AN ESTIMAND-BASED EVALUATION FRAMEWORK FOR COMPOSITIONAL REASONING PIPELINES

Anonymous authors
Paper under double-blind review

ABSTRACT

Stage-level scores for compositional reasoning depend on which examples are scored, what upstream information is available, and which chance reference is used for comparison. These are distinct choices: scoring a stage only on examples whose upstream stages were answered correctly selects a model-specific population and changes the reference against which its score should be read, whereas supplying ground-truth upstream answers changes what the model receives while retaining all applicable examples. We present an estimand-based framework that makes these choices explicit. It distinguishes natural, prefix-conditional, and oracle-conditioned accuracy, pairing each with a chance reference matched to both its population and its upstream information. It additionally measures a signed selection gap by scoring the same oracle predictions on different populations. In a controlled shape-grammar study with five multimodal models, final-stage conditional accuracy reaches 98.45–99.67% but covers only 6.2–15.1% of applicable examples and ranks models differently from full-population accuracy. Holding oracle predictions fixed for two models, restricting evaluation to natural survivors raises accuracy by 7.15 and 12.97 percentage points at the rule stage and by 7.07 and 11.30 points at the final stage. Final-stage survivors reach 93.37–96.49% accuracy under fixed oracle predictions, still below their natural prefix-conditional scores. Furthermore, survivor-matched chance references deviate from full-cohort baselines. In a reanalysis of published results, accounting for coverage changes the preferred retrieval context size in five of eight comparisons. These results suggest the necessity of reporting population, input regime, coverage, and a matched chance reference alongside every stage score.

1 INTRODUCTION

Aggregate task scores measure whether a system succeeds, while stage-level evaluation describes performance on intermediate outputs. This diagnostic perspective is useful for multimodal systems and agents that combine perception, state construction, reasoning, tool use, and action (Liu et al., 2024; Zhang et al., 2025; Barke et al., 2026; Chen et al., 2026a). However, a stage score also depends on how the evaluation is constructed. Under *prefix conditioning*, stage d is scored only on examples with correct predictions at every earlier applicable stage. The resulting score describes that selected population, which can differ across models and need not represent all applicable examples.

Prefix conditioning brings at least two consequences: First, under prefix conditioning, coverage at a deep stage is the product of the preceding conditional pass rates, so high downstream scores may describe only small, model-specific subsets. This coverage loss increases the sample budget needed for deep-stage evaluation (Section 3.3.1). Section 5.1 reports the observed attrition. Second, it affects both the guessing reference and the evaluated population: 1) correct upstream answers can restrict downstream choices: once rule identity is given, only the arguments associated with that rule remain admissible, changing the appropriate guessing reference; 2) prefix-correct examples form a model-specific subset whose downstream performance can be higher, equal to, or lower than that of the full applicable population.

The practical consequences extend beyond our synthetic setting. In our reanalysis of FinRAG-QA (Miola et al., 2026), choosing retrieval context size by joint retrieval-gate-and-answer success changes the preferred configuration in five of eight retrieval–generator comparisons (Section 5.5),

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

demonstrating that the distinction between conditional performance and population-level success affects benchmark-based decisions.

Our contribution is methodological and empirical: we specify comparisons that distinguish changes in supplied information from changes in the population being scored, and demonstrate their consequences through a controlled synthetic study and an external benchmark reanalysis. The framework distinguishes natural full-population, prefix-conditional, and oracle-conditioned accuracy. For chance references, it varies the guessing information while holding the population fixed, then varies the population while holding the guessing rule fixed. For model accuracy, it retains natural survivor membership and compares the same oracle predictions on survivors, non-survivors, and the full applicable cohort. In summary, we make two main contributions:

1. **An estimand-based framework for stage-level evaluation.** First, we specify natural, prefix-conditional, and oracle-conditioned accuracy together with their evaluated populations and input regimes. Second, paired chance references separate guessing-information changes from population changes. Third, a signed selection gap compares populations under fixed oracle predictions. Finally, standard probability identities connect these measurements and make explicit which comparison each quantity supports.
2. **Empirical evidence that the distinctions affect interpretation.** Across five models, near-perfect operand scores describe small survivor subsets and yield different rankings from full-population scores (Section 5.2). The two-model audits at the rule and operand stages quantify selection gaps under fixed oracle predictions (Section 5.3). The matched-cohort argument results show that using the full-cohort reference for a survivor score overstates performance relative to its matched chance reference (Section 5.4).

Stage decomposition, oracle substitution, selection effects, and chance normalization are established ideas. The chain rule, mixture identity, and additive decompositions provide accounting relations for our framework. The paper's value lies in making the evaluation choices independently inspectable and showing their practical consequences, rather than establishing a general selection mechanism or a model ranking across tasks.

2 Related Work

Diagnostic evaluation of reasoning. CLEVR (Johnson et al., 2017) and CARV (Du et al., 2026) support structured diagnosis of visual reasoning, while ProcessBench (Zheng et al., 2025) and Who&When (Zhang et al., 2025) locate errors in reasoning steps and multi-agent tasks, respectively. These studies move evaluation beyond final-answer accuracy by examining intermediate reasoning and locating failures. Our focus is the estimand underlying a diagnostic score: which population it represents, what upstream information is available, and which guessing reference matches that evaluation.

Coverage and selection in staged evaluation. Wünsch et al. (2025) show that excluding data sets on which a method fails changes performance comparisons from unconditional evaluation to evaluation conditional on non-failure, potentially favoring methods assessed on easier subsets. FinRAG-QA (Miola et al., 2026) evaluates generation accuracy only when the ground-truth answer is present in the retrieved context, while GOBench (Chen et al., 2026b) documents intermediate-stage dropout that limits downstream verification coverage and end-to-end success. Together, these examples motivate reporting coverage alongside conditional scores, while recognizing that coverage alone does not resolve survivor selection or differences in reference difficulty. We build on this perspective by quantifying upstream correctness filtering and comparing natural survivors with non-survivors under identical oracle inputs, separating survivor selection from differences in supplied information. Conditioning on an intermediate outcome is a familiar problem in causal inference, where principal stratification defines effects within strata of a post-treatment variable rather than conditioning on its observed value (Frangakis & Rubin, 2002). Our setting differs in aim: we characterize what a reported stage score measures under each evaluation choice, rather than identify a causal effect within a latent stratum.

Oracle evaluation and error attribution. PuzzleVQA (Chia et al., 2024) progressively supplies ground-truth guidance for perception, induction, and deduction. CARV (Du et al., 2026) supplies

atomic transformations to test subsequent composition and application. Related precedents include scene-graph protocols with ground-truth object information (Xu et al., 2017) and TIDE's corrections of detection errors in existing predictions (Bolya et al., 2020). We use oracle inputs as an established diagnostic intervention and add an explicit population comparison: oracle accuracy is evaluated on all applicable examples, while natural survivor membership is retained to score the same oracle predictions on different subsets.

Chance correction and reference choice. Opitz (2024) analyzes chance correction and invariance to class prevalence as separate properties of classification metrics. Our score uses the established chance-correction form (Cohen, 1960), with a task-defined guessing reference. We focus on how upstream conditioning changes this reference: correct upstream answers can restrict the admissible answers, while selecting upstream-correct examples changes the evaluated population. We separate these effects and match the reference to both the available information and the scored examples.

3 Measurement Framework

We consider ordered evaluation stages with known ground-truth answers, including components scored after a single structured prediction as in Section 4. An estimand specifies the quantity an evaluation is intended to measure. For a fixed model and target stage, the framework makes the evaluated population, input regime, and scoring rule explicit, and separately specifies the guessing reference. We define the resulting accuracy estimands and reference comparisons, then use standard probability identities to describe coverage and signed population-selection contrasts.

3.1 What the three stage-level scores estimate

For a fixed model and cohort, let $\mathcal{D}_d$ contain the examples to which stage d applies, and let $\mathcal{P}_d \subseteq \mathcal{D}_d$ contain those examples for which every earlier applicable stage was correct in the *natural run*, where no ground-truth upstream fields are supplied. Let $c_{id}^{\mathrm{nat}} \in \{0,1\}$ indicate whether example i's stage-d prediction is correct in the natural run. Let $c_{id}^{\mathrm{orc}} \in \{0,1\}$ indicate correctness under a separate *oracle run* that supplies the complete ground-truth prefix required by stage d. We distinguish unconditioned, prefix-conditional, and oracle-conditioned accuracy:

$$\widehat{A}_d^{\mathrm{uncond}} = \frac{\sum_{i\in\mathcal{D}_d} c_{id}^{\mathrm{nat}}}{|\mathcal{D}_d|}, \qquad \widehat{A}_d^{\mathrm{prefix}} = \frac{\sum_{i\in\mathcal{P}_d} c_{id}^{\mathrm{nat}}}{|\mathcal{P}_d|}, \qquad \widehat{A}_d^{\mathrm{oracle}} = \frac{\sum_{i\in\mathcal{D}_d} c_{id}^{\mathrm{orc}}}{|\mathcal{D}_d|}. \tag{1}$$

Prefix conditioning selects a model-dependent population with coverage $|\mathcal{P}_d|/|\mathcal{D}_d|$. Oracle conditioning changes the input while retaining all applicable examples. Importantly, oracle scores are not filtered by whether the model reproduces the supplied prefix, but are computed across all applicable examples.

3.2 Separating information change from population change

The chance reference depends on both the information available to a guesser and the population over which it is evaluated. We separate the information change and population change using random guessing without the correct upstream answers (q_{id}^0) and with the complete correct prefix (q_{id}^+). For correct target answer $y_{id}^\star$, chance success is

$$b_{id}^s = \Pr_{\widetilde{Y}\sim q_{id}^s}\left(\widetilde{Y} = y_{id}^\star\right), \qquad s \in \{0,+\}.$$

We average the per-example chance probabilities to obtain three references: B_{all}^0 uses all applicable examples without correct upstream answers; B_{all}^+ uses the same examples with the complete ground-truth prefix; and $B_{\mathrm{survivors}}^+$ uses only natural survivors with that same prefix information. These references are defined as

$$B_{\mathrm{all}}^0 = \frac{\sum_{i\in\mathcal{D}_d} b_{id}^0}{|\mathcal{D}_d|}, \qquad B_{\mathrm{all}}^+ = \frac{\sum_{i\in\mathcal{D}_d} b_{id}^+}{|\mathcal{D}_d|}, \qquad B_{\mathrm{survivors}}^+ = \frac{\sum_{i\in\mathcal{P}_d} b_{id}^+}{|\mathcal{P}_d|}. \tag{2}$$

The full-cohort prefix-informed reference B^+_{all} provides the intermediate comparison: $B^+_{\text{all}} - B^0_{\text{all}}$ changes information on a fixed population, while $B^+_{\text{survivors}} - B^+_{\text{all}}$ changes population under a fixed guessing rule. Thus, equation 3 decomposes the total shift in chance accuracy:

$$B^+_{\text{survivors}} - B^0_{\text{all}} = \underbrace{(B^+_{\text{all}} - B^0_{\text{all}})}_{\text{information change}} + \underbrace{(B^+_{\text{survivors}} - B^+_{\text{all}})}_{\text{population change}}. \tag{3}$$

For example i at stage d, let $\mathcal{V}^0_{id}$ and $\mathcal{V}^+_{id}$ denote the finite sets of candidate answers available to the guesser without and with the complete ground-truth prefix, respectively. If $\mathcal{V}^+_{id} \subseteq \mathcal{V}^0_{id}$ and both sets contain the correct answer, uniform guessing gives $b^+_{id} = 1/|\mathcal{V}^+_{id}| \geq 1/|\mathcal{V}^0_{id}| = b^0_{id}$. This is answer-space contraction: Supplying correct upstream answers narrows or preserves the candidate set. Under uniform guessing, fewer candidates mean a higher probability of selecting the correct answer. This guarantees a nonnegative information-change term. In contrast, the population-change term $B^+_{\text{survivors}} - B^+_{\text{all}}$ can be positive, zero, or negative: survivors can have higher or lower chance success under the same guessing rule. Answer-space contraction therefore does not imply a positive population-selection effect.

3.3 Population consequences of prefix conditioning

Under a fixed natural input regime, restricting evaluation to $\mathcal{P}_d$ can reduce coverage and change model accuracy, with the accuracy change not determined by the chance-reference contrast.

3.3.1 Multiplicative coverage decay

For the coverage analysis, write $N_d = |\mathcal{D}_d|$ and $n_{\text{prefix},d} = |\mathcal{P}_d|$. Let A_d denote the applicability of stage d and C_i correctness at a preceding applicable stage i (treated as satisfied if stage i does not apply). Defining $q_i^{(d)} = P(C_i \mid A_d, C_1, \ldots, C_{i-1})$, the chain rule yields:

$$E[n_{\text{prefix},d}] = N_d \prod_{i<d} q_i^{(d)}. \tag{4}$$

No independence assumption is required for equation 4. The empirical counterpart holds exactly when pass rates are computed on the same stage-d cohort, and results are demonstrated in Section 5.1.

3.3.2 Selection among survivors

Prefix conditioning selects examples based on upstream correctness, but does not determine the direction of their target-stage performance difference. Survivors may perform better, equally well, or worse than excluded examples. A shared-difficulty account predicts positive selection only with additional assumptions linking upstream success to target success. Stage-specific difficulties or opposing associations can instead produce negative selection. Our framework does not assume a common difficulty ordering across stages.

To quantify the population contrast while holding the input regime fixed, we partition the oracle predictions using correctness from the *natural run*. For nonempty survivor and non-survivor groups, let $p = |\mathcal{P}_d|/|\mathcal{D}_d|$, and let r_S, r_F, and r_{all} be the oracle target accuracies on natural survivors, natural non-survivors, and all applicable examples, respectively. The following empirical identities hold:

$$r_{\text{all}} = p\, r_S + (1-p)\, r_F, \qquad \widehat{\Delta}_{\text{sel}} = r_S - r_{\text{all}} = (1-p)(r_S - r_F). \tag{5}$$

For $0 < p < 1$, the sign of $\widehat{\Delta}_{\text{sel}}$ is exactly the sign of $r_S - r_F$: positive selection raises the reported score when $r_S > r_F$; zero selection leaves it unchanged when $r_S = r_F$; and negative selection lowers it when $r_S < r_F$. With full coverage ($p = 1$), the gap is zero; with no survivors ($p = 0$), survivor accuracy and the gap are undefined.

Equation 5 also applies to natural-run predictions: prefix-conditional accuracy scores only survivors, whereas unconditioned accuracy scores all applicable examples. Their difference measures how restricting evaluation to survivors changes the reported natural-run accuracy. This differs from the

oracle comparison because it uses the model's original predictions rather than predictions generated with supplied ground-truth inputs.

Finally, the difference between natural accuracy on survivors and oracle accuracy on all applicable examples can be split into two parts:

$$\widehat{A}_d^{\text{prefix}} - r_{\text{all}} = \underbrace{(r_S - r_{\text{all}})}_{\text{selection within the oracle regime}} + \underbrace{(\widehat{A}_d^{\text{prefix}} - r_S)}_{\text{natural–oracle difference on survivors}} . \tag{6}$$

The first term changes population under fixed oracle predictions, while the second compares input regimes on the same natural survivors. Equation 6 then separates changing which examples are scored from changing the predictions being scored on the same survivors.

4 Empirical Design

We design an original visual reasoning dataset based on shape grammar (Stiny & Gips, 1972) to instantiate our measurement framework in a controlled setting. Figure 1 illustrates the task and structured prediction.

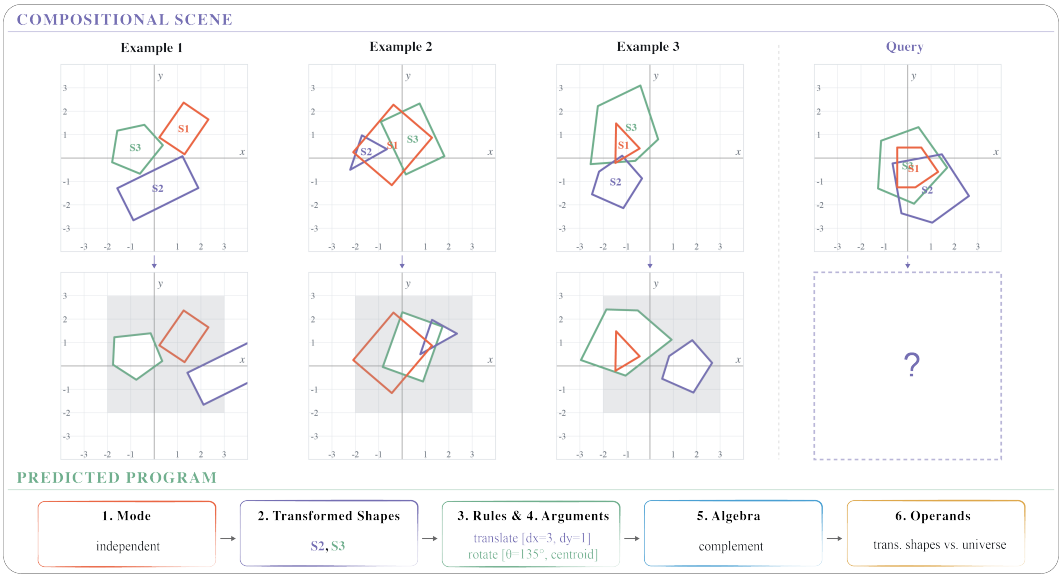


Figure 1: An example of the visual reasoning task and prediction pipeline. The model infers a shared transformation program from N demonstration pairs. We set N to one plus the number of parameters in the rule arguments.

4.1 Task, Predictions, and Staged Evaluation

In our shape grammar test, each puzzle contains multiple demonstration input–output pairs and a held-out query whose output is masked. In one test, all pairs share the same transformation program. For each model and prompt regime, the runner submits all demonstration images and the masked-query image together in one model-completion call per attempt, requesting one JSON program. The stage order is a scoring protocol applied after generation, and it does not correspond to separate model calls or to the model's internal reasoning order.

We evaluate six ordered stages of this prediction: *mode*, *shapes*, *rules*, *arguments*, *algebra*, and *operands*. Each stage contributes one correctness indicator per applicable puzzle. Mode specifies sequential geometric rules, independent per-shape rules, or no geometric transformation. Shapes identifies the named objects receiving geometric rules. Rules identifies the transformation primitives, and arguments specifies their parameters and reference frames. Algebra specifies the Boolean

operation, and operands specifies the inputs to that combination. Appendix D lists the geometric primitives, algebra operations, prompt-defined answer spaces and parameter domains.

Applicability is determined by ground truth. Shapes, rules, and arguments apply when shapes are transformed; algebra when a Boolean operation is present; operands for binary operations. An inapplicable stage passes the example through without exclusion. Coverage is the retained fraction of the current stage's applicable population (Section 3.1). Thus, coverage loss measures exclusion from conditional scoring, not termination of model execution. Supplementary scene-composition and component-level localization diagnostics are described in Appendices E and F.

4.2 Models and Evaluation Cohorts

We evaluate five leading multimodal models on a primary cohort of 4,100 examples: Claude Sonnet 5 (Anthropic, 2026), Gemini 3.5 Flash (Google DeepMind, 2026), GPT-5.6 Terra (OpenAI, 2026), Kimi K3 (Kimi Team, 2026), and Qwen 3.8 Max (Qwen, 2026). Separate oracle runs change the information supplied in the prompt. On the full dataset, GPT-5.6 Terra and Kimi K3 receive ground-truth mode and shapes (MS). A separate matched cohort of 1,000 GPT-5.6 Terra examples uses identical example IDs across natural, MS, and MSR regimes, where MSR additionally supplies rule identities. MS supplies the complete prefix for rule identity, and MSR supplies rule identity for arguments; later-stage results under these regimes remain partial-oracle measurements.

5 Results

We first examine coverage decay and how scoring only these survivors changes model comparisons. We then measure how selecting survivors changes accuracy, and examine how supplied upstream answers and the evaluated population affect chance baselines. Finally, Section 5.5 reanalyzes FinRAG-QA to show how accounting for coverage can change the preferred retrieval configuration in an existing benchmark.

5.1 Coverage collapses across the five primary models

The theoretical coverage decay outlined in Section 3.3.1 manifests clearly in our primary evaluation. Figure 2 reports the number of applicable examples entering the stage after surviving the prefix. While all models begin with 4,100 applicable mode examples, attrition is rapid. Only 492–1,371 of the 3,750 argument examples remain eligible. At the algebra stage, this drops to 319–667 out of 2,600. Finally, at the operands stage, only 129–314 predictions remain eligible out of 2,080 applicable examples. Thus, final-stage effective coverage is merely 6.2–15.1%, representing a loss of 84.9–93.8% of the evaluation population.

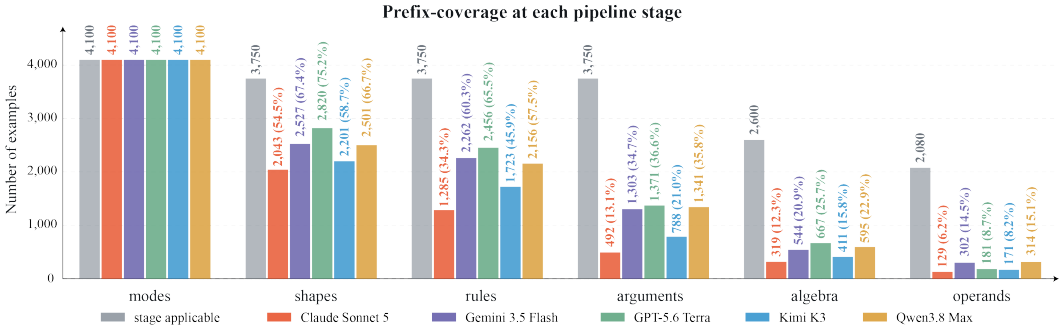


Figure 2: Prefix-surviving coverage of five models across six reasoning stages in the natural evaluation starting all with 4,100 examples.

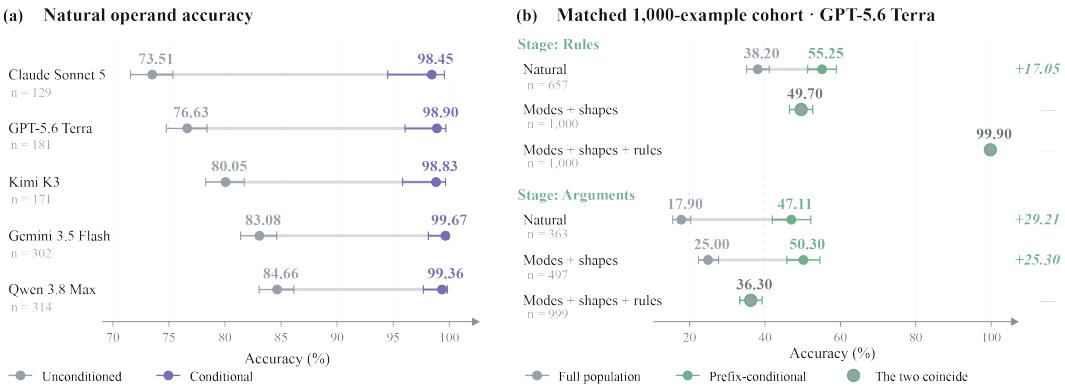


Figure 3: Natural operand scores and oracle input comparisons. Error bars show 95% confidence intervals (CIs). **(a)** Natural operand accuracy on all applicable examples and on prefix-correct subsets; n is the number of prefix-eligible examples. **(b)** Rule and argument accuracy on the matched 1,000-example GPT-5.6 Terra cohort, with oracle inputs supplying mode and shapes, then additionally rule identities. The separate comparison of populations under fixed oracle inputs is reported in Table 1.

5.2 Model rankings change and accuracy differences narrow under prefix conditioning

Figure 3(a) compares unconditioned and prefix-conditional operand accuracy in the natural run. Prefix-conditional accuracy ranges from 98.45% to 99.67% across the five models, compared with 73.51–84.66% unconditioned. These conditional scores are computed on different, model-specific populations: each model's operand denominator contains only the examples for which that model answered every upstream stage correctly. Eligible counts range from 129 to 314.

Because the evaluated populations differ, prefix-conditional operand scores are not directly comparable as measures of the same downstream capability. Two properties of the scores show the practical consequence. First, rankings reorder: unconditioned accuracy ranks the models Qwen 3.8 Max $>$ Gemini 3.5 Flash $>$ Kimi K3 $>$ GPT-5.6 Terra $>$ Claude Sonnet 5, whereas prefix-conditional accuracy ranks them Gemini 3.5 Flash $>$ Qwen 3.8 Max $>$ GPT-5.6 Terra $>$ Kimi K3 $>$ Claude Sonnet 5. Second, conditional scores lose discriminative power: Unconditioned accuracy spans 11.15 percentage points, with non-overlapping 95% CIs for every pair of models separated by at least two ranking positions. By comparison, prefix-conditional accuracy spans only 1.22 percentage points, and all five 95% CIs overlap over a common interval. Figure 3(a) visualizes these CIs and Table 3 reports the accuracy estimates and CI bounds.

5.3 Measuring population selection under fixed oracle inputs

Following Section 3.3.2, we compare matched natural and oracle predictions for GPT-5.6 Terra and Kimi K3. In this comparison, only the ground-truth mode and shapes are supplied for prediction. Operand survivors are identified by further filtering examples for correctness at rule, arguments, and algebra stage.

At the rule stage, which applies to 3,750 examples, natural survivors are examples with both mode and shapes predicted correctly. Table 1A shows that survivors achieve higher oracle rule accuracy than non-survivors for both models. The survivor–non-survivor gaps $r_S - r_F$ are 20.72 and 23.99 percentage points, respectively. Restricting the same oracle predictions to survivors raises GPT-5.6 Terra's rule accuracy by $+7.15$ points and Kimi K3's by $+12.97$ points. Equation 5 accounts for the difference between models: Kimi K3 excludes a larger fraction of examples, and its excluded examples perform worse relative to survivors. Equation 6 further decomposes the gap between natural prefix-conditional rule accuracy (Table 3) and full-population oracle accuracy. For GPT-5.6 Terra, the total gap is $+7.48$ percentage points, comprising 7.15 points of selection and 0.33 points of natural–oracle difference on survivors. For Kimi K3, the total $+15.87$ gap comprises

Table 1: Oracle accuracy and selection gaps at the rule and operand stages. 95% CIs in brackets: Newcombe hybrid score intervals (Wilson-based) for $r_S - r_F$, scaled by $1 - p$ for the selection gap.

	GPT-5.6 Terra		Kimi K3	
Natural-run group	Correct / total	Accuracy (%)	Correct / total	Accuracy (%)
A Rule stage				
Survivors (S)	1,363 / 2,456	55.50	738 / 1,723	42.83
Non-survivors (F)	450 / 1,294	34.78	382 / 2,027	18.85
All examples	1,813 / 3,750	48.35	1,120 / 3,750	29.87
Survivor coverage p (%)	65.49		45.95	
Accuracy gap $r_S - r_F$	20.72	[17.43, 23.93]	23.99	[21.08, 26.85]
Selection gap $r_S - r_{\text{all}}$	**7.15**	[6.01, 8.26]	**12.97**	[11.39, 14.52]
B Operand stage				
Survivors (S)	169 / 181	93.37	165 / 171	96.49
Non-survivors (F)	1,626 / 1,899	85.62	1,607 / 1,909	84.18
All examples	1,795 / 2,080	86.30	1,772 / 2,080	85.19
Survivor coverage p (%)	8.70		8.22	
Accuracy gap $r_S - r_F$	7.75	[2.91, 10.99]	12.31	[8.08, 14.86]
Selection gap $r_S - r_{\text{all}}$	**7.07**	[2.65, 10.04]	**11.30**	[7.41, 13.64]

12.97 points of selection and 2.90 points of natural–oracle difference. All differences are calculated before rounding.

At the operand stage, which applies to 2,080 examples, natural survivors are examples whose natural run answered every upstream stage correctly. Under the same oracle input regime, survivors again achieve higher operand accuracy than non-survivors: 93.37% versus 85.62% for GPT-5.6 Terra, and 96.49% versus 84.18% for Kimi K3 (Table 1B). The selection gaps are $+7.07$ and $+11.30$ percentage points, respectively. Applying Equation 6, the GPT-5.6 Terra gap ($+12.60$ points) comprises 7.07 points of survivor selection and 5.52 points of natural–oracle difference on survivors; the Kimi K3 gap ($+13.64$) comprises 11.30 and 2.34 points.

Survivor selection therefore accounts for part, but not all, of the near-ceiling conditional operand scores in Section 5.2. Selection is also not uniformly positive. At algebra, natural conditional minus full-population accuracy is -3.21, -7.11, and $+1.03$ percentage points for Claude Sonnet 5, Kimi K3, and GPT-5.6 Terra, respectively (Appendix A). These comparisons hold natural predictions fixed and show that selection can have either sign.

5.4 Separating information and population changes in the baseline

On the matched 1,000-example GPT-5.6 Terra cohort (seed 42), supplying upstream answers raises model accuracy: supplying mode and shapes raises rule accuracy from 38.20% to 49.70%, and additionally supplying rule identities raises argument accuracy from 17.90% to 36.30% (Figure 3(b); full results for both cohorts in Appendix B and C). Figure 3(b) also shows that prefix-conditional accuracy can exceed full-population accuracy under each regime. We now ask how the same upstream information changes the chance reference.

At the arguments stage, Table 2 reports chance baselines of 0.23% for natural full-cohort evaluation, 1.62% for oracle (MSR) evaluation, and 2.71% for natural prefix-conditional evaluation. Using these chance baselines, Equation 3 splits the total increase of $2.71 - 0.23 = 2.48$ percentage points into $1.62 - 0.23 = 1.39$ points from changing the guessing information on the full cohort and $2.71 - 1.62 = 1.09$ points from changing the evaluated population (restricting the same prefix-informed guessing rule to natural survivors). Both changes contribute to the reference used to interpret argument accuracy.

Table 2: Accuracy and population- and information-matched chance references on the 1,000-example cohort. For model and chance proportions a, b, chance-corrected accuracy(CCA) is $(a - b)/(1 - b)$ and lift is a/b.

			Accuracy (%)			
Stage	**Evaluated population**	n	**Model**	**Chance**	**CCA**	**Lift**
Rules	Natural, full	1,000	38.20	1.13	**37.49**	33.81
	Natural, prefix-correct	657	55.25	5.12	**52.84**	10.79
	Oracle (MS)	1,000	49.70	4.81	**47.16**	10.33
Arguments	Natural, full	1,000	17.90	0.23	**17.71**	77.83
	Natural, prefix-correct	363	47.11	2.71	**45.64**	17.38
	Oracle (MSR)	1,000	36.30	1.62	**35.25**	22.41

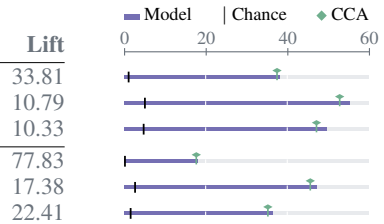


Table 2 also illustrates the danger of using the wrong chance baseline. Within the Arguments stage, the "Natural, prefix-correct" row pairs 47.11% accuracy on natural survivors with their matched 2.71% chance accuracy, giving a lift of 17.38. Substituting the lower 1.62% baseline from the "Oracle (MSR)" row would overstate this lift: that baseline describes all 1,000 examples, not the 363 survivors being scored. The chance reference must therefore match the evaluated population as well as the available upstream information (Appendix D).

5.5 Applying to an existing benchmark

FinRAG-QA reports answer-generation accuracy on a retrieval-gated subset of 999 financial-document questions (Miola et al., 2026). We reconstruct joint gate-and-answer success as coverage times conditional accuracy. For VoyageAI contextualized retrieval with o1-high, one chunk gives 85.2% conditional accuracy at 51.95% coverage, whereas twenty chunks gives 81.1% at 97.80% coverage. The preferred context size reverses under joint success: 79.31% for twenty chunks versus 44.26% for one. Across all eight retrieval–generator combinations, the maximizing context size changes in five. This changes configuration choice for a joint-success objective; it does not identify unrestricted answer accuracy. Appendix G gives all 24 configurations and reconstruction checks.

6 Limitations

Fixed-oracle selection gaps in learned-model accuracy are demonstrated at rule and operand stage for two models on one synthetic task. Generalization to natural images and agent trajectories remains untested. The FinRAG-QA reanalysis addresses coverage accounting, but it does not validate survivor-matched chance or fixed-oracle selection gaps. Oracle conditioning changes the prompt distribution, so oracle accuracy is not natural accuracy with selection bias removed. We lack shared-survivor target-agreement and difficulty-stratified analyses, and the observed gaps neither establish a shared latent-difficulty mechanism nor attribute the entire natural conditional–oracle difference to selection. At the operand stage, the natural–oracle difference on survivors combines the change in input regime with within-response agreement between upstream and operand answers.

7 Conclusion

The contribution of this work is an explicit evaluation framework and controlled evidence for the consequences of its choices. Specifying population, input regime, and guessing reference shows that 1) near-ceiling conditional scores describe small, model-specific subsets; 2) survivor selection raises reported accuracy even under fixed oracle predictions, explaining part of those scores; and 3) chance references must match the scored population. The negative natural-run algebra contrasts further show why selection must be treated as signed and specific to a stage and regime. These comparisons support reporting the stage-applicable and prefix-eligible counts, the input regime, raw accuracy, and a population- and information-matched chance reference together. Concretely, each stage score should carry four fields: *the stage-applicable count, the evaluated count with its population definition, the accuracy, and the chance reference computed on those same examples under the declared guessing rule.* The underlying probability identities organize these measurements; the controlled comparisons establish their relevance in the synthetic testbed, while the FinRAG-QA reanalysis shows that coverage accounting can change context choice for a joint-success objective.

AI USE STATEMENT

In this work, we used generative AI tools to assist in formulating formal mathematical definitions and contextualizing the interpretation of our empirical findings. Additionally, we used generative AI tools to improve the clarity, organization, and prose of the manuscript. We have not used generative AI tools for data collection or the design and execution of the original shape-grammar experiments. AI tools also assisted the implementation and interpretation of the public-data reanalysis of FinRAG-QA. We have reviewed all AI-assisted work: all AI-assisted outputs underwent independent verification by the authors, and all AI-edited text was manually audited. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Code and instructions for reproducing the experiments and evaluation analyses are submitted in the **supplementary material**, which will be publicly released upon publication. The evaluation dataset is available at `https://huggingface.co/datasets/submission28654/shape_grammar_4100`.

To support reproducibility, Section 3 provides the formal definitions and assumptions, while Section 4 describes the task, models, evaluation cohorts, and oracle input regimes. Appendices A–C report the applicable and eligible counts, stage-level scores, and available CIs needed to check the reported comparisons; Appendix D specifies the chance references and normalization formulas. Appendices E and F provide the detailed diagnostic results. Appendix G identifies the versioned source, cohort definitions, reconstruction formula, and complete results for FinRAG-QA. Supporting these evaluation protocols, all shape-grammar evaluation data are generated by Python scripts with a fixed random seed and can be regenerated using the same scripts and generation settings.

ETHICS STATEMENT

Our controlled experiments use procedurally generated geometric shapes and transformations. They involve no human participants, personal information, or sensitive demographic attributes, and no experiments are conducted on people or social systems. The external reanalysis uses published statistics, involves no new human-subject data collection, and makes no financial decisions or recommendations. In line with the ICLR Code of Ethics, our principal ethical considerations are scientific integrity, transparency, and accurate communication of model capabilities. We make the evaluated populations and input regimes explicit, report the limitations of our comparisons, and distinguish observed associations from causal claims. Our findings concern the tested evaluation settings and do not establish model reliability or safety in applications affecting people.

REFERENCES

Anthropic. Introducing Claude Sonnet 5, June 2026. URL https://www.anthropic.com/news/claude-sonnet-5.

Shraddha Barke, Arnav Goyal, Alind Khare, Avaljot Singh, Suman Nath, and Chetan Bansal. AgentRx: Diagnosing AI agent failures from execution trajectories. *arXiv preprint arXiv:2602.02475*, 2026. doi: 10.48550/arXiv.2602.02475. URL https://arxiv.org/abs/2602.02475. Accepted to Findings of EMNLP 2026.

Daniel Bolya, Sean Foley, James Hays, and Judy Hoffman. TIDE: A general toolbox for identifying object detection errors. In *European Conference on Computer Vision*, 2020. URL https://arxiv.org/abs/2008.08115.

Jing Chen, Yang Sun, Li Zhang, Lin Xu, and Jie Shi. Long-horizon agent trajectory attribution: A unified benchmark and fine-grained annotation framework. *arXiv preprint arXiv:2608.06909*, 2026a. doi: 10.48550/arXiv.2608.06909. URL https://arxiv.org/abs/2608.06909.

Yinghao Chen, Wantong Xie, Shuli Zeng, Sijia Zhang, Xiaotian Pan, Feng Wu, and Xiangyang Li. GOBench: Stage-wise diagnostics and the visual paradox in multimodal graph optimization. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 6144–6167. Association for Computational Linguistics, July 2026b. doi: 10.18653/v1/2026.findings-acl.306. URL https://aclanthology.org/2026.findings-acl.306/.

Yew Ken Chia, Vernon Toh, Deepanway Ghosal, Lidong Bing, and Soujanya Poria. PuzzleVQA: Diagnosing multimodal reasoning challenges of language models with abstract visual patterns. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 16259–16273, 2024. doi: 10.18653/v1/2024.findings-acl.962. URL https://aclanthology.org/2024.findings-acl.962/.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960. doi: 10.1177/001316446002000104. URL https://journals.sagepub.com/doi/abs/10.1177/001316446002000104.

Yongkang Du, Xiaohan Zou, Minhao Cheng, and Lu Lin. CARV: A diagnostic benchmark for compositional analogical reasoning in multimodal LLMs. *arXiv preprint arXiv:2603.27958*, 2026. URL https://arxiv.org/abs/2603.27958.

Constantine E. Frangakis and Donald B. Rubin. Principal stratification in causal inference. *Biometrics*, 58(1):21–29, 2002. doi: 10.1111/j.0006-341X.2002.00021.x.

Google DeepMind. Gemini 3.5 Flash. Model card, May 2026. URL https://deepmind.google/models/model-cards/gemini-3-5-flash/.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. URL https://arxiv.org/abs/1612.06890.

Kimi Team. Kimi K3: Open frontier intelligence. *arXiv preprint arXiv:2607.24653*, 2026. URL https://arxiv.org/abs/2607.24653.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. AgentBench: Evaluating LLMs as agents. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/e9df36b21ff4ee211a8b71ee8b7e9f57-Abstract-Conference.html.

Arianna Miola, Bruno Spaccavento, Lorenzo Silotto, Marco Bianchetti, and Luca Cagliero. Enhancing financial question answering: A novel benchmark dataset of banks' financial statements. *arXiv preprint arXiv:2609.03654*, 2026. doi: 10.48550/arXiv.2609.03654. URL https://arxiv.org/abs/2609.03654.

OpenAI. GPT-5.6 Terra. Model documentation, 2026. URL https://developers.openai.com/api/docs/models/gpt-5.6-terra. Accessed September 10, 2026.

Juri Opitz. A closer look at classification evaluation metrics and a critical reflection of common evaluation practice. *Transactions of the Association for Computational Linguistics*, 12:820–836, 2024. doi: 10.1162/tacl_a_00675. URL https://aclanthology.org/2024.tacl-1.46/.

Qwen. Qwen3.8-Max: Model description. Qwen Studio model documentation, 2026. URL https://chat.qwen.ai/legal-agreement/models. Accessed September 10, 2026.

George Stiny and James Gips. Shape grammars and the generative specification of painting and sculpture. In C. V. Freiman (ed.), *Information Processing 71*, pp. 1460–1465, Amsterdam, 1972. North-Holland. URL https://www.researchgate.net/publication/221329330.

Milena Wünsch, Moritz Herrmann, Elisa Noltenius, Mattia Mohr, Tim P. Morris, and Anne-Laure Boulesteix. Rethinking the handling of method failure in comparison studies. *Statistics in Medicine*, 44(23–24):e70257, 2025. doi: 10.1002/sim.70257. URL https://doi.org/10.1002/sim.70257.

Danfei Xu, Yuke Zhu, Christopher B. Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5410–5419, 2017. URL https://arxiv.org/abs/1701.02426.

Shaokun Zhang, Ming Yin, Jieyu Zhang, Jiale Liu, Zhiguang Han, Jingyang Zhang, Beibin Li, Chi Wang, Huazheng Wang, Yiran Chen, and Qingyun Wu. Which agent causes task failures and when? on automated failure attribution of LLM multi-agent systems. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 76583–76599, 2025. URL https://proceedings.mlr.press/v267/zhang25cq.html.

Chujie Zheng, Zhenru Zhang, Beichen Zhang, Runji Lin, Keming Lu, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. ProcessBench: Identifying process errors in mathematical reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1009–1024, 2025. doi: 10.18653/v1/2025.acl-long.50. URL https://aclanthology.org/2025.acl-long.50/.

A COMPLETE NATURAL STAGE ACCURACIES

Stage accuracy does not monotonically decrease as upstream failures accumulate. Unconditioned rule accuracy ranges from 17.71% to 39.12%, and argument accuracy drops to 3.60%–18.99%. However, accuracy rebounds to 45.15%–73.58% at algebra and 73.51%–84.66% at operands (Table 3). All rates and supplied 95% CIs are percentages. N is the stage-applicable count; n is the prefix-eligible denominator.

Table 3: Natural stage results for all five primary models.

Stage	N	n	Unconditioned [95% CI]	Conditional [95% CI]
Claude Sonnet 5				
Mode	4100	4100	56.68 [55.16, 58.19]	56.68 [55.16, 58.19]
Shapes	3750	2043	63.25 [61.70, 64.78]	62.90 [60.78, 64.97]
Rules	3750	1285	17.71 [16.52, 18.96]	38.29 [35.67, 40.98]
Arguments	3750	492	3.60 [3.05, 4.25]	17.07 [14.01, 20.65]
Algebra	2600	319	62.46 [60.58, 64.30]	59.25 [53.78, 64.50]
Operands	2080	129	73.51 [71.57, 75.36]	98.45 [94.52, 99.57]
Gemini 3.5 Flash				
Mode	4100	4100	68.24 [66.80, 69.65]	68.24 [66.80, 69.65]
Shapes	3750	2527	88.35 [87.28, 89.33]	89.51 [88.26, 90.65]
Rules	3750	2262	35.39 [33.87, 36.93]	57.60 [55.56, 59.63]
Arguments	3750	1303	16.19 [15.04, 17.40]	45.43 [42.75, 48.15]
Algebra	2600	544	73.58 [71.85, 75.24]	73.53 [69.67, 77.06]
Operands	2080	302	83.08 [81.40, 84.63]	99.67 [98.15, 99.94]
GPT-5.6 Terra				
Mode	4100	4100	76.95 [75.64, 78.21]	76.95 [75.64, 78.21]
Shapes	3750	2820	83.68 [82.46, 84.83]	87.09 [85.80, 88.28]
Rules	3750	2456	38.53 [36.99, 40.10]	55.82 [53.85, 57.78]
Arguments	3750	1371	18.99 [17.76, 20.27]	50.18 [47.54, 52.82]
Algebra	2600	667	45.15 [43.25, 47.07]	46.18 [42.43, 49.97]
Operands	2080	181	76.63 [74.77, 78.40]	98.90 [96.06, 99.70]
Kimi K3				
Mode	4100	4100	61.00 [59.50, 62.48]	61.00 [59.50, 62.48]
Shapes	3750	2201	75.57 [74.17, 76.92]	78.28 [76.51, 79.96]
Rules	3750	1723	24.08 [22.74, 25.47]	45.73 [43.39, 48.09]
Arguments	3750	788	6.88 [6.11, 7.74]	26.78 [23.80, 29.98]
Algebra	2600	411	63.31 [61.44, 65.14]	56.20 [51.37, 60.92]
Operands	2080	171	80.05 [78.28, 81.71]	98.83 [95.84, 99.68]
Qwen 3.8 Max				
Mode	4100	4100	68.73 [67.30, 70.13]	68.73 [67.30, 70.13]
Shapes	3750	2501	83.71 [82.49, 84.85]	86.21 [84.80, 87.50]
Rules	3750	2156	39.12 [37.57, 40.69]	62.20 [60.13, 64.22]
Arguments	3750	1341	16.67 [15.51, 17.89]	41.83 [39.22, 44.49]
Algebra	2600	595	73.46 [71.73, 75.12]	73.45 [69.75, 76.84]
Operands	2080	314	84.66 [83.05, 86.15]	99.36 [97.71, 99.83]

B FULL-DATASET PARTIAL-ORACLE RESULTS

All runs supply mode and shapes. Full-population rows score every applicable example without prefix filtering.

Table 4: Full-dataset rule-identity intervention. The oracle supplies mode and shapes. Accuracies are reported as percentages. Oracle rates are calculated from the exact counts in Table 1.

	Natural run		Oracle run	
Model	Eligible n	Full-pop. accuracy	Eligible n	Full-pop. accuracy
GPT-5.6 Terra	2,456	38.53	3,750	48.35
Kimi K3	1,723	24.08	3,750	29.87

Table 5: Full-dataset oracle summaries, with full and prefix-eligible denominators. Accuracies are percentages.

Model	Score	Mode	Shapes	Rules	Arguments	Algebra	Operands
GPT-5.6 Terra	Full population	100.0	100.0	48.3	24.7	51.2	86.3
GPT-5.6 Terra	Conditional	100.0	100.0	48.3	51.1	52.4	99.2
Kimi K3	Full population	100.0	100.0	29.9	8.8	64.1	85.2
Kimi K3	Conditional	100.0	100.0	29.9	29.5	62.2	99.5
Full applicable N		4100	3750	3750	3750	2600	2080
GPT prefix-eligible n		4100	3750	3750	1813	815	249
Kimi prefix-eligible n		4100	3750	3750	1120	508	212

C COMPLETE MATCHED-COHORT RESULTS

MS supplies mode and shapes; MSR also supplies rule identities. N denotes the applicable count and n the prefix-eligible count.

Table 6: All available matched-cohort stage results. Accuracies are reported as percentages.

Regime	Stage	N	n	Full population [95% CI]	Conditional [95% CI]
Natural	Mode	1000	1000	74.10 [71.30, 76.72]	74.10 [71.30, 76.72]
Natural	Shapes	1000	741	84.50 [82.13, 86.61]	88.66 [86.18, 90.75]
Natural	Rules	1000	657	38.20 [35.24, 41.25]	55.25 [51.43, 59.01]
Natural	Arguments	1000	363	17.90 [15.65, 20.40]	47.11 [42.03, 52.25]
Natural	Algebra	627	81	45.61 [41.75, 49.53]	49.38 [38.77, 60.05]
Natural	Operands	502	23	79.48 [75.73, 82.78]	100.00 [85.69, 100.00]
MS	Mode	1000	1000	100.00 [99.62, 100.00]	100.00 [99.62, 100.00]
MS	Shapes	1000	1000	100.00 [99.62, 100.00]	100.00 [99.62, 100.00]
MS	Rules	1000	1000	49.70 [46.61, 52.79]	49.70 [46.61, 52.79]
MS	Arguments	1000	497	25.00 [22.42, 27.78]	50.30 [45.92, 54.68]
MS	Algebra	627	119	50.72 [46.81, 54.61]	52.10 [43.20, 60.87]
MS	Operands	502	38	86.65 [83.40, 89.35]	100.00 [90.82, 100.00]
MSR	Mode	1000	1000	100.00 [99.62, 100.00]	100.00 [99.62, 100.00]
MSR	Shapes	1000	1000	100.00 [99.62, 100.00]	100.00 [99.62, 100.00]
MSR	Rules	1000	1000	99.90 [99.44, 99.98]	99.90 [99.44, 99.98]
MSR	Arguments	1000	999	36.30 [33.38, 39.33]	36.34 [33.41, 39.37]
MSR	Algebra	627	193	48.80 [44.91, 52.71]	48.70 [41.75, 55.71]
MSR	Operands	502	59	84.66 [81.25, 87.55]	98.31 [91.00, 99.70]

D ANSWER SPACES AND CHANCE REFERENCES

Output domains. Table 7 summarizes the output domains associated with the six reported components. A chance reference additionally specifies which information is available to the guesser and how predictions are sampled. Let K be the number of labeled input shapes and T the number receiving geometric transformations.

Table 7: Answer domains and applicability for the six reported components.

Stage	Output domain and constraints	Applicability
Mode	`rule_mode` takes one of `seq`, `ind`, or `none`: two ordered rules per transformed shape, one rule per transformed shape, or no geometric rule, respectively.	All puzzles.
Shapes	`transformed_shapes` identifies a subset of $\{S_1, \ldots, S_K\}$. There are 2^K subsets including the empty set, and $2^K - 1$ nonempty subsets. (Boolean combination does not make a shape geometrically transformed.)	$T > 0$ for the reported stage; `none` requires an empty list.
Rules	Six canonical primitives: `translate`, `rotate`, `mirror`, `scale`, `nonuniform_scale`, and `shear`. In `seq`, two top-level identities specify the ordered rules; in `ind`, each transformed shape has one identity and both top-level identities are null. In `none`, both are null.	$T > 0$.
Arguments	Primitive-specific dictionaries (Table 8): two per transformed shape in `seq`, one per transformed shape in `ind`. The `shapes` dictionary contains one entry for each name in `transformed_shapes`; it is empty in `none`.	$T > 0$; all required argument sets must match for puzzle-level correctness.
Algebra	`op` is `sum` (union), `difference`, `product` (intersection), `symmetric_difference`, `complement`, or `none`. The schema has six labels; the reported applicable cohort contains the five non-`none` operations.	A Boolean operation is present.
Operands	`combine_targets` specifies original/original, original/transformed, or transformed/transformed states for binary operations. The schema also uses `none` for no algebra and JSON null for complement.	Binary Boolean operations; excludes no algebra.

Geometric parameter domains. Define $D = \{-3, -2, -1, 0, 1, 2, 3\}$, $Q = \{0.5, 0.75, 1.25, 1.5, 2.0\}$, and $\mathcal{F} = \{\texttt{global}, \texttt{centroid}\}$. A global frame uses the origin $(0, 0)$; a centroid frame uses the shape centroid. Translation is measured in displayed grid-cell units, and positive rotation angles are counterclockwise.

Table 8 counts the possible parameter settings for one transformation. The distributions q_{id}^0 and q_{id}^+ describe these guessing rules without and with the correct upstream answers, respectively. For complement, the prompt additionally requests an axis-aligned universe rectangle as four corners in counterclockwise order, starting at the bottom-left. Coordinates lie in $\{-4, -3, \ldots, 4\}$ and at least one side spans more than four grid cells. Its bounds are analyzed separately in Appendix F; this auxiliary output is outside the six-stage summary.

Table 8: Parameter domains in the task prompt. M counts settings for one canonical primitive application using the listed Cartesian products and explicit constraints. It does not count complete multi-shape programs or assert that every setting occurs in the generated cohort.

Primitive	Required keys	Allowed values and constraints	M
Translation	`dx`, `dy`	$(dx, dy) \in D^2 \setminus \{(0, 0)\}$.	48
Rotation	`theta`, `frame`	$\theta \in \{45, 90, 135, 180, 225, 270, 315\}$ degrees; frame in $\mathcal{F}$.	14
Reflection	`axis`, `frame`	Axis is `x` (horizontal) or `y` (vertical); frame in $\mathcal{F}$.	4
Uniform scaling	`s`, `frame`	$s \in Q$; frame in $\mathcal{F}$.	10
Non-uniform scaling	`sx`, `sy`, `frame`	$(s_x, s_y) \in Q^2$; frame in $\mathcal{F}$. The prompt states no $s_x \neq s_y$ restriction.	50
Shear	`k`, `orientation`, `frame`	$k \in \{-0.5, -0.25, 0.25, 0.5\}$; orientation is horizontal or vertical; frame in $\mathcal{F}$.	16

Table 9 lists the natural-run references in the 1,000-example cohort. The columns distinguish all examples without correct upstream answers (B^0_{all}), all examples with them (B^+_{all}), and natural survivors with them ($B^+_{\text{survivors}}$). The distinction is substantive at rules and arguments and smaller at shapes. For MS, the survivor-matched rule and argument references are 4.81% and 2.55%; for MSR they are 4.81% and 1.62%. The other supplied-regime survivor-matched references equal their full-cohort prefix-informed values.

Table 9: Chance references for the natural N=1000 cohort.

Stage	B^0_{all} (%)	B^+_{all} (%)	$B^+_{\text{survivors}}$ (%)
Mode	33.33	33.33	33.33
Shapes	25.33	25.33	25.18
Rules	1.13	4.81	5.12
Arguments	0.23	1.62	2.71
Algebra	20.00	20.00	20.00
Operands	33.33	33.33	33.33

E SCENE-COMPOSITION DIAGNOSTIC

This supplementary analysis examines object-level discrimination in the synthetic testbed. It addresses a task-construction question separate from the main comparisons of evaluation populations and chance references.

Motivation and design. If every object in a scene changes, an "all changed" policy obtains perfect recall without object-level discrimination. If no object changes, a "none changed" policy obtains perfect specificity. We therefore vary the geometrically transformed fraction T/K for $T \in \{0, \ldots, K\}$ and report recall and specificity on mixed scenes. These scenes make both metrics observable, distinguishing object-level discrimination from uniform scene-level responses.

Results. Figure 4 summarizes the scene-composition sweep. For $K = 3, T = 2$, Claude Sonnet 5 has 89.5% recall and 46.0% specificity, indicating frequent transformation predictions for unchanged objects. Gemini 3.5 Flash has 92.8% recall and 93.6% specificity in the same condition. At fixed $K = 3$, specificity decreases from $T = 1$ to $T = 2$ for all five tested models. These descriptive contrasts characterize errors on unchanged objects in this testbed.

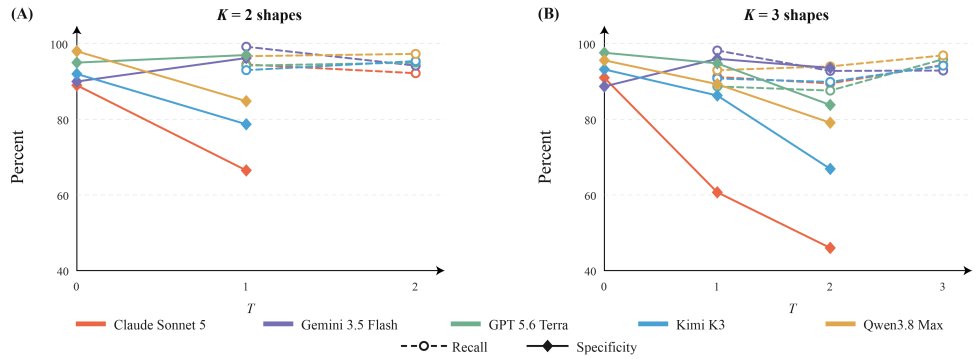


Figure 4: Scene-composition sweep for the five primary models.

Complete results. Table 10 reports recall and specificity (percent) for all seven (K, T) conditions for each model. A dash indicates that the metric is undefined because one object class is absent.

Table 10: All reported scene-composition conditions.

Model	K	T	recall	specificity
Claude Sonnet 5	2	0	—	89.0
Claude Sonnet 5	2	1	94.5	66.5
Claude Sonnet 5	2	2	92.2	—
Claude Sonnet 5	3	0	—	91.0
Claude Sonnet 5	3	1	91.2	60.7
Claude Sonnet 5	3	2	89.5	46.0
Claude Sonnet 5	3	3	94.4	—
Gemini 3.5 Flash	2	0	—	90.0
Gemini 3.5 Flash	2	1	99.2	96.2
Gemini 3.5 Flash	2	2	94.2	—
Gemini 3.5 Flash	3	0	—	88.7
Gemini 3.5 Flash	3	1	98.2	96.0
Gemini 3.5 Flash	3	2	92.8	93.6
Gemini 3.5 Flash	3	3	92.9	—
GPT-5.6 Terra	2	0	—	95.0
GPT-5.6 Terra	2	1	94.2	97.0
GPT-5.6 Terra	2	2	95.0	—
GPT-5.6 Terra	3	0	—	97.6
GPT-5.6 Terra	3	1	88.7	94.8
GPT-5.6 Terra	3	2	87.6	83.8
GPT-5.6 Terra	3	3	95.9	—
Kimi K3	2	0	—	92.0
Kimi K3	2	1	93.0	78.7
Kimi K3	2	2	95.4	—
Kimi K3	3	0	—	93.2
Kimi K3	3	1	90.8	86.3
Kimi K3	3	2	89.9	66.9
Kimi K3	3	3	94.2	—
Qwen 3.8 Max	2	0	—	98.0
Qwen 3.8 Max	2	1	96.7	84.8
Qwen 3.8 Max	2	2	97.3	—
Qwen 3.8 Max	3	0	—	95.6
Qwen 3.8 Max	3	1	93.0	89.3
Qwen 3.8 Max	3	2	94.0	79.1
Qwen 3.8 Max	3	3	96.9	—

F STRUCTURED OUTPUT DECOMPOSITION REVEALS DISTINCT LOCALIZATION GAPS

By decomposing the universe prediction into its constituent boundary lines, we identify model-dependent spatial grounding gaps (Table 11).

Claude Sonnet 5 and Kimi K3 display a large border advantage, scoring 43.9 and 26.2 percentage points higher on panel-border bounds than on interior bounds. In contrast, Gemini 3.5 Flash and GPT-5.6 Terra exhibit gaps of just 0.9 and 1.3 points. This structural decomposition suggests that these specific models are not struggling with the broader conceptual operation of drawing a bounding box, but rather suffer from a specific spatial anchoring defect when features float in the image interior.

Table 11: Complete universe-bound accuracy results by ground-truth location for all five models. Accuracy is reported as a percentage; the gap is panel-border minus interior accuracy in percentage points. Counts n refer to individual bounds, not independent scenes.

	Panel border		Interior		
Model	Accuracy (%)	n	Accuracy (%)	n	Gap
Claude Sonnet 5	82.7	625	38.8	1,455	43.9
Gemini 3.5 Flash	97.8	625	96.9	1,455	0.9
GPT-5.6 Terra	85.6	625	84.3	1,455	1.3
Kimi K3	85.9	625	59.7	1,455	26.2
Qwen 3.8 Max	97.9	625	87.5	1,455	10.4

G EXTERNAL REANALYSIS: FINRAG-QA

Source. FinRAG-QA (Miola et al., 2026) evaluates retrieval and answer generation on 999 financial-document questions. It scores answer generation only on questions whose retrieval passes a gate, and for each of its 24 generation configurations (four retrieval setups, two generators, and three context sizes $k \in \{1, 10, 20\}$) it reports the number of scored questions and the accuracy on them (version 1, Table 5). We use these published values directly, without new model queries.

Estimands. For a configuration with n_k scored questions and published conditional accuracy a_k, evaluation coverage is $p_k = n_k/999$, and joint success is $J_k = p_k\, a_k$: the fraction of all 999 questions that both pass the gate and are answered correctly. Questions excluded by the gate count as failures under joint success.

Results. Table 13 lists all 24 configurations, and Table 12 compares the context size preferred by each criterion. The preferred k changes in five of eight retrieval–generator combinations. Because published accuracies are rounded to 0.1 percentage points, each J_k is approximate, with rounding error below 0.05 points. The smallest within-combination difference between the best and second-best J_k is 1.15 points, so every ordering is unaffected.

Scope. Joint success is a different estimand from the source's conditional accuracy, and both remain valid: the result shows that the preferred configuration depends on which one is the objective. Changing k changes both the retrieved context and the scored population, so this comparison does not isolate a causal effect of context size. Nor does joint success identify full-cohort answer accuracy, since excluded questions might be answered correctly without passing the gate.

Table 12: Preferred context size under conditional accuracy and joint success for each retrieval–generator combination in FinRAG-QA.

Retrieval setup	Generator	Best conditional k	Best joint k	Changes?
OpenAI baseline	4o	10	20	Yes
OpenAI baseline	o1-high	20	20	No
OpenAI + headings	4o	10	20	Yes
OpenAI + headings	o1-high	20	20	No
OpenAI contextualized	4o	10	20	Yes
OpenAI contextualized	o1-high	10	20	Yes
VoyageAI contextualized	4o	10	10	No
VoyageAI contextualized	o1-high	1	20	Yes

Table 13: All 24 FinRAG-QA generation configurations. n and a are taken from the source's Table 5; $p = n/999$ and $J = p\,a$. Rates are percentages.

Retrieval setup	Generator	k	n	a	p	J
OpenAI baseline	4o	1	180	30.6	18.02	5.51
OpenAI baseline	4o	10	545	40.2	54.55	21.93
OpenAI baseline	4o	20	670	39.9	67.07	26.76
OpenAI baseline	o1-high	1	180	54.4	18.02	9.80
OpenAI baseline	o1-high	10	545	74.7	54.55	40.75
OpenAI baseline	o1-high	20	670	75.5	67.07	50.64
OpenAI + headings	4o	1	183	39.9	18.32	7.31
OpenAI + headings	4o	10	529	45.4	52.95	24.04
OpenAI + headings	4o	20	672	40.6	67.27	27.31
OpenAI + headings	o1-high	1	183	62.3	18.32	11.41
OpenAI + headings	o1-high	10	529	79.8	52.95	42.26
OpenAI + headings	o1-high	20	672	80.8	67.27	54.35
OpenAI contextualized	4o	1	263	49.0	26.33	12.90
OpenAI contextualized	4o	10	744	50.4	74.47	37.54
OpenAI contextualized	4o	20	851	49.7	85.19	42.34
OpenAI contextualized	o1-high	1	263	72.2	26.33	19.01
OpenAI contextualized	o1-high	10	744	82.3	74.47	61.29
OpenAI contextualized	o1-high	20	851	81.2	85.19	69.17
VoyageAI contextualized	4o	1	519	45.5	51.95	23.64
VoyageAI contextualized	4o	10	950	46.7	95.10	44.41
VoyageAI contextualized	4o	20	977	43.6	97.80	42.64
VoyageAI contextualized	o1-high	1	519	85.2	51.95	44.26
VoyageAI contextualized	o1-high	10	950	82.2	95.10	78.17
VoyageAI contextualized	o1-high	20	977	81.1	97.80	79.31