04	47.12	44.23	46.80	-2.57
Gemma-4-26B-A4B-I	35.58	35.58	25.00	34.62	31.73	+3.85
Olmo-3.1-32B-I	18.27	20.19	16.35	16.35	17.63	+0.64
Aya-Expanse-32B	22.12	16.35	22.12	12.50	16.99	+5.13
Llama-3.3-70B-I	29.81	30.77	33.65	25.00	29.81	+0.00
Qwen3.6-Plus	52.88	56.73	55.77	54.81	55.77	-2.89
DeepSeek-V3.1	20.19	21.15	15.38	18.27	18.27	+1.92
Command-A	19.23	16.35	24.04	19.23	19.87	-0.64
o4-mini	25.00	21.15	18.27	23.08	20.83	+4.17
GPT-5.2	63.46	62.50	64.42	64.42	63.78	-0.32
Gemini-3-Flash	33.65	31.73	22.12	25.96	26.60	+7.05
Claude Haiku 4.5	27.88	37.50	42.31	26.92	35.58	**-7.70**
Human (Base)	85.58	93.27	92.31	92.31	92.63	-7.05
Human (Expert)	**100.00**	**100.00**	**100.00**	**99.04**	**99.68**	+0.32

Table 88. Invariance stability — Sub-Type: *C4. Graph Isomorphism and Structural Mapping*.

Model	*Global Stability (%)* IMI↑	CIR↑	ICR↓	*Novel (%)* BSS↑	CE↓	*Pairwise Failure Dep. (%)* $P(\mathcal{I}_1{=}0\|\mathcal{I}_2{=}0)$	$P(\mathcal{I}_1{=}0\|\mathcal{I}_3{=}0)$	$P(\mathcal{I}_2{=}0\|\mathcal{I}_3{=}0)$
Llama-3.2-1B-I	0.96	12.50	3.85	82.35	13.81	94.85	93.81	94.85
Ministral-3B-2512	3.85	18.18	12.50	92.31	32.78	83.53	86.05	88.37
Gemma-3-4B-I	1.92	18.18	6.73	54.55	27.38	88.04	86.67	88.89
Qwen3-8B	32.69	73.91	4.81	82.76	23.89	83.64	81.03	84.48
Gemma-4-26B-A4B-I	7.69	21.62	13.46	90.57	44.43	71.79	79.41	80.88
Olmo-3.1-32B-I	3.85	21.05	9.62	78.95	31.46	82.76	85.06	86.21
Aya-Expanse-32B	2.88	13.04	15.38	**94.74**	32.55	87.65	89.01	81.32
Llama-3.3-70B-I	13.46	45.16	8.65	80.95	34.58	85.51	80.77	75.64
Qwen3.6-Plus	43.27	81.82	1.92	90.91	18.43	84.78	80.85	87.23
DeepSeek-V3.1	1.92	9.52	9.62	88.24	29.06	84.09	85.88	94.12
Command-A	3.85	20.00	5.77	78.05	32.53	91.14	90.48	79.76
o4-mini	5.77	23.08	13.46	94.74	32.36	85.88	86.25	90.00
GPT-5.2	51.92	81.82	3.85	80.00	16.51	86.49	91.89	86.49
Gemini-3-Flash	9.62	28.57	12.50	75.00	33.02	77.78	84.42	89.61
Claude Haiku 4.5	8.65	31.03	10.58	71.43	46.95	76.67	73.68	69.74
Human (Base)	67.31	78.65	2.88	88.24	27.25	0.00	12.50	25.00
Human (Expert)	**99.04**	**99.04**	**0.00**	0.00	**0.78**	0.00	0.00	0.00

Table 89. Novel metrics — Sub-Type: *C4. Graph Isomorphism and Structural Mapping*.

Model	*Bilateral Symmetry*			*Entropy*	*Invariant Recovery (%)*			*Fragility Gradient*	
	BSS↑	$F^+(\%)$	$F^-(\%)$	CE↓	IRR_{full} ↑	IRR_{part} ↑	IRR_{Δ}	FG	$\overline{g}$
Llama-3.2-1B-I	82.35	6.73	9.62	13.81	0.00	10.42	-87.50	**2**	+0.48
Ministral-3B-2512	92.31	17.31	20.19	32.78	1.22	25.61	-80.60	2	+1.44
Gemma-3-4B-I	54.55	8.65	23.08	27.38	0.00	25.81	-81.82	1	+1.44
Qwen3-8B	82.76	11.54	16.35	23.89	5.17	29.31	-20.92	2	+2.40
Gemma-4-26B-A4B-I	90.57	27.88	23.08	44.43	2.99	35.82	-75.39	1	+0.48
Olmo-3.1-32B-I	78.95	14.42	22.12	31.46	0.00	27.06	-78.95	2	+1.92
Aya-Expanse-32B	**94.74**	19.23	17.31	32.55	1.23	22.22	-85.73	1	+1.92
Llama-3.3-70B-I	80.95	16.35	24.04	34.58	0.00	34.25	-54.84	1	+2.88
Qwen3.6-Plus	90.91	9.62	11.54	18.43	4.08	24.49	-14.10	2	+0.96
DeepSeek-V3.1	88.24	18.27	14.42	29.06	2.41	18.07	-88.07	1	+1.44
Command-A	78.05	15.38	24.04	32.53	1.19	29.76	-78.81	1	-1.44
o4-mini	94.74	19.23	17.31	32.36	1.28	23.08	-75.64	1	-0.96
GPT-5.2	80.00	11.54	7.69	16.51	10.53	21.05	-7.65	1	-0.96
Gemini-3-Flash	75.00	24.04	14.42	33.02	2.90	21.74	-68.53	1	+2.89
Claude Haiku 4.5	71.43	19.23	34.62	46.95	8.00	48.00	-60.97	1	+5.29
Human (Base)	88.24	18.27	14.42	27.25	**93.33**	**100.00**	+71.98	2	+0.48
Human (Expert)	0.00	0.96	0.00	**0.78**	—	—	—	2	+0.48

Table 90. Sensitivity profile — Sub-Type: *C4. Graph Isomorphism and Structural Mapping*.

Model	*Cond. Inv. Retention (%)*			*Collapse Contribution (%)*			*Rel. Sensitivity (pp)*		
	TCIR_1	TCIR_2	TCIR_3	$\text{TCC}_1 \downarrow$	$\text{TCC}_2 \downarrow$	$\text{TCC}_3 \downarrow$	RIS_1	RIS_2	RIS_3
Llama-3.2-1B-I	44.44	38.89	38.89	71.43	71.43	64.29	+0.48	-0.48	-0.48
Ministral-3B-2512	48.84	44.19	41.86	57.58	66.67	75.76	+0.96	-0.96	-1.92
Gemma-3-4B-I	48.57	34.29	40.00	58.06	67.74	67.74	+3.37	-1.44	+0.48
Qwen3-8B	80.95	77.78	73.02	57.89	63.16	68.42	+2.89	+0.97	-1.92
Gemma-4-26B-A4B-I	60.66	42.62	59.02	61.11	66.67	63.89	+2.89	-7.69	+1.93
Olmo-3.1-32B-I	50.00	40.48	40.48	63.64	63.64	72.73	+2.40	-1.44	-1.44
Aya-Expanse-32B	41.46	56.10	31.71	69.70	48.48	81.82	-1.92	+3.85	-5.77
Llama-3.3-70B-I	57.14	62.50	46.43	70.59	52.94	73.53	+0.96	+3.84	-4.81
Qwen3.6-Plus	88.06	86.57	85.07	41.67	58.33	58.33	+1.68	+0.72	-0.24
DeepSeek-V3.1	61.11	44.44	52.78	43.48	78.26	60.87	+2.40	-3.37	-0.48
Command-A	37.78	55.56	44.44	86.67	56.67	66.67	-3.36	+4.33	-0.48
o4-mini	50.00	43.18	54.55	61.29	74.19	61.29	-0.73	-3.60	+1.21
GPT-5.2	87.84	90.54	90.54	87.50	37.50	62.50	-1.20	+0.72	+0.72
Gemini-3-Flash	66.00	46.00	54.00	53.85	73.08	84.62	+3.36	-6.25	-2.41
Claude Haiku 4.5	60.00	67.69	43.08	56.10	43.90	82.93	+3.85	+8.66	-6.73
Human (Base)	93.27	92.31	92.31	25.00	75.00	75.00	+2.40	+1.44	+1.44
Human (Expert)	**100.00**	**100.00**	**99.04**	**0.00**	**0.00**	**0.00**	+0.24	+0.24	-0.72

H.3.5. C5. LANGUAGE, SEMANTICS AND INFORMATION STRUCTURE-BASED REASONING

Table 91. Core accuracy — Sub-Type: *C5. Language, Semantics and Information Structure-based Reasoning*.

Model	Base	*Accuracy (%)* $\mathcal{I}_1$	$\mathcal{I}_2$	$\mathcal{I}_3$	*Mean Inv.* $\overline{\mathcal{I}}$	*Base*$-\overline{\mathcal{I}}$ Δ
Llama-3.2-1B-I	16.67	13.89	16.67	19.44	16.67	+0.00
Ministral-3B-2512	41.67	16.67	27.78	16.67	20.37	+21.30
Gemma-3-4B-I	36.11	30.56	27.78	22.22	26.85	+9.26
Qwen3-8B	50.00	44.44	36.11	33.33	37.96	+12.04
Gemma-4-26B-A4B-I	69.44	55.56	55.56	41.67	50.93	+18.51
Olmo-3.1-32B-I	44.44	36.11	25.00	16.67	25.93	+18.51
Aya-Expanse-32B	38.89	36.11	27.78	22.22	28.70	+10.19
Llama-3.3-70B-I	52.78	44.44	33.33	36.11	37.96	+14.82
Qwen3.6-Plus	75.00	61.11	55.56	44.44	53.70	+21.30
DeepSeek-V3.1	55.56	33.33	30.56	25.00	29.63	+25.93
Command-A	61.11	44.44	36.11	36.11	38.89	+22.22
o4-mini	66.67	38.89	36.11	27.78	34.26	+32.41
GPT-5.2	83.33	52.78	61.11	63.89	59.26	+24.07
Gemini-3-Flash	86.11	63.89	58.33	52.78	58.33	+27.78
Claude Haiku 4.5	69.44	55.56	38.89	36.11	43.52	+25.92
Human (Base)	88.89	88.89	91.67	91.67	90.74	**-1.85**
Human (Expert)	**100.00**	**100.00**	**100.00**	**97.22**	**99.07**	+0.93

Table 92. Invariance stability — Sub-Type: *C5. Language, Semantics and Information Structure-based Reasoning*.

Model	*Global Stability (%)* IMI↑	CIR↑	ICR↓	*Novel (%)* BSS↑	CE↓	*Pairwise Failure Dep. (%)* $P(\mathcal{I}_1{=}0\|\mathcal{I}_2{=}0)$	$P(\mathcal{I}_1{=}0\|\mathcal{I}_3{=}0)$	$P(\mathcal{I}_2{=}0\|\mathcal{I}_3{=}0)$
Llama-3.2-1B-I	5.56	33.33	5.56	**80.00**	23.58	90.00	96.55	89.66
Ministral-3B-2512	0.00	0.00	27.78	42.11	47.01	84.62	90.00	73.33
Gemma-3-4B-I	5.56	15.38	11.11	62.50	38.15	84.62	78.57	75.00
Qwen3-8B	11.11	22.22	13.89	52.63	44.91	73.91	75.00	66.67
Gemma-4-26B-A4B-I	11.11	16.00	19.44	38.46	63.83	62.50	42.86	38.10
Olmo-3.1-32B-I	2.78	6.25	19.44	57.14	48.90	70.37	70.00	76.67
Aya-Expanse-32B	5.56	14.29	11.11	58.82	40.41	80.77	71.43	71.43
Llama-3.3-70B-I	5.56	10.53	25.00	69.23	63.83	54.17	73.91	60.87
Qwen3.6-Plus	22.22	29.63	19.44	34.78	53.93	62.50	55.00	50.00
DeepSeek-V3.1	0.00	0.00	33.33	51.85	64.52	80.00	70.37	66.67
Command-A	13.89	22.73	27.78	45.45	53.77	73.91	65.22	69.57
o4-mini	2.78	4.17	36.11	29.63	65.04	73.91	69.23	57.69
GPT-5.2	27.78	33.33	22.22	26.09	55.50	71.43	53.85	38.46
Gemini-3-Flash	27.78	32.26	25.00	25.00	57.23	53.33	47.06	47.06
Claude Haiku 4.5	8.33	12.00	30.56	37.04	66.61	50.00	52.17	56.52
Human (Base)	69.44	78.12	2.78	72.73	26.36	0.00	33.33	0.00
Human (Expert)	**97.22**	**97.22**	**0.00**	0.00	**2.25**	0.00	0.00	0.00

Table 93. Novel metrics — Sub-Type: *C5. Language, Semantics and Information Structure-based Reasoning*.

Model	*Bilateral Symmetry*			*Entropy*	*Invariant Recovery (%)*			*Fragility Gradient*	
	BSS↑	$F^{+}(\%)$	$F^{-}(\%)$	CE↓	IRR_{full} ↑	IRR_{part} ↑	IRR_{Δ}	FG	$\overline{g}$
Llama-3.2-1B-I	**80.00**	11.11	16.67	23.58	0.00	20.00	-66.67	0	-2.77
Ministral-3B-2512	42.11	41.67	11.11	47.01	0.00	19.05	-100.00	1	+0.00
Gemma-3-4B-I	62.50	30.56	13.89	38.15	0.00	21.74	-84.62	**2**	+4.17
Qwen3-8B	52.63	38.89	13.89	44.91	5.56	27.78	-72.22	2	+5.56
Gemma-4-26B-A4B-I	38.46	58.33	13.89	63.83	0.00	45.45	-84.00	2	+6.94
Olmo-3.1-32B-I	57.14	41.67	16.67	48.90	0.00	30.00	-93.75	2	+9.72
Aya-Expanse-32B	58.82	33.33	13.89	40.41	0.00	22.73	-85.71	2	+6.94
Llama-3.3-70B-I	69.23	47.22	25.00	63.83	0.00	52.94	-89.47	1	+4.17
Qwen3.6-Plus	34.78	52.78	11.11	53.93	11.11	44.44	-59.26	2	+8.33
DeepSeek-V3.1	51.85	55.56	19.44	64.52	6.25	43.75	-93.75	2	+4.16
Command-A	45.45	47.22	13.89	53.77	0.00	35.71	-77.27	2	+4.16
o4-mini	29.63	63.89	11.11	65.04	8.33	33.33	-87.50	2	+5.56
GPT-5.2	26.09	55.56	8.33	55.50	0.00	50.00	-66.67	0	-5.56
Gemini-3-Flash	25.00	58.33	8.33	57.23	0.00	60.00	-67.74	2	+5.56
Claude Haiku 4.5	37.04	61.11	13.89	66.61	0.00	45.45	-88.00	2	+9.72
Human (Base)	72.73	19.44	11.11	26.36	**50.00**	**100.00**	+28.12	1	-1.39
Human (Expert)	0.00	2.78	0.00	**2.25**	—	—	—	2	+1.39

Table 94. Sensitivity profile — Sub-Type: *C5. Language, Semantics and Information Structure-based Reasoning*.

Model	*Cond. Inv. Retention (%)*			*Collapse Contribution (%)*			*Rel. Sensitivity (pp)*		
	TCIR_1	TCIR_2	TCIR_3	$\text{TCC}_1 \downarrow$	$\text{TCC}_2 \downarrow$	$\text{TCC}_3 \downarrow$	RIS_1	RIS_2	RIS_3
Llama-3.2-1B-I	41.67	50.00	58.33	87.50	50.00	62.50	-2.78	+0.00	+2.77
Ministral-3B-2512	31.58	52.63	31.58	78.57	50.00	85.71	-9.02	+2.09	-9.02
Gemma-3-4B-I	61.11	55.56	44.44	77.78	55.56	66.67	+1.39	-1.39	-6.95
Qwen3-8B	69.57	56.52	52.17	77.78	66.67	66.67	+3.47	-4.86	-7.64
Gemma-4-26B-A4B-I	66.67	66.67	50.00	75.00	58.33	41.67	+0.00	+0.00	-13.89
Olmo-3.1-32B-I	59.09	40.91	27.27	61.54	76.92	92.31	+5.55	-5.56	-13.89
Aya-Expanse-32B	68.42	52.63	42.11	66.67	66.67	66.67	+4.86	-3.47	-9.03
Llama-3.3-70B-I	57.14	42.86	46.43	66.67	55.56	72.22	+2.77	-8.34	-5.56
Qwen3.6-Plus	70.97	64.52	51.61	80.00	70.00	80.00	+2.08	-3.47	-14.59
DeepSeek-V3.1	44.44	40.74	33.33	77.78	72.22	77.78	-2.78	-5.55	-11.11
Command-A	59.26	48.15	48.15	73.33	73.33	66.67	+-0.00	-8.33	-8.33
o4-mini	50.00	46.43	35.71	87.50	68.75	75.00	-3.47	-6.25	-14.58
GPT-5.2	57.58	66.67	69.70	90.91	72.73	45.45	-12.50	-4.17	-1.39
Gemini-3-Flash	67.65	61.76	55.88	75.00	75.00	75.00	-1.39	-6.95	-12.50
Claude Haiku 4.5	66.67	46.67	43.33	56.25	68.75	75.00	+5.56	-11.11	-13.89
Human (Base)	88.89	91.67	91.67	66.67	33.33	33.33	-1.39	+1.39	+1.39
Human (Expert)	**100.00**	**100.00**	**97.22**	**0.00**	**0.00**	**0.00**	+0.69	+0.69	-2.09

I. Error and Reasoning Failure Mode Analysis

I.1. Setup and aggregate behavior

We analysed every response of seven contemporary instruction-tuned language models on a 1,200-item reasoning benchmark covering three task families — *A. Formal Logic & Discrete Structure*, *B. Quantitative Reasoning & Optimization*, and *C. Relational, Planning, Algorithmic and Semantic Reasoning* — each presented in four equivalent surface forms (a `base` item plus three semantically-preserving rephrasings, `inv1`, `inv2`, `inv3`), giving 4,800 scored question–response pairs per model. The seven models studied are CLAUDE-HAIKU-4.5, DEEPSEEK-V3.1, GPT-5, GEMMA-4-26B, LLAMA-3.3-70B, OLMO-3.1-32B, and QWEN3.6-PLUS. Aggregate failure counts are summarised in Table 95.

Model	Failures / 4800	Failure rate
GPT-5	1501	31.3%
QWEN3.6-PLUS	1597	33.3%
CLAUDE-HAIKU-4.5	2463	51.3%
GEMMA-4-26B	3496	72.8%
LLAMA-3.3-70B	3595	74.9%
DEEPSEEK-V3.1	3941	82.1%
OLMO-3.1-32B	4052	84.4%

Table 95. Aggregate failure counts per model across all 4,800 items.

We performed a two-pass qualitative coding of these failures, stratified by model, task sub-type, and invariant axis, and supplemented this with corpus-level lexical analysis of the rubric judges' reasoning notes. The result is the eight-category, 48-subcategory taxonomy summarised in this section. The eight top-level categories partition the space of failure *loci* along the reasoning pipeline: input parsing (E1), formalisation (E2), local computation (E3), chain-coherence (E4), invariance robustness (E5), lexical-semantic selection (E6), output specification (E7), and generation pathology (E8). The taxonomy is intentionally pipeline-aligned: each category corresponds to a stage where a model can fail *even if all preceding stages succeeded*.

I.2. E1. Premise-comprehension failure

E1 failures arise when the model does not correctly internalize the problem statement before reasoning begins. In such cases, constraints are dropped, quantifiers are misinterpreted, the question target is misidentified, or surface-level artifacts such as premise order or distractors distort the internal representation. Since all subsequent reasoning is conditioned on this initial encoding, even small E1 errors propagate and lead to confidently incorrect final answers.

Why it happens. We identify two dominant mechanisms. First, *long-context attention dilution*, where load-bearing constraints are overshadowed in prompts containing many structurally similar tokens (e.g., repeated rates, quantities, or conditions), causing critical modifiers to be weakly encoded in the working representation. Second, *prior-driven solution substitution*, where the model recognizes a familiar problem template and applies a memorized solution pattern while ignoring problem-specific modifications.

Subcategories.

E1.1 Dropped or ignored constraint. A semantically critical constraint is omitted during internalization, leading to a systematically incorrect solution. EXAMPLE. Item `e8014b69456947d6` (`B1. Multi-Stage Algebraic Systems`) describes three workers with pairwise combined rates ($A{+}B{=}28$, $B{+}C{=}26$, $A{+}C{=}26$) operating for different durations ($3, 2, 3$ hours), with an additional constraint that *worker A loses 2 units/hour due to overhead*. The correct solution is 100. However, DEEPSEEK-V3.1, OLMO-3.1-32B, and GEMMA-4-26B produce 36, 58, and 70 respectively, indicating complete omission of the loss term.

E1.2 Quantifier / scope misreading. The model incorrectly interprets logical scope, particularly for universal and existential quantifiers. EXAMPLE. Item `ec5b349677f64d38` requires formalizing the statement "Everyone has exactly one father," whose correct representation is $\forall y\, \exists! x\, F(x, y)$. CLAUDE-HAIKU-4.5 instead produces $\forall x\, \exists! y\, F(y, x)$, preserving quantifier structure but swapping variable roles, thereby altering the intended meaning.

E1.3 Goal / question-target misidentification. The model answers a related but incorrect target within the same problem. EXAMPLE. Item `b1d1239a08fe4fbe` (`B5. Matrix/Linear Algebra`) asks which property must hold for a 4×4 matrix A satisfying $A^3=0$. The correct answer is that all eigenvalues are 0. CLAUDE-HAIKU-4.5 instead states that A is not invertible, which is true but not the queried property.

E1.4 Distractor capture. Information introduced as distractors is incorrectly treated as constraint-bearing, leading to corrupted reasoning. This failure mode is most pronounced in OLMO-3.1-32B (11.5% breakage), LLAMA-3.3-70B (14.0%), and GEMMA-4-26B (14.0%).

E1.5 Premise-order brittleness. We observe systematic failure under premise reordering, where logically equivalent problems lead to inconsistent outputs. Per-model breakage rates are: CLAUDE-HAIKU-4.5 (19.6%), OLMO-3.1-32B (10.6%), and LLAMA-3.3-70B (10.6%). EXAMPLE. For item `16aa090e103e4f85`, the *base* version asks for the number of paths from S to T in a directed acyclic graph, and CLAUDE-HAIKU-4.5 correctly outputs 4. However, under invariant transformation `inv2`, where the same edges are presented in a permuted order (e.g., $D\to G$, $S\to B$, $I\to T$, $A\to C$, etc.), the model fails to preserve equivalence and incorrectly concludes that there are no valid paths, reasoning that "nothing points to I ... there are 0 total paths." This illustrates sensitivity to superficial ordering despite unchanged underlying structure.

E1.6 Implicit-assumption injection. The model introduces unstated or default assumptions that override explicitly provided values in the problem statement. EXAMPLE. Item `85cf10cde3b247d5` (*base*) considers a vertical circular motion setup with $m=4\,\mathrm{kg}$, $r=1.5\,\mathrm{m}$, and $v_{\text{top}}=6\,\mathrm{m/s}$, asking for the tension at the bottom of the circle. The correct solution is $292\,\mathrm{N}$ using $g=9.81\,\mathrm{m/s^2}$. However, CLAUDE-HAIKU-4.5 replaces the specified gravitational constant with $g=10$, yielding $296\,\mathrm{N}$, while also noting an alternative value of $\sim 300\,\mathrm{N}$ for $g=9.8$. This indicates awareness of the correct convention but a failure to consistently adhere to the problem-specified physical parameters.

I.3. E2. Symbolic / logical translation failure

This category captures errors at the natural-language $\to$ formalism interface, where the model has generally understood the underlying problem but fails to express the solution in the required formal language, or produces a representation that is structurally close yet semantically or syntactically misaligned with the expected form.

Why it happens. A key source of failure is the mismatch between sparsely represented formal logics (first-order, modal, epistemic) and richly represented natural-language reasoning. In addition, rubric constraints enforce canonical conventions (e.g., predicate signatures, quantifier ordering, normal forms) that are not mathematically necessary for correctness but are required for evaluation consistency. As a result, even logically equivalent answers may be marked incorrect. A frequent manifestation is *translation refusal*, where the model retains correct semantics but avoids formal notation and instead responds in natural language.

Subcategories.

E2.1 Variable role swap. Incorrect alignment of argument order, producing expressions such as $F(y,x)$ instead of $F(x,y)$. This is particularly prevalent in A1-style logical tasks across DEEPSEEK-V3.1, OLMO-3.1-32B, and LLAMA-3.3-70B.

E2.2 Quantifier-order / type confusion. Errors involving $\forall/\exists$ scope or ordering, such as confusing $\forall x\exists y$ with $\exists y\forall x$. EXAMPLE. Item `0bcf8d391e1a4985`: "Every driver is certified for at least one route category." The correct formalisation is $\forall d\exists r(\mathrm{Driver}(d) \to \mathrm{Certified}(d,r))$. CLAUDE-HAIKU-4.5 instead produces $\forall x(\mathrm{Driver}(x) \to \exists y(\mathrm{Certified}(x,y) \land \mathrm{RouteCategory}(y)))$, which is semantically close but introduces an unintended predicate constraint, leading to rubric mismatch.

E2.3 Refusal-to-formalise / natural-language fallback. The model produces a correct paraphrase but fails to emit the required formal expression. EXAMPLE. Item `0b080fc57e474451` (`inv1`, A1): instead of $\neg\exists a\,\forall m\,(\mathrm{Analyst}(a) \land \mathrm{Covered}(a,m))$, CLAUDE-HAIKU-4.5 answers: "Every analyst failed to cover at least one market." The semantics are correct, but the required formal representation is omitted.

E2.4 Predicate / signature drift. Introduction of extra predicates or decomposition of a single relation into multiple components, yielding structurally inconsistent but semantically equivalent representations that violate the expected schema.

E2.5 Modal / epistemic operator mishandling. Failures involving $\Box$, $\Diamond$, and epistemic operators such as K_A, K_B. EXAMPLE. Item `db3546f3eeae45a4` (A5): given a set of knowledge constraints, the correct conclusion is inconsistency. CLAUDE-HAIKU-4.5 instead derives $K_A(K_B(P)) \wedge K_A(K_B(\neg Q))$, incorrectly treating distributive knowledge across agents as valid, violating standard epistemic closure principles.

E2.6 Logical-equivalence rejection. The model produces a logically equivalent form in a different representation (e.g., DNF vs. CNF) or a structurally valid but rubric-misaligned classification. EXAMPLE. Item `bb403d5df2cc4c51`: the canonical answer identifies both graphs as isomorphic to the Petersen graph, but CLAUDE-HAIKU-4.5 returns "No," indicating failure to recognize equivalence under the required canonical form.

I.4. E3. Computation & arithmetic failure

This category captures cases where the high-level reasoning structure is correct, but a local computational step is executed incorrectly. Typical errors include linear-system miscalculation, sign inversion, off-by-one mistakes, incorrect intermediate aggregation, and unit inconsistencies. These failures are generally traceable from the model's intermediate steps and do not require deep semantic interpretation to diagnose.

Why it happens. In current models, arithmetic is not performed via explicit algorithmic execution but through learned pattern completion over latent representations. As a result, intermediate quantities beyond a few significant digits are unstable, and errors accumulate through propagation rather than discrete computation. Additionally, sign and unit inconsistencies arise when the model conflates multiple valid conventions (e.g., coordinate direction, physics sign conventions, or unit scaling), leading to internally coherent but externally incorrect outputs.

Subcategories.

E3.1 System-of-equations solving error. Errors in solving linear systems, particularly prevalent in B1 and B4-style constraint problems involving bundled or coupled equations. EXAMPLE. Item `ae7024cda8854d76` (*B1 base*): solving $2t + c + 3h = 53$, $t + 3c + 2h = 68$, $3t + 2c + h = 59$ yields the correct solution $t = 13, c = 14, h = 9$. LLAMA-3.3-70B instead produces $a = 20, b = 12, c = 4$ and computes a final value of 90 instead of the correct 102, due to a sign error during elimination.

E3.2 Sign / direction error. Incorrect handling of directional or signed quantities, often triggered by representation changes. EXAMPLE. Item `3eb72f494b0c4741` (*C2, lens problem*, `inv2` encoding): the model correctly solves the base form as $v = 100$ cm, but under tabular encoding incorrectly applies $m = -v/u \Rightarrow v = -3u$, resulting in $v = -50$ cm. The transformation preserves structure but flips physical sign convention.

E3.3 Off-by-one / boundary error. Errors in discrete counting or indexing, especially in A2-style cardinality and constraint enumeration tasks. These arise from incomplete boundary tracking during iterative counting.

E3.4 Combinatorial mis-enumeration. Incorrect counting of valid assignments in logical or constraint-satisfaction problems. EXAMPLE. Item `8dd9b134cb41411a` (`A4`): the correct number of satisfying assignments for $(p \vee q) \wedge (\neg p \vee r) \wedge (\neg q \vee \neg r)$ is 3. CLAUDE-HAIKU-4.5 correctly enumerates all 8 cases but drops one valid configuration during verification, yielding 2 instead of 3.

E3.5 Floating-point / precision drift. Loss of numerical fidelity due to intermediate rounding. EXAMPLE. Item `39a88038785c42a8`: the correct value is 0.737, but CLAUDE-HAIKU-4.5 returns 0.74, which is rejected by the rubric due to insufficient precision despite being numerically close.

E3.6 Unit / dimensional inconsistency. Mixing or misalignment of measurement units (e.g., cm vs. m, percent vs. decimal form), or inconsistent use of physical constants such as $g = 9.81$ vs. $g = 10$. This failure mode is widespread across all models, with only minor variation in severity.

I.5. E4. Multi-step reasoning / chain-coherence failure

This category captures failures where individual intermediate steps are locally correct, but the global reasoning trajectory becomes inconsistent. The model fails to maintain a stable latent state across multiple inference steps, causing later operations to violate earlier constraints or overwrite previously correct deductions. As a result, intermediate verifications are either ignored or actively contradicted in subsequent reasoning.

Why it happens. In long chain-of-thought generation, models progressively shift attention from the original input to their own recently generated tokens. This leads to a drift in the effective working memory, where earlier constraints are no longer reliably retained. The failure is particularly pronounced in problems involving conditional logic, where the model must simultaneously track a running computation and a gating condition (e.g., thresholds, discount rules, or validity checks). When these two streams are not jointly maintained, later reasoning steps override earlier correct state estimates.

Subcategories.

E4.1 State-tracking loss in graph traversal. Failure to maintain correct dynamic programming state across nodes in graph or path-counting problems. EXAMPLE. Item `24d0727b6d73472a` (*C3 base*): the correct number of paths from S to T is 5. CLAUDE-HAIKU-4.5 correctly computes intermediate paths up to node E (yielding 2 paths) but incorrectly re-binds state at node F, double-counting a path via $D \to E \to F$, and outputs 8.

E4.2 Conditional-rule sequencing error. Incorrect ordering or application of constraints that depend on prior conditions. EXAMPLE. Item `9ea2e088806a4b4c` (*B2 base*): a discount applies only when quantity ≥ 7. Although CLAUDE-HAIKU-4.5 correctly solves the system and identifies a consistency check failure, it proceeds by overriding the condition and returns 196 instead of the correct 164.

E4.3 Bracket/tier transition error. Failures in maintaining correct segment boundaries in piecewise or tiered systems (e.g., tax brackets or pricing tiers). EXAMPLE. Item `907c047889ef4c64` (*C4*): correct computation yields 75.50. The model initially produces this value, then re-segments the brackets mid-response and outputs 82.50, before finally stating 65.50 as the reported answer, indicating unstable tier tracking.

E4.4 Self-contradiction / mid-chain reversal. The model explicitly revises a correct intermediate step without justification, leading to degraded final outputs. This pattern is frequently observed in CLAUDE-HAIKU-4.5, with 26 documented cases where a correct computation is followed by "Wait, let me recalculate," resulting in a worse final answer.

E4.5 Verification-failure ignored. The model correctly identifies an inconsistency during self-checking but proceeds regardless. EXAMPLE. A computed equality check shows $2(14) + 21 + 3(11) = 82 \neq 74$, yet the model continues with the incorrect branch, discarding its own verification signal. This reflects a failure in using verification as a control mechanism rather than generation-only behavior.

E4.6 Unjustified expert-name shortcut. The model prematurely assigns a known structure without full structural verification. EXAMPLE. Item `d9bd5ae675384c7c` (*A3 base*): after correctly verifying properties consistent with multiple candidate graphs (including 3-regularity and edge count), CLAUDE-HAIKU-4.5 asserts the graph is the Möbius–Kantor graph or similar, and concludes equivalence with H, despite the canonical solution identifying it as the cuboctahedral graph with distinct structural properties.

I.6. E5. Invariance brittleness

This is the defining phenomenon studied in the benchmark: a model produces different answers for semantically equivalent reformulations of the same underlying problem. As shown in Table 96, no model is fully invariant. Even the strongest system, GPT-5, exhibits a non-trivial loss of approximately 7–12% of base-correct answers under at least one invariance axis, indicating that correctness is not stable under representation-preserving transformations.

Why it happens. The failure arises from the interaction of two mechanisms. First, *shallow surface anchoring*, where the model conditions its reasoning path on lexical patterns that change under paraphrase, leading to a different implicit solution strategy even when the underlying problem is unchanged. Second, *role-by-position binding*, where renamed or reordered entities are re-associated based on surface order rather than structural constraints, causing systematic role misalignment under transformations such as entity relabeling or permutation.

Invariance axis	Breakage rate (base pass → inv fail)					
	GPT-5	Qwen	Claude	Llama	Gemma	Olmo
Abstraction Level Shift	11.6	14.9	21.0	14.9	15.2	14.6
Distractor Inj./Removal	10.8	9.1	14.3	14.0	14.0	11.5
Entity Relabeling	9.2	7.2	14.5	14.1	8.2	6.9
Equiv. Logical Reform.	7.4	7.1	13.0	11.8	16.7	8.6
Equiv. Math. Reform.	10.7	12.3	17.6	10.2	11.9	6.1
Perspective / Role	8.4	10.1	19.4	13.2	14.1	8.8
Premise Reordering	9.7	9.7	19.6	10.6	9.0	10.6
Redundant Restatement	3.1	5.6	10.1	7.0	8.1	6.1
Repr./Encoding Swap	11.9	11.5	21.1	10.1	8.8	7.0
Style/Register	9.8	8.1	14.0	9.8	13.3	8.8
Synonym/Paraphrase	8.8	10.0	15.8	9.6	11.9	11.2
Variable Renaming	9.4	9.4	14.5	11.4	11.4	10.0

Table 96. Breakage rate (%): base item passed but the corresponding invariance variant failed. DEEPSEEK-V3.1 is omitted from the table because its "base pass" set is small (n=859) and rates fluctuate, but its breakage profile is qualitatively similar to OLMO-3.1-32B. CLAUDE-HAIKU-4.5 is highest along nearly every axis — a function of its high *base* pass rate creating more opportunities to break, not necessarily of more fragile reasoning.

Subcategories.

E5.1 Premise-reordering brittleness. Sensitivity to clause or premise permutation without any semantic change. Up to 19.6% of base-correct answers in CLAUDE-HAIKU-4.5 are lost under simple reordering. EXAMPLE. Item `16aa090e103e4f85` (see E1.5), where permutation of identical constraints changes the model's conclusion despite identical logical content.

E5.2 Entity-relabeling brittleness. Failure under renaming of objects, agents, or variables while preserving structure. EXAMPLE. Item `464ad049f9aa427d` (*B2*): the same bundle-pricing problem under `inv2` causes CLAUDE-HAIKU-4.5 to flip from 172 (correct) to 178, purely due to entity renaming.

E5.3 Variable-renaming brittleness. Sensitivity to purely symbolic renaming (e.g., $A, B, C \rightarrow P, Q, R$ or $n \rightarrow m$). Most affected models are CLAUDE-HAIKU-4.5 (14.5%), DEEPSEEK-V3.1 (12.5%), and LLAMA-3.3-70B (11.4%), indicating weak abstraction over variable identity.

E5.4 Abstraction-level shift brittleness. The most widespread invariance failure, occurring when the same problem is expressed at different abstraction levels (e.g., narrative ↔ formal notation). Across models, 11.6%–21.0% of base-correct solutions break under this shift, with CLAUDE-HAIKU-4.5 showing the highest sensitivity (21.0%).

E5.5 Representation / encoding swap. Failure under conversion between equivalent representations (e.g., prose to table or equation form). This effect is especially pronounced when structural cues are redistributed across formats, as illustrated in the lens problem example in E3.2.

E5.6 Perspective / role reframing. Instability under changes in agent perspective or role assignment, where constraints remain identical but viewpoint changes. Error rates reach 19.4% for CLAUDE-HAIKU-4.5 and 14.1% for GEMMA-4-26B, indicating inconsistent role grounding under reframing.

I.7. E6. Lexical / semantic selection failure

This category captures failures on reading-comprehension and word-choice tasks (primarily C5), where the model selects a near-neighbour of the correct answer that is semantically plausible but not rubric-acceptable. The error manifests as substitution with a synonym, antonym, thematic associate, incorrect analytical framing, or mismatched register. This failure mode dominates performance degradation on C5 across all evaluated models.

Why it happens. Many SAT-style and cloze-form items admit multiple semantically related candidates, but only one is pragmatically licensed under the rubric. Models trained on large corpora tend to prefer high-likelihood lexical completions, which often correspond to frequent synonyms (e.g., "reflected in", "consistent with") rather than the precise target phrase (e.g., "manifest in", "co-extensive with"). This bias is amplified in negation-heavy or polarity-sensitive contexts, where embedded logical structure is not reliably tracked. As a result, models optimize for fluency rather than rubric-constrained semantic precision.

Subcategories.

E6.1 Synonym substitution. Replacement of the target phrase with a semantically close but non-equivalent synonym. EXAMPLE. Item `92ecc3eefc044b01`: "Morris's repudiation of industrialization is ___ the Kelmscott editions' use of handmade materials." Correct answer: "manifest in". OLMO-3.1-32B answers "consistent with", while CLAUDE-HAIKU-4.5 answers "reflected in", both failing rubric specificity.

E6.2 Antonym / polarity inversion. Selection of a semantically opposite term due to polarity misinterpretation. EXAMPLE. Item `b1f7163280af43eb`: correct answer is "homogeneous". OLMO-3.1-32B returns "heterogeneous", inverting the intended density condition and violating the underlying inference.

E6.3 Wrong-axis inference. Correctly understanding the domain but selecting an incorrect abstraction level or response type (e.g., instance vs. category). EXAMPLE. Item `5efcb6c6d66a4fe2`: the correct classification is "Free from". LLAMA-3.3-70B instead provides a descriptive statement ("contains no allergen X"), which is semantically related but fails to match the required categorical output.

E6.4 Document-cross-reference failure. Inability to integrate constraints across multiple sources or passages. EXAMPLE. Item `6d5f1072770543d9`: combining Doc1 and Doc2 yields a correct total of 52%, implying disagreement with Doc2 alone. OLMO-3.1-32B answers "Yes" by relying solely on Doc2, ignoring cross-document reconciliation.

E6.5 Idiomatic / register mismatch. Selection of an answer with incorrect stylistic or register alignment (e.g., informal vs. formal usage), even when semantically similar to the correct option.

E6.6 Pragmatic-cue blindness. Failure to incorporate discourse-level qualifiers such as "typically", "independently", or "with little in common", leading to systematic overgeneralization of the candidate answer space.

I.8. E7. Format / specification compliance failure

In this class of errors, the model often arrives at a correct or near-correct solution, but fails at the level of output specification. The final answer is rejected not due to reasoning mistakes, but due to violations of formatting, structural, or response-contract constraints imposed by the benchmark rubric.

Why it happens. This failure mode arises from two main sources. First, many evaluation rubrics operate as strict-match or format-sensitive graders, where mathematically equivalent expressions may still be marked incorrect if they deviate from the expected representation. Second, models carry strong default output priors (e.g., simplifying fractions, rounding to fixed precision, or preferring natural language explanations) that may conflict with dataset-specific output requirements, leading to mismatches even when the underlying reasoning is sound.

Subcategories.

E7.1 Decimal vs. fraction mismatch. EXAMPLE. Item `043c9dee9e8b436b`: computing $P(\text{div by 8} \mid \text{div by 4})$ for $n \in [1, 100]$. The correct value is 0.48 (equivalently $12/25$). GPT-5 returns $12/25$, which is mathematically identical but depends on whether the judge accepts fractional form equivalence, highlighting sensitivity to grading policy.

E7.2 Multi-part answer incompleteness. Item `ff22b72bc81e4cba` requires both the number of defective items and the identification of which retailer reports first. Several models correctly solve one component but omit the second, resulting in partial but invalid submissions under the rubric.

E7.3 Spurious extra element. EXAMPLE. Item `f002a3f1e73c4b08`, symmetric difference of $((L \cup M) \cap N)$ and O. The correct answer is $\{1, 4, 5, 6, 7\}$. CLAUDE-HAIKU-4.5 outputs $\{1, 4, 5, 6, 7, 12\}$, propagating an extraneous element from an intermediate step without re-verification.

E7.4 Notation / format drift. Structural equivalence holds, but notation differs from the required canonical form (e.g., determinant expansions written in permuted index order rather than the expected Levi-Civita expansion format), leading to rejection despite correctness.

E7.5 Refusal / "insufficient information" on solvable items. EXAMPLE. Item `0fa6761d40d942d6` (*C2*, heel-diameter problem). CLAUDE-HAIKU-4.5 responds that the heel diameter cannot be determined, citing missing geometric constraints, despite the problem being fully specified. This reflects overcautious abstention rather than true underdetermination.

I.9. E8. Generation / output pathology

The output trajectory itself fails. The model may loop, garble, truncate, abandon a recalculation, or return no usable response at all. These failures are distinct from classical reasoning errors: in these cases, the model does not produce a complete, well-formed answer that can even be meaningfully evaluated by the rubric.

Why it happens. These are primarily decoding and generation-stability pathologies rather than reasoning defects. Token-level corruption in LLAMA-3.3-70B on long structured prompts (e.g., bundle-pricing or multi-constraint algebra items) suggests entropy collapse under high-complexity contexts. Phrase-level loops in CLAUDE-HAIKU-4.5 (e.g., item `d940bd577e104035`: "the 1-UIP is $(\neg Z \vee \neg A \vee \neg B)$ is wrong, the 1-UIP is $(\neg Z \vee \neg A \vee \neg B)$ is wrong...") indicate a failure mode where the sampler remains trapped in a locally high-probability repetition cycle after the model has internally lost a stable continuation path. Mid-step truncation, in contrast, reflects hard token-budget constraints during aggressive enumerations (e.g., item `83994f0c8a9643d8`: "$k = 0 : x = 23$, $k = 1 : x = 64$, ..."), where the output is structurally correct but cut off before completion.

Subcategories.

E8.1 Token-level garble. Concentrated in LLAMA-3.3-70B, with 24 visible instances primarily on B1/B2 items involving multilingual interference and escape-character corruption. EXAMPLE. Item `e8014b69456947d6` produces an incoherent mixed-token output of the form *"201 Usageak qeller\\\\\\\cp:* experiment PART... "* in LLAMA-3.3-70B, indicating breakdown at the token-stream level rather than semantic reasoning.

E8.2 Phrase-level loop. Characterized by repetitive or cyclic re-generation of partial phrases without progression toward a final answer. This failure is most prevalent in CLAUDE-HAIKU-4.5 (77 visible cases) and also appears in LLAMA-3.3-70B (33 cases), typically manifesting as repeated restatements of intermediate reasoning steps.

E8.3 Abandoned recalculation. The model initiates a corrective step (e.g., "Wait, let me recalculate") but fails to complete the recomputation or output a final result. This is strongly model-specific to CLAUDE-HAIKU-4.5, with 26 visible cases, and reflects incomplete recovery after self-detected inconsistencies.

E8.4 Empty / null output. Rare failure mode in which the model produces either an empty response or a structurally null completion. Observed sporadically in both CLAUDE-HAIKU-4.5 and LLAMA-3.3-70B, typically under high-context or multi-step constraint-heavy items.

I.10. Overall Observations

Reasoning-frontier vs. mid-tier behaviour diverges qualitatively. The two frontier models in our evaluation, GPT-5 and QWEN3.6-PLUS, predominantly fail along E5 (invariance brittleness) and E6 (lexical selection). In other words, their failures are largely *post-reasoning distortions*: they generally reach the correct solution in their native representation, but lose accuracy under surface-form changes or fine-grained lexical constraints. By contrast, mid-tier models such as CLAUDE-HAIKU-4.5 exhibit a different failure profile dominated by E1, E4, and E8, where errors emerge earlier in the reasoning trajectory. These include premature constraint loss (E1), unstable multi-step coherence (E4), and generation-level breakdowns such as recalc-abandon and looping (E8.4). Lower-tier models (GEMMA-4-26B, LLAMA-3.3-70B, DEEPSEEK-V3.1, OLMO-3.1-32B) are primarily governed by E1 and E3 failures, indicating that they often fail at constraint integration and basic computation before higher-order reasoning effects become visible.

The most consistent invariance break is abstraction shift. Across all evaluated models, *Abstraction Level Shift* remains the most systematically damaging transformation, with breakage rates consistently in the 11–21% range. This exceeds all other invariance axes and holds even for top-performing systems. The consistency of this effect provides strong evidence that the benchmark is not probing memorization artifacts, but rather genuine representational robustness: models that correctly solve a problem in one representational form frequently fail when the same underlying structure is expressed at a different level of abstraction.

The most informative single test is the bundle-pricing problem under entity relabeling. A canonical 3-by-3 linear system with an embedded conditional discount, when combined with entity relabeling between `base` and `inv` variants, jointly stresses multiple failure modes simultaneously. This single sub-type (B1/B2/B4) activates E1.1 (dropped constraints), E2.1 (variable re-binding), E3.1 (linear-system computation error), E4.2 (rule sequencing failure), and E5.2 (entity-relabeling brittleness). Empirically, performance on this task cluster correlates nearly monotonically with overall model ranking, making it one of the strongest single predictors of aggregate invariance robustness in the benchmark.

I.11. Root Cause Analysis and Mitigation Strategies

Across the eight error families identified in INVARIANTBENCH, we observe that failures are not isolated artifacts but emerge from a small set of recurring systemic weaknesses: (i) fragile input representation binding, (ii) weak cross-form consistency, (iii) limited multi-step state persistence, (iv) over-reliance on surface priors, and (v) decoding instability under structured constraints. Below we outline targeted mitigation strategies that directly correspond to these failure modes, rather than treating errors as independent phenomena.

RC1. Explicit invariant training across surface forms. A primary driver of E5 (invariance brittleness) is the lack of exposure to semantically identical but structurally diverse inputs. Training pipelines should enforce multi-view supervision, where each problem is paired with all invariant transformations during both supervised and reinforcement phases. This reduces reliance on surface anchors and encourages representation-level encoding. Empirically, this directly targets abstraction shift and entity-relabeling failures. It is most effective when invariants are treated as first-class training objects rather than augmentation artifacts.

RC2. Constraint-aware attention regularization. E1 failures arise from dropped or ignored constraints. A mitigation is to explicitly regularize attention mass toward constraint-bearing tokens using auxiliary supervision signals or gradient-based attribution alignment. This ensures that numerically or logically critical modifiers are not diluted in long contexts. Such mechanisms reduce long-context thinning and improve robustness in multi-constraint algebraic and logic tasks. This is especially relevant for E1.1 and E1.2 subcategories.

RC3. Structured state tracking for multi-step reasoning. E4 failures indicate breakdowns in global state persistence. Introducing an explicit scratchpad or latent state memory that is updated deterministically at each reasoning step can mitigate state-tracking loss. Unlike free-form chain-of-thought, structured intermediate representations enforce consistency checks across steps. This reduces errors in graph traversal, conditional logic, and recursive dependency tasks. It directly addresses E4.1 and E4.4 failure modes.

RC4. Decoding-time constraint enforcement. Many E7 and E8 failures are not reasoning errors but decoding violations. Introducing constrained decoding or post-generation validation layers can enforce format compliance, completeness, and termination conditions. For example, verifying that all required fields are present or that numerical outputs satisfy constraints before finalization. This reduces format drift, empty outputs, and mid-chain truncation without modifying the underlying model.

RC5. Adversarial paraphrase and relabel training. To address E5.1–E5.3, models should be trained on adversarially generated paraphrases, including entity renaming, variable swaps, and structural reordering. This forces abstraction over symbol identity and reduces role-by-position binding. The key objective is to ensure that semantic roles are grounded in relational structure rather than lexical position. This approach is particularly effective for linear systems and logic-based tasks.

RC6. Dual-path verification mechanisms. E3 and E4 failures often persist because intermediate verification is ignored. A dual-path architecture, where one branch performs reasoning and another independently verifies intermediate results, can reduce arithmetic and logical drift. Disagreement triggers recomputation rather than continuation. This mechanism directly targets system-of-equations errors and combinatorial miscounting.

RC7. Calibration of uncertainty-aware abstention. E7.5 and parts of E6 indicate overconfident responses under ambiguity. Introducing calibrated uncertainty estimation allows models to abstain or request clarification rather than hallucinate or overcommit. This reduces incorrect categorical selection and improper completion of underdetermined tasks. Importantly, calibration should be task-conditioned rather than global to avoid excessive refusal rates.

RC8. Token-level entropy stabilization. E8.1 and E8.2 suggest instability in decoding dynamics under long structured inputs. Techniques such as entropy penalties, repetition suppression, or nucleus sampling scheduling can reduce loop formation and token garbling. Stabilizing decoding trajectories prevents collapse into repetitive or noisy sequences. This is particularly important for long reasoning chains in constrained combinatorial tasks.

RC9. Canonical form normalization layer. To reduce E2 (translation failures), outputs should be mapped through a canonicalization layer that converts semantically equivalent expressions into a standard form before evaluation. This reduces false negatives caused by predicate ordering, notation drift, or equivalent logical reformulations. It aligns model outputs with rubric expectations without penalizing valid alternative representations.

RC10. Hierarchical difficulty-aware prompting. Many failures arise when models misallocate effort across reasoning steps. Hierarchical prompting that explicitly separates coarse reasoning from fine computation can reduce overload in multi-constraint problems. This structure helps prevent premature commitment to incorrect intermediate conclusions and reduces cascading errors in E4-type chains.

RC11. Multi-view consistency training across invariance axes. Rather than training on isolated transformations, models should jointly optimize consistency across all invariance families (symbolic, structural, abstraction). This encourages global invariance manifolds in representation space rather than axis-specific robustness. It directly addresses the systematic degradation observed in abstraction shift and encoding swap scenarios.

RC12. Post-hoc reasoning trace alignment. Finally, aligning intermediate reasoning traces with final answers using supervision or preference optimization reduces self-contradiction. This discourages "correct-then-abandon" behaviors seen in E4.4 and improves coherence between verification steps and final outputs. It ensures that once a correct intermediate state is reached, it is preserved through to finalization.

Collectively, these interventions suggest that robustness in INVARIANTBENCH is not a single capability but an interaction between representation stability, constraint preservation, and decoding control. Improvements in any one dimension are insufficient unless accompanied by cross-axis consistency enforcement.

J. Details on Improving Invariant Reasoning Capabilities

To enhance logical reasoning performance on our diverse benchmark, we explored SFT and GRPO (RL).

J.1. Supervised Fine-Tuning (SFT)

J.1.1. SETUP AND IMPLEMENTATION

We fine-tuned GEMMA-3-1B-IT using the TRL library with default training hyperparameters unless otherwise noted. Training followed a parameter-efficient LoRA protocol applied to attention and feed-forward projection layers, with the model loaded in 4-bit precision to accommodate GPU memory constraints. We constructed two training regimes to disentangle memorization from generalization. In the *All Invariants* setting, a stratified 40:60 train–test split is applied across all task sub-types, and all four surface forms (`base`, `inv1`, `inv2`, `inv3`) are included during training. In the more constrained *Base + Inv1* setting, training uses only `base` and `inv1` variants, and evaluation is additionally conducted on the held-out `inv2` and `inv3` forms, which the model has never seen, providing a clean probe of whether fine-tuning induces genuine cross-form invariance rather than variant-specific pattern fitting. All models are trained for five epochs with accuracy on each invariant form tracked independently at every epoch checkpoint.

J.1.2. SFT RESULTS AND ANALYSIS

The learning curves in Figure 3 reveal a stark divergence between the two training regimes. Under All Invariants training, performance rises sharply and uniformly across all four surface forms, climbing from a shared baseline of 6.2% to 52.8% (Base), 52.8% (Inv1), 52.4% (Inv2), and 52.4% (Inv3) by epoch 5, a near-identical convergence pattern consistent with memorization of problem content through cross-variant leakage rather than internalization of invariant structure. The tight alignment of all four curves provides strong evidence that the model is exploiting the statistical overlap between training variants and test instances, and that the observed accuracy gains do not reflect a generalized capacity for invariance-robust reasoning.

The Base + Inv1 setting exposes this limitation directly. While Base and Inv1 performance improves substantially, reaching 41.4% and 41.8% respectively by epoch 5, accuracy on the unseen forms Inv2 and Inv3 remains critically low throughout training, plateauing at only 17.4% and 14.6%. This gap of approximately 25 percentage points between seen and unseen invariants confirms that SFT does not robustly transfer reasoning patterns across surface reformulations: the model learns a form-specific mapping that fails to generalize to logically equivalent but syntactically novel variants. The findings collectively indicate that standard SFT, even when applied to diverse invariant forms, primarily induces task-surface memorization rather than the representational abstraction required for invariance-robust reasoning. Full per-epoch accuracy breakdowns are provided in Table 97.

Table 97. SFT performance across epochs under full and partial invariant training settings.

Epoch	Setup	Base	Inv1	Inv2	Inv3
0	all invariants	6.25	6.08	6.42	6.17
1	all invariants	36.08	38.75	38.92	38.42
2	all invariants	49.08	46.67	48.17	50.08
3	all invariants	49.17	47.67	49.33	50.33
4	all invariants	52.58	51.83	51.00	51.92
5	all invariants	52.75	52.75	52.42	52.42
0	Base + Inv1	6.25	6.08	6.42	6.17
1	Base + Inv1	23.58	27.75	7.83	7.25
2	Base + Inv1	36.33	34.25	12.00	14.08
3	Base + Inv1	38.83	35.92	11.08	17.08
4	Base + Inv1	41.50	40.83	14.50	15.92
5	Base + Inv1	41.42	41.75	17.42	14.58

J.2. Reinforcement Learning via GRPO

J.2.1. SETUP AND IMPLEMENTATION

We trained GEMMA-3-1B-IT using Group Relative Policy Optimization (GRPO) (**?**) as implemented in TRL, with default hyperparameters. The reward signal is binary correctness evaluated against the ground-truth answer for each invariant variant, and the model generates multiple rollouts per problem to estimate group-relative advantages. As with SFT, we train under two regimes: All Invariants (all four surface forms available during training) and Base + Inv1 (training restricted to `base` and `inv1`, with `inv2` and `inv3` held out entirely for evaluation). The same stratified 40:60 split across task sub-types is used in both settings, ensuring that SFT and GRPO results are directly comparable. Training proceeds for five epochs, with per-invariant accuracy tracked at each checkpoint.

J.2.2. GRPO RESULTS AND ANALYSIS

GRPO training exhibits a different trajectory from SFT across both regimes, as shown in Figure 3. Under *All Invariants*, accuracy increases steadily but remains lower overall, reaching 31.94 (Base) and 31.67 (Inv1) at epoch 5, with Inv2 and Inv3 at 27.78 and 28.06. Unlike the tightly clustered SFT curves, Inv2/Inv3 lag behind Base/Inv1 throughout training, with only partial convergence, indicating gradual and incomplete cross-form generalization.

The generalization effect of GRPO is more visible in the *Base + Inv1* regime. While SFT leaves Inv2 and Inv3 at 17.42 and 14.58, GRPO improves them to 19.17 and 20.69 by epoch 5, a modest gain of a few points. Base and Inv1 reach 29.72 and 31.53, lower than SFT on seen variants. These results indicate that reward-driven optimization provides limited improvement in transfer to unseen forms but does not substantially close the invariance gap. Full per-epoch results are reported in Table 98.

J.3. On SFT/GRPO Diagnostic Scope

A potential limitation is that the SFT and GRPO experiments are conducted on a single model (GEMMA-3-1B-IT) without multiple seeds or cross-architecture replication. We emphasize that these experiments are intended as *diagnostic interventions* rather than definitive claims about training paradigms at scale. The goal is to isolate whether standard supervised imitation versus reward-driven optimization exhibits qualitatively different behavior under controlled invariance settings. As such, the focus is on identifying directional effects (e.g., memorization vs. generalization patterns) rather than establishing statistically robust performance trends across models.

Table 98. GRPO (RL) performance across epochs under full and partial invariant training settings.

Epoch	Setup	Base	Inv1	Inv2	Inv3
0	all invariants	6.25	6.08	6.42	6.17
1	all invariants	11.53	9.17	11.53	8.47
2	all invariants	18.61	19.03	16.25	14.86
3	all invariants	21.39	20.56	23.75	21.67
4	all invariants	28.75	28.33	29.72	28.89
5	all invariants	31.94	31.67	27.78	28.06
0	Base + Inv1	6.25	6.08	6.42	6.17
1	Base + Inv1	8.75	10.83	6.67	7.08
2	Base + Inv1	16.53	22.50	11.53	7.64
3	Base + Inv1	23.06	25.97	14.86	11.11
4	Base + Inv1	26.25	29.44	17.22	15.14
5	Base + Inv1	29.72	31.53	19.17	20.69

Importantly, the observed divergence between SFT and GRPO is large and structurally consistent across training regimes within this controlled setup, making it unlikely to be purely attributable to noise or initialization variance. However, we do not claim that these findings universally generalize across model sizes, architectures, or training seeds. Extending this analysis to larger models (e.g., 7B, 8B, or beyond), multiple random initializations, and alternative architectures remains an important direction for future work.

We therefore position these results as a proof-of-concept demonstrating that training objectives can materially affect invariance behavior, rather than as a conclusive comparison of optimization strategies. A comprehensive study across scales and settings is left for subsequent investigation.

K. Discussion

K.1. Overall Discussion

High base accuracy vs. true reasoning. Our results demonstrate that strong performance on fixed benchmarks does not necessarily reflect robust reasoning ability. Across models, high accuracy on original problem instances often degrades substantially when semantically equivalent but syntactically transformed variants are introduced. This indicates that many reported gains in reasoning benchmarks may be driven by surface-form pattern matching rather than stable semantic understanding. The observed performance collapse under invariance-preserving transformations exposes an illusion of competence, where models appear capable only within narrow representational contexts. In contrast, true reasoning should remain invariant under all semantics-preserving transformations. Our framework operationalizes this requirement and shows that it yields a significantly stricter evaluation axis than conventional accuracy-based metrics.

Severity of the invariance gap across model scales. Across 15 evaluated models, we observe a persistent and non-trivial invariance gap that does not vanish with scale. Smaller models exhibit extreme brittleness, with degradation exceeding 10 points under transformation, while even frontier systems show consistent drops between 1 and 5 points. Despite improvements in base accuracy with scaling, full-consistency remains low, with most models achieving below 5 percent invariant stability. This suggests that scaling alone improves surface competence but does not resolve representational fragility. The gap between base performance and invariant performance therefore remains structurally persistent across architectures and training regimes. These findings indicate that invariance robustness is not an emergent property of scale alone.

The challenge of abstraction-level shifts. Among all invariance axes, abstraction-level shifts emerge as the most consistently damaging transformation class. These transformations systematically vary the level of conceptual granularity while preserving underlying semantics, such as converting concrete instances into parameterized forms. We observe failure rates between 11 and 21 percent even for the strongest models, indicating a failure to maintain conceptual stability across representational scales. This suggests that models are often anchored to the abstraction level of training data rather than learning level-invariant structures. As a result, correct reasoning at one level of detail does not reliably transfer to alternative formulations of the same problem. This highlights a fundamental limitation in how current models encode and generalize abstract structure.

Structural and positional fragility. Our analysis reveals significant sensitivity to structural rearrangements of problem representation. In particular, premise reordering induces accuracy losses of up to 19.6 percent in strong models, despite no change in semantic content. Similarly, representation and encoding swaps across text, symbolic, and tabular formats lead to substantial degradation in performance. These results suggest that models rely heavily on positional and format-specific cues rather than constructing fully abstract representations of problem structure. Such behavior is inconsistent with a robust notion of reasoning, which should be invariant to permutation of independent premises. Overall, these findings indicate that structural generalization remains a key weakness in current LLMs.

Divergent failure modes across model scales. Failure analysis reveals a clear stratification of error types across model scales. Smaller models primarily fail at early stages of reasoning, including premise comprehension and basic computation, leading to cascading downstream errors. In contrast, larger models generally succeed in initial parsing but fail at higher-level invariance preservation. Their errors are dominated by invariance brittleness and lexical-semantic selection issues under transformation. This shift indicates that scaling improves low-level reasoning competence but does not resolve higher-order stability across representations. Consequently, different model scales fail for fundamentally different structural reasons.

Limitations of supervised fine-tuning. Supervised fine-tuning produces strong improvements in in-distribution accuracy but fails to induce robust invariance. When trained on all variants, models exhibit apparent convergence across forms; however, restricted training leads to a persistent generalization gap of approximately 25 percentage points on unseen invariants. This demonstrates that SFT largely promotes memorization of overlapping surface patterns rather than abstraction of underlying structure. The resulting representations remain tied to training-specific formulations. As a consequence, SFT alone is insufficient for learning invariance-preserving reasoning.

Strengths and limits of reinforcement learning. Reinforcement learning via GRPO exhibits a more favorable generalization profile compared to SFT. Under partial invariant exposure, GRPO more than doubles performance on unseen variants relative to supervised baselines, while maintaining competitive accuracy on seen forms. This suggests that reward-based optimization encourages exploration of alternative reasoning paths that are less sensitive to surface form. However, despite these gains, a substantial invariance gap persists, indicating incomplete abstraction of semantic structure. These results imply that RL improves robustness but does not fully resolve representational sensitivity.

Human vs. machine robustness. Human evaluators exhibit near-perfect invariance across all transformation types, maintaining consistency rates above 98 percent. In contrast, even the strongest models fail to exceed approximately 48 percent full consistency. This large discrepancy highlights a fundamental gap between human semantic abstraction and machine pattern-based reasoning. Humans naturally operate over invariant conceptual representations that remain stable under paraphrasing or restructuring. Current models, however, remain sensitive to surface form perturbations despite high base accuracy. These results establish human performance as a strong upper bound for invariant reasoning systems.

Implications for benchmark contamination. Our findings also provide a natural defense against benchmark contamination concerns. Even if a model has seen a base problem during training, it must still generalize correctly to unseen invariant transformations. This ensures that performance on INVARIANTBENCH cannot be explained solely through memorization of training instances. The gap between base and invariant performance thus serves as a direct measure of true generalization ability. This makes invariance-based evaluation substantially more robust to data leakage effects. Consequently, invariance provides a principled alternative to static benchmark evaluation.

Broader impacts and risks of invariance optimization. While invariance evaluation improves diagnostic rigor, it also introduces potential risks if used improperly. On the positive side, it enables more reliable assessment of models in high-stakes domains where input variability is unavoidable. However, the detailed taxonomy of failures could be exploited for adversarial prompt engineering or targeted system probing. There is also a risk that models are optimized specifically for benchmark invariance without achieving true semantic robustness. This could lead to superficial improvements that mask underlying reasoning deficiencies. Therefore, invariance should be treated as a diagnostic framework rather than a direct optimization target.

K.2. Future Work Directions

Joint-invariance optimization strategies. Current approaches such as SFT and GRPO improve performance in a largely axis-specific manner, where gains on one invariance type do not reliably transfer to others. A key direction is to develop joint-invariance optimization objectives that enforce consistency across all semantics-preserving transformations simultaneously. This requires learning objectives that explicitly minimize divergence between latent representations of a single problem under different invariant views. Rather than treating each transformation as an independent augmentation, the model should be optimized over the full invariance manifold induced by a seed problem. Such a formulation would encourage convergence toward representation spaces that are stable under reparameterization, reordering, and abstraction shifts. This shift is necessary to move from brittle, surface-dependent generalization to globally invariant reasoning systems.

Constraint-aware attention mechanisms. A major source of reasoning failure arises from premature loss or dilution of critical constraints during encoding, particularly in multi-constraint problems. Future architectures should incorporate constraint-aware attention mechanisms that explicitly preserve high-salience logical tokens throughout computation. This may be achieved through auxiliary supervision signals that align attention distributions with formally important elements such as quantifiers, bounds, and conditional clauses. Alternatively, gradient-based regularizers can be used to enforce stability of attention mass over constraint-defining tokens. Strengthening this early-stage representation is essential to prevent cascading failures in downstream reasoning steps. Overall, this direction targets improving the fidelity of the model's internal problem representation before inference begins.

Multi-step state persistence via latent tracking. Many invariance failures emerge from instability in intermediate reasoning states, where early correct deductions are overwritten or ignored in later steps. This motivates the development of explicit latent state tracking mechanisms that persistently store intermediate conclusions across generation steps. A promising direction is hybrid architectures combining neural generation with structured scratchpad memory or differentiable working memory modules. Such mechanisms would allow the model to verify consistency between newly generated steps and previously established constraints. This would directly mitigate self-contradiction and reasoning drift in long-horizon tasks such as planning and combinatorial search. Improving state persistence is critical for enabling reliable multi-step reasoning beyond shallow heuristic approximations.

Dual-path verification for computational stability. Our error analysis highlights frequent failures in arithmetic and logical consistency that are not self-corrected during generation. To address this, future systems should incorporate dual-path architectures consisting of a generative reasoning module and an independent verification module. The verifier would continuously audit intermediate steps for arithmetic correctness, logical consistency, and constraint satisfaction. Upon detecting inconsistencies, it would trigger recomputation rather than allowing error propagation. This design mirrors human problem-solving strategies, where intermediate results are routinely checked and corrected. Such enforced verification is essential for improving robustness in domains requiring precise computation and formal correctness.

Abstraction-level invariant representations. One of the most persistent weaknesses observed is sensitivity to abstraction-level shifts, where performance degrades when problems are reformulated at different granularities. Future research should aim to develop abstraction-invariant representations that decouple reasoning operations from specific instantiations. This may involve training curricula that systematically pair concrete and abstract formulations of the same underlying problem. Additional constraints can enforce symmetric performance across abstraction pairs during optimization. By encouraging alignment across abstraction hierarchies, models can learn more transferable and structural representations of reasoning tasks. Achieving abstraction invariance is a key step toward closing the gap between surface pattern matching and genuine conceptual reasoning.

Robustness to role-by-position binding. Structural perturbations such as premise reordering and entity relabeling expose a strong tendency in current models to bind semantic roles to positional or lexical cues. Future training strategies should explicitly disrupt these shortcuts through adversarial reordering and systematic relabeling during pretraining and fine-tuning. The goal is to ensure that semantic roles are inferred purely from relational structure rather than token position or identity. This requires exposing models to aggressively permuted inputs where surface cues are unreliable or misleading. By enforcing invariance to syntactic arrangement, models can develop more stable relational reasoning capabilities. This is particularly important for multi-entity and graph-structured reasoning tasks.

Lexical-semantic precision and grounding. A significant fraction of failures in language-based reasoning arises from incorrect but plausible lexical substitutions that violate strict semantic constraints. Future work should focus on improving fine-grained semantic control, particularly around negation scope, quantifier interpretation, and pragmatic modifiers. This can be achieved through loss functions that penalize near-synonym substitutions that alter truth conditions. Additionally, training data should include contrastive examples that distinguish subtle semantic differences between closely related expressions. Enhancing lexical precision will reduce errors that stem from fluent but semantically incorrect outputs. This is essential for improving reliability in tasks requiring exact natural language interpretation.

Canonical normalization for format invariance. A non-trivial subset of errors arises not from reasoning failure but from format mismatch between model output and evaluation criteria. Future evaluation pipelines should incorporate canonical normalization layers that map equivalent representations into a standardized form prior to scoring. This includes normalizing mathematical expressions, symbolic outputs, and structured answers into consistent canonical formats. Such a mechanism would reduce noise introduced by superficial formatting differences and better isolate true reasoning performance. It also enables fair comparison across models that may prefer different valid output conventions. Separating semantic correctness from syntactic form is crucial for accurate benchmarking.

K.3. Broader Impacts

This work introduces INVARIANTBENCH, a systematic framework for evaluating invariance in large language model reasoning under semantics-preserving transformations. While the primary goal is diagnostic—measuring whether models are robust to changes in representation—it has several broader societal implications, both positive and negative.

On the positive side, the benchmark provides a more reliable lens for assessing *true reasoning capability* beyond surface-level accuracy. By exposing invariance gaps that remain hidden in standard evaluations, it can improve model selection for high-stakes applications such as education, scientific reasoning, legal analysis, and decision-support systems. Models that are invariant-robust are less likely to fail due to superficial rephrasings or formatting differences, which is important in real-world settings where inputs are naturally heterogeneous. More broadly, the framework encourages the development of models that reason over semantics rather than memorized patterns, potentially improving robustness, fairness, and generalization in downstream applications.

On the negative side, the same diagnostic capability can also be misused for model optimization purely toward benchmark performance without improving true reasoning, leading to "invariance gaming" where systems are tuned to specific transformation families rather than genuine semantic understanding. Additionally, by revealing systematic failure modes (e.g., sensitivity to entity relabeling, structural reformulation, or abstraction shifts), the benchmark could inadvertently guide adversarial prompt engineering or targeted attacks that exploit these weaknesses in deployed systems. There is also a broader risk that improved models trained to close invariance gaps could be applied in harmful contexts such as more persuasive disinformation generation, more robust automated manipulation across paraphrased content, or scalable generation of misleading but semantically consistent narratives.

Mitigation of these risks primarily lies in responsible release practices and evaluation framing. We emphasize that INVARIANTBENCH is intended as an evaluation and diagnostic tool rather than a direct optimization target, and recommend that improvements based on it be combined with broader safety and robustness evaluations. Overall, while the work does not introduce a new application domain, it meaningfully affects how reasoning systems are evaluated, which has downstream implications for both beneficial and potentially harmful deployment settings. =

L. Representative Benchmark Samples with Model Answers and Evalautions

A2. Set Algebra + Cardinality with Constraints

ID: 83994f0c8a9643d8 | Base

Question: Let S = {x in N | 1 <= x <= 100000, x mod 41 = 23, x is prime}. How many elements are in S?

Answer: 17

Claude Haiku 4.5: By the prime number theorem for arithmetic progressions and computational verification, the actual count is: 80.

DeepSeek-V3.1: 108.

Gemma-4-25B-A4B-I: 237.

GPT-5.2: 240.

Llama-3.3-70B-I: 0.

Olmo-3.1-32B-I: 0.

Qwen3.6-Plus: 240.

A3. Conditional Override Transformation Systems

ID: 8d3aeea181294a0d | Base

Question: At a music shop, Bundle A contains 2 strings, 1 reeds, and 3 sticks for 86; Bundle B contains 1 strings, 3 reeds, and 2 sticks for 82; Bundle C contains 3 strings, 2 reeds, and 1 sticks for 90. If an order includes at least 6 reeds, each such unit is discounted by 5. One order contains 5 strings, 6 reeds, and 5 sticks. What is total charge?

Answer: 198

Claude Haiku 4.5: 198.

DeepSeek-V3.1: 337.

Gemma-4-25B-A4B-I: 261.

GPT-5.2: 198.

Llama-3.3-70B-I: 134.

Olmo-3.1-32B-I: 244.

Qwen3.6-Plus: 198.

Figure 4. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

A4. Constraint Satisfaction in Propositional Logic

ID: 7e19ebe57e9d4eee | Base

Question: Four switches connect to a motor: Q, R, S, T. Rules:
1. If Q is ON, then R is OFF ($Q \Rightarrow \neg R$)
2. If R is OFF, then S is ON ($\neg R \Rightarrow S$)
3. T is ON if and only if Q is ON ($T \Leftrightarrow Q$)
4. S is OFF ($\neg S$)

Find the true state of the switches.

Answer: Q=OFF, R=ON, S=OFF, T=OFF

Claude Haiku 4.5: Q: OFF, R: ON, S: OFF, T: OFF.

DeepSeek-V3.1: Q off R on S off T off.

Gemma-4-25B-A4B-I: Q is OFF, R is ON, S is OFF, T is OFF.

GPT-5.2: Q OFF R ON S OFF T OFF.

Llama-3.3-70B-I: Q is OFF R is ON S is OFF T is OFF.

Olmo-3.1-32B-I: Q OFF R ON S OFF T OFF.

Qwen3.6-Plus: Q and S and T are off and R is on.

A5. Inductive Rule + Formal Deduction

ID: 476db78e286a4519 | inv1

Question: In Boolean logic, the rule $p \wedge (p \vee q)$ is known as the "Absorption Law." If you simplify this rule, what does it exactly equal?

Answer: p

Claude Haiku 4.5: p.

DeepSeek-V3.1: The same as the first thing being true.

Gemma-4-25B-A4B-I: It is just the first thing being true.

GPT-5.2: Just the first true thing A.

Llama-3.3-70B-I: It is true.

Olmo-3.1-32B-I: True.

Qwen3.6-Plus: That something is true.

Figure 5. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

B1. Sequential Algebraic Systems with Layered Constraints

ID: 86ec7e1d1d574ddf | inv1

Question: A startup's cap table before Series A: Founders hold 80% (4M shares), Angel investors hold 15% (750K shares), ESOP pool remains 5% (250K shares), Total: 5M shares, Series A investor buys 2M new shares at $5/share, investing $10M, After Series A, what percentage does each group own?

Answer: Founders 57.1%, Angels 10.7%, ESOP 3.6%, Series A 28.6%

Claude Haiku 4.5: After Series A (7M total shares), founders hold 57.14% (4M/7M), angel investors hold 10.71% (750K/7M), the ESOP pool holds 3.57% (250K/7M), and the Series A investor holds 28.57% (2M/7M).

DeepSeek-V3.1: Founders 50% Angels 9.375% ESOP 3.125% Series A 25%.

Gemma-4-25B-A4B-I: Founders 57.14%, Angel 10.71%, ESOP 3.57%, Series A 28.57%.

GPT-5.2: Founders 57.14% Angels 10.71% ESOP 3.57% Series A 28.57%.

Llama-3.3-70B-I: Founders 40%, Angel investors 11.25%, ESOP pool 3.75%, Series A investor 45%.

Olmo-3.1-32B-I: Founders: 48%, Angel: 9%, ESOP: 2.5%, Series A: 40.5%.

Qwen3.6-Plus: Founders 57.1% Angels 10.7% ESOP 3.6% Series A 28.6%.

A6. Unit/Representation Invariance in Physics-Style Word Problems

ID: 5bd6c212f0ef49d1 | inv1

Question: A block of mass m = 100 kg moves a distance s = 10 m on a rough surface with friction coefficient μ = 0.4. Find work W done against friction.

Answer: 4000

Claude Haiku 4.5: W = μmgs = 0.4 × 100 × 10 × 10 = 4000 J.

DeepSeek-V3.1: 3920.

Gemma-4-25B-A4B-I: 3924 J.

GPT-5.2: 3920 J.

Llama-3.3-70B-I: 392.

Olmo-3.1-32B-I: 400 J.

Qwen3.6-Plus: 3920 J.

Figure 6. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

B2. Optimization over Coupled Algebraic Equations

ID: ae7024cda8854d76 | inv1

Question: Let the production rates of three workers be x, y, and z. The combined rates are x + y = 32, y + z = 28, and x + z = 24. Calculate the total production if worker x works 3 hours, worker y works 2 hours at a rate reduced by 3 units/hour, and worker z works 3 hours.

Answer: 102

Claude Haiku 4.5: 102.

DeepSeek-V3.1: 84.

Gemma-4-25B-A4B-I: 75.

GPT-5.2: 102.

Llama-3.3-70B-I: 102.

Olmo-3.1-32B-I: 63.

Qwen3.6-Plus: 102.

B3. Probabilistic Reasoning with Conditional Dependence

ID: 07710e3bde024aae | inv2

Question: Evaluate the following conditions for a competitive event: Format: strictly single-elimination; Total participants: exactly 64 players; Skill level: perfectly equal (every match is a 50/50 coin flip); Seeding: initial bracket placements are assigned completely at random. Based strictly on these parameters, what is the exact mathematical probability that Player Alpha and Player Beta will face each other in the final championship match?

Answer: 1/2,016

Claude Haiku 4.5: 1/32.

DeepSeek-V3.1: 1/63.

Gemma-4-25B-A4B-I: 1/1024.

GPT-5.2: Exact probability Player Alpha and Player Beta meet in final is 1/2016.

Llama-3.3-70B-I: 1/63.

Olmo-3.1-32B-I: 1/63.

Qwen3.6-Plus: 1/2016.

Figure 7. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

B4. Interdependent Equation Systems with Structural Constraints

ID: 9c315c930a1a4788 | inv2

Question: Member's pre-discount cart total: $600. Item price: $100. Discount order: (3) 5% loyalty on running total; (2) 10% if cart >$200 before discounts; (1) 20% member discount. Apply in correct sequence (1→2→3). What is the final price?

Answer: $68.40

Claude Haiku 4.5: $408.

DeepSeek-V3.1: 468.3.

Gemma-4-25B-A4B-I: $384.

GPT-5.2: Final price $68.40.

Llama-3.3-70B-I: 384.

Olmo-3.1-32B-I: 420.

Qwen3.6-Plus: $410.40.

B5. Constraint-Driven Algebraic Reasoning Problems

ID: fabcb364f8bf4676 | inv2

Question: At an aquarium store, Bundle C contains 3 filters, 2 lights, and 1 pump for 105; Bundle A contains 2 filters, 1 light, and 3 pumps for 91; Bundle B contains 1 filter, 3 lights, and 2 pumps for 104. If an order includes at least 5 filters, each such unit is discounted by 3. One order contains 5 filters, 5 lights, and 5 pumps. What is the total charge?

Answer: 235

Claude Haiku 4.5: 230.

DeepSeek-V3.1: 157.

Gemma-4-25B-A4B-I: 252.

GPT-5.2: Total charge is 235 after applying the filter discount in this order.

Llama-3.3-70B-I: 289.

Olmo-3.1-32B-I: 182.

Qwen3.6-Plus: 235.

Figure 8. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

C1. Algebra, Pseudocode, and Algorithmic Reasoning

ID: 1af17e20e2d84e8b | inv2

Question: Matrices A (6×4, rank 4) and B (4×6, rank 4) satisfy Sylvester's Law for shared dimension n = 4. What is rank(AB)?

Answer: rank(AB) = 4

Claude Haiku 4.5: 6.

DeepSeek-V3.1: 4.

Gemma-4-25B-A4B-I: 4.

GPT-5.2: 4.

Llama-3.3-70B-I: 4.

Olmo-3.1-32B-I: 4.

C2. Graph Search and Traversal Reasoning

ID: 5934c42b54254ee8 | inv3

Question: Graph Alpha is a mathematically defined "Strongly Regular Graph" where every pair of connected nodes shares exactly 0 common neighbors, and every pair of disconnected nodes shares exactly 1 common neighbor. In Graph Beta, you find a pair of connected nodes that share exactly 1 common neighbor. Can they be structurally isomorphic?

Answer: No, because the specific counts of common neighbors for adjacent and non-adjacent nodes are strict structural invariants.

Claude Haiku 4.5: No.

DeepSeek-V3.1: No.

Gemma-4-25B-A4B-I: No.

GPT-5.2: No they cannot be isomorphic since adjacency common neighbor counts differ.

Llama-3.3-70B-I: No.

Olmo-3.1-32B-I: No.

Qwen3.6-Plus: No.

Figure 9. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

C3. Path-Based Search and Multi-Agent Interaction Reasoning

ID: fbb64f7348104e0c | inv3

Question: A piecewise linear function f(x) has slopes 0.05 ($0 \leq x \leq 10k$), 0.08 ($10k < x \leq 50k$), 0.12 ($x > 50k$) with continuity at breakpoints. Evaluate f(65000) − f(0).

Answer: $5,500

Claude Haiku 4.5: 5,500.

DeepSeek-V3.1: 1240.

Gemma-4-25B-A4B-I: 8100.

GPT-5.2: 5500.

Llama-3.3-70B-I: 0.05*10000 + 0.08*40000 + 0.12*60000 = 500 + 3200 + 7200 = 10900.

Olmo-3.1-32B-I: 6500.

Qwen3.6-Plus: 5500.

C4. Graph Isomorphism and Structural Mapping

ID: b55d59a1e4c6400a | inv3

Question: As a Compliance Auditor, you are tracking a record that moves from Purchasing View through Accounting, Plant/Storage, Quality, and Sales Views before hitting the Approval stage. How many handovers occur in this specific sequence?

Answer: 5

Claude Haiku 4.5: 4.

DeepSeek-V3.1: 4.

Gemma-4-25B-A4B-I: 5.

GPT-5.2: Five handovers occur in this specific sequence.

Llama-3.3-70B-I: 5.

Olmo-3.1-32B-I: 4.

Qwen3.6-Plus: Five.

Figure 10. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

C5. Language, Semantics & Information Structure-based Reasoning

ID: deba8761b42d4852 | inv3

Question: New and interesting research conducted by Suleiman A. Al-Sweedan and Moath Alhaj is inspired by their observation that though there have been many studies of the effect of high altitude on blood chemistry, there is a _____ studies of the effect on blood chemistry of living in locations below sea level, such as the California towns of Salton City and Seeley. Therefore, research on high-altitude effects is more abundant than research on effects found in below-sea-level environments.

Answer: paucity of

Claude Haiku 4.5: scarcity of.

DeepSeek-V3.1: dearth.

Gemma-4-25B-A4B-I: scarcity of.

GPT-5.2: paucity of.

Llama-3.3-70B-I: lack.

Olmo-3.1-32B-I: few.

Qwen3.6-Plus: dearth of.

A1. Multi-Quantifier First-Order Logic

ID: 93dd1344e1514daa | Base

Question: Let Networks = {Net_1, Net_2, Net_3} be a set of three networks and Nodes = {N, n_1, n_2, ...} be a set of nodes. Let InNetwork(x, net) denote that node x is in network net, and Connected(x, y, net) denote that nodes x and y are connected in network net. The following constraints are given: 1. Node N is a member of every network: $\forall$net $\in$ Networks, InNetwork(N, net). 2. In each network, there exists at least one node n (where $n \neq N$) that is not connected to node N: $\forall$net $\in$ Networks, $\exists$n $\in$ Nodes ($n \neq N \wedge \neg$Connected(N, n, net)). Given these constraints, determine the minimum number of other nodes (excluding N) that must exist in the domain.

Answer: 1

Claude Haiku 4.5: 3.

DeepSeek-V3.1: 3.

Gemma-4-25B-A4B-I: 3.

GPT-5.2: 1.

Llama-3.3-70B-I: 3.

Olmo-3.1-32B-I: 3.

Figure 11. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

A2. Set Algebra + Cardinality with Constraints

ID: 5a19cfd636674901 | Base

Question: A secure server requires a 4-digit passcode strictly utilizing the digits 0 through 9.The system enforces three absolute rules.Every digit in the passcode must be completely distinct (no repeats).The number '0' is strictly forbidden from being the first digit. The final (fourth) digit must be an odd number. Exactly how many valid, unique 4-digit passcodes can be generated under these rules?

Answer: 2240

Claude Haiku 4.5: 2016.

DeepSeek-V3.1: 1296.

Gemma-4-25B-A4B-I: 1296.

GPT-5.2: 2240.

Llama-3.3-70B-I: 2634.

Olmo-3.1-32B-I: 4536.

Qwen3.6-Plus: 2240.

A3. Conditional Override Transformation Systems

ID: 45b528fd05a04f14 | Base

Question: At a hospital supply desk, Bundle A contains 2 masks, 1 gloves, and 3 gowns for 67; Bundle B contains 1 masks, 3 gloves, and 2 gowns for 60; Bundle C contains 3 masks, 2 gloves, and 1 gowns for 77. If an order includes at least 7 gowns, each such unit is discounted by 3. One order contains 5 masks, 4 gloves, and 7 gowns. What is total charge?

Answer: 156

Claude Haiku 4.5: 156.

DeepSeek-V3.1: 131.

Gemma-4-25B-A4B-I: 253.

GPT-5.2: 156.

Llama-3.3-70B-I: 133.

Olmo-3.1-32B-I: 186.

Qwen3.6-Plus: 156.

Figure 12. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

A4. Constraint Satisfaction in Propositional Logic

ID: 123f8a0e5ca4404c | Base

Question: Professor Lin supervises Wei, Adebayo, Petrova. Committee memberships 2014 Wei: Lin, Hassan, Morales. Adebayo: Lin, Morales. Petrova: Lin, Hassan, Morales, Chen. External examiners cannot be on the student's committee. Dr. Chen is available. For which students can Chen serve as examiner?

Answer: Wei and Adebayo only

Claude Haiku 4.5: Petrova.

DeepSeek-V3.1: Wei and Adebayo.

Gemma-4-25B-A4B-I: Adebayo.

GPT-5.2: Wei and Adebayo.

Llama-3.3-70B-I: Wei and Adebayo.

Olmo-3.1-32B-I: Petrova.

Qwen3.6-Plus: Wei and Adebayo.

B1. Sequential Algebraic Systems with Layered Constraints

ID: e8014b69456947d6 | inv1

Question: Three workers have rates x, y, and z. Given x+y=28, y+z=26, and x+z=26, calculate the total production if worker x works 3 hours at a rate reduced by 2 units/hour, worker y works 2 hours, and worker z works 3 hours.

Answer: 100

Claude Haiku 4.5: 100.

DeepSeek-V3.1: 32.

Gemma-4-25B-A4B-I: 138.

GPT-5.2: 100.

Llama-3.3-70B-I:96.

Olmo-3.1-32B-I: 128.

Qwen3.6-Plus: 100.

Figure 13. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

B2. Optimization over Coupled Algebraic Equations

ID: f7a2da94afe9408c | inv1

Question: At a supply desk, Package X contains 2 masks, 1 gloves, and 3 gowns for 50; Package Y contains 1 masks, 3 gloves, and 2 gowns for 56; Package Z contains 3 masks, 2 gloves, and 1 gowns for 56. If an order includes at least 6 gloves, each such unit is discounted by 3. One order contains 5 masks, 6 gloves, and 5 gowns. What is total charge?

Answer: 128

Claude Haiku 4.5: 128.

DeepSeek-V3.1: 68.

Gemma-4-25B-A4B-I: 175.

GPT-5.2: 128.

Llama-3.3-70B-I: 133.

Olmo-3.1-32B-I: 208.

Qwen3.6-Plus: 128.

B3. Probabilistic Reasoning with Conditional Dependence

ID: 043c9dee9e8b436b | inv2

Question: A number is chosen randomly from 1 to 100. If the number belongs to the set of multiples of 4, what is the probability that it belongs to the subset of multiples of 8?

Answer: 0.48

Claude Haiku 4.5: 1/2.

DeepSeek-V3.1: 1/2.

Gemma-4-25B-A4B-I: 1/2.

GPT-5.2: 12/25.

Llama-3.3-70B-I: 1/2.

Olmo-3.1-32B-I: One half.

Qwen3.6-Plus: 12/25.

Figure 14. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

B4. Interdependent Equation Systems with Structural Constraints

ID: fc2cce2f671843dd | inv2

Question: Bundle C contains 3 paperbacks, 2 hardcovers, and 1 magazine for 84; Bundle A contains 2 paperbacks, 1 hardcover, and 3 magazines for 73; Bundle B contains 1 paperback, 3 hardcovers, and 2 magazines for 71. If an order includes at least 5 paperbacks, each such unit is discounted by 2. One order contains 5 paperbacks, 5 hardcovers, and 4 magazines. What is the total charge?

Answer: 171

Claude Haiku 4.5: 171.

DeepSeek-V3.1: 124.

Gemma-4-25B-A4B-I: 189.

GPT-5.2: 171.

Llama-3.3-70B-I: 129.

Olmo-3.1-32B-I: 144.

Qwen3.6-Plus: 171.

B5. Constraint-Driven Algebraic Reasoning Problems

ID: b9c86db40db347ad | inv2

Question: A 45-liter tank contains a 25% dye solution. Perform the following operations: 1) Replace 9 liters of the current mixture with 9 liters of 50% dye solution. 2) Replace 7 liters of the resulting mixture with 7 liters of pure solvent. Calculate the final percentage of dye in the tank.

Answer: 0.2533

Claude Haiku 4.5: 25.33%.

DeepSeek-V3.1: 10.

Gemma-4-25B-A4B-I: 31.36%.

GPT-5.2: 25.33%.

Llama-3.3-70B-I: 29.5%.

Olmo-3.1-32B-I: 14.44.

Qwen3.6-Plus: 25.33%.

Figure 15. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

C1. Algebra, Pseudocode, and Algorithmic Reasoning

ID: eb35ff9d094045d6 | inv2

Question: A cylinder has radius 2 and height 9 and is cut into 3 equal parts. The area of one circular base is $\pi \times \text{radius}^2$. If the number of parts is greater than 1, each cut creates two new circular faces, so the total number of additional faces is (parts − 1) × 2, and the extra surface area is the number of new faces multiplied by the area of one base. If no cuts are made, there is no extra surface area. What is the new surface area?

Answer: rank(AB) = 16π

Claude Haiku 4.5: 60π.

DeepSeek-V3.1: 24π.

Gemma-4-25B-A4B-I: 36π.

GPT-5.2: 60π.

Llama-3.3-70B-I: $2\pi rh + 2\pi r^2$.

Olmo-3.1-32B-I: 72π.

Qwen3.6-Plus: 16π.

C2. Graph Search and Traversal Reasoning

ID: 7e4ff377c00041ac | inv3

Question: Hubs: A, B, C, D, E, F, G, H, I, J. Routes: A→B(5), A→C(7), B→D(6), C→D(4), D→E(3), B→F(10), E→G(2), F→H(5), G→I(3), H→I(4), I→J(2). D acts as a necessary gateway for cargo originating at A to access the middle segment of the network. What is the shortest path from A to J?

Answer: A-B-D-E-G-I-J

Claude Haiku 4.5: A-C-D-E-G-I-J.

DeepSeek-V3.1: 13.

Gemma-4-25B-A4B-I: A-C-D-E-G-I-J.

GPT-5.2: A-B-D-E-G-I-J total cost 21

Llama-3.3-70B-I: A-C-D-E-G-I-J.

Olmo-3.1-32B-I: A-C-D-E-G-I-J.

Qwen3.6-Plus: A-C-D-E-G-I-J costs 21.

Figure 16. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.