

Beyond Accuracy: Diagnosing Algebraic Reasoning Failures in LLMs Across Nine Complexity Dimensions

Parth Shrikant Patil¹, Yash Sinha², Murari Mandal³, Dhruv Kumar²

¹Nanyang Technological University (NBS)

²Birla Institute of Technology and Science, Pilani

³Kalinga Institute of Industrial Technology (KIIT), Bhubaneswar

parth010@e.ntu.edu.sg

Abstract

Aggregate mathematical accuracy does not usually identify the source of a model failure. In algebraic reasoning, errors may arise from nesting depth, operator identity, simultaneous intermediate values, branch-local composition, repeated-token counting, sequential dependence, or numeric magnitude. We introduce a nine-dimension algebraic complexity framework that converts these factors into controlled diagnostic variables. For each dimension, a parametric generator increases one target variable while holding the remaining variables at minimum values. Every retained problem is verified by a computer algebra system. We evaluate eight models, including six standard instruction-tuned models and two reasoning-specialised models. The results show that standard models exhibit a scale-limited D4 working-memory collapse in the 20–40 parallel-branch region. Notably, Qwen3 235B does not escape the transition band observed for Llama 3 8B, despite its much larger parameter scale. Reasoning-specialised models shift this limit to higher levels, but they do not remove it. D5 branch-local composition and D7 repeated-token counting produce distinct failures that are not reducible to aggregate length or depth.

1 Introduction

Mathematical reasoning is a central evaluation setting for large language models (LLMs). Existing benchmarks include grade-school word problems (Cobbe et al., 2021), competition mathematics (Hendrycks et al., 2021), graduate-level applied problems (Fan et al., 2024), and symbolic computation (Lample and Charton, 2020; Saxton et al., 2019). These benchmarks are valuable for comparing models, but their aggregate scores often conceal the cause of an incorrect response. A model may fail because an expression contains deeply nested subexpressions, an unfamiliar operator, many simultaneous intermediate values, repeated terms that

must be counted, or a long chain in which early errors compound.

This attribution problem becomes more important as static benchmarks saturate or become contamination-sensitive. Dynamic variants such as GSM-Symbolic (Mirzadeh et al., 2025) and Putnam-AXIOM (Gulati et al., 2025) show that surface changes can alter performance. Surveys of mathematical reasoning failures also emphasise that LLM errors are heterogeneous rather than a single capability gap (Song et al., 2026). Prior studies isolate individual mechanisms. Multiplication can induce $O(n^2)$ cross-token interactions (Dziri et al., 2023). Nesting depth causes recursive propagation losses (Saxton et al., 2019). Counting is sensitive to tokenisation and positional structure (Chang and Bisk, 2024; Zhang et al., 2024; Yehudai et al., 2024). Compositional reasoning can also break when subproblems are chained (Zhao et al., 2024; Hosseini et al., 2024). However, existing work does not yet provide a unified algebraic benchmark in which these mechanisms are separated, scaled systematically, and compared across model families.

We address this gap by defining nine algebraic complexity dimensions. They are syntactic length (D1), tree depth (D2), operator hardness (D3), working-memory load (D4), compositional branching (D5), solution ambiguity (D6), counting load (D7), sequential chain length (D8), and numeric magnitude (D9). Each dimension is implemented as a generator control parameter. Increasing D4, for example, adds independent single-digit products that must be stored before summation. Increasing D5 leaves the number of branches fixed but makes each branch internally more compositional. Increasing D7 repeats an identical operand many times so that the active burden is enumeration rather than arithmetic. The appendix provides full definitions, examples at multiple levels, and the structural coupling analysis (Appendix A (dimension guide); Appendix A (example items); Appendix D (cou-

plings)).

Our main empirical result is that standard instruction-tuned models share a scale-limited D4 working-memory failure. All standard models remain strong at small branch counts, but every standard model in the updated D4 battery reaches at least one zero-accuracy point by 40 branches. A 235B standard model does not escape the same transition band as the open 8B/12B baselines, which suggests that parameter count alone does not create register-like algebraic state. Reasoning-specialised models substantially shift D4 and D5 thresholds, but the extended batteries still locate failure. Qwen QwQ fails at 500 D4 branches and 50 D5 operations per branch, while DeepSeek R1 0528 fails at 3000 D4 branches and 70 D5 operations per branch. Counting remains especially idiosyncratic: Claude 3.5 Haiku stays perfect through $K = 300$, but DeepSeek R1 0528 fails at $K = 300$ and several models show non-monotonic recoveries.

Contributions. We make three contributions.

- **A nine-dimension algebraic diagnostic framework.** We operationalise nine structurally distinct algebraic complexity dimensions and connect each one to a documented LLM failure mechanism.
- **Automated, systematically scalable problem generation.** We introduce a parametric generator-and-verifier pipeline that raises the difficulty of one dimension at a time. The pipeline verifies every retained problem with a computer algebra system, avoiding human annotation and fixed-corpus ceilings.
- **Cross-model diagnostic evaluation.** We evaluate six standard instruction-tuned models and two reasoning-specialised models under the same controlled batteries, identifying which failures are shared, which are shifted by reasoning traces, and which remain model-specific.

2 Related Work

Mathematical benchmarks and attribution. GSM8K (Cobbe et al., 2021), MATH (Hendrycks et al., 2021), HardMath (Fan et al., 2024), and symbolic-mathematics benchmarks (Lample and Charton, 2020; Saxton et al., 2019) evaluate broad mathematical competence. These benchmarks usually combine several structural pressures inside

each item, which makes failure attribution difficult. GSM-Symbolic (Mirzadeh et al., 2025) and Putnam-AXIOM (Gulati et al., 2025) show that benchmark performance can change under controlled variants. However, they do not decompose algebra into independently scaled failure dimensions. Our benchmark uses controlled generation to produce per-dimension failure curves rather than a single aggregate score.

Automated problem generation. Prior work has generated algebra variants through syntactic transformation and verification (Singh et al., 2012) and procedural elementary-math templates (Xu et al., 2021). Other systems use large-scale LLM-assisted data generation (Tang et al., 2024; Manem et al., 2025; Chen et al., 2025). These systems primarily support practice or training. Our pipeline is diagnostic. Each generator is tied to a named complexity variable, and each level is used to estimate a failure threshold.

Controlled complexity and reasoning models.

Length-generalisation studies show that models may solve short algorithmic instances while failing on longer ones (Anil et al., 2022). CLRS-Text varies problem size and trace length in algorithmic reasoning (Markeeva et al., 2024); seqBench varies sequential depth and finds collapse as action length increases (Ramezanali et al., 2025). Recent trace-aware benchmarks further isolate depth-scaled sequential decay across task regimes and intermediate-state execution errors in algorithmic reasoning (Chapra et al., 2026; Kapoor et al., 2026). Reasoning-specialised models such as DeepSeek R1 (DeepSeek-AI, 2025) and Qwen QwQ (Alibaba Cloud, 2025) improve aggregate mathematical scores through extended reasoning traces. However, aggregate gains do not reveal which structural failures are eliminated and which are merely displaced. Controlled puzzle studies similarly find thresholded collapse even for reasoning models (Shojaee et al., 2025). We bring this controlled-complexity view to CAS-verified algebra.

3 Benchmark Design

The benchmark follows two principles. First, each dimension must describe a distinct algebraic pressure rather than a surface statistic alone. Second, only one variable should change inside a battery. The generator fixes all non-target dimensions at their minimum viable values, then increases the tar-

get variable across a level schedule chosen to reveal both gradual degradation and sharp thresholds.

Table 1 summarises the dimensions. The table is organised around failure mechanisms rather than surface difficulty labels. The same expression can be long because it has many repeated terms, a deep tree, or a long dependent chain. These cases impose different computational burdens and should not be collapsed into a single category of difficult algebra. The benchmark therefore treats each dimension as a testable hypothesis about why a model might fail.

Length, depth, and sequential dependence. D1 measures syntactic length in prefix notation. It is a useful stressor, but it is also the most confounded dimension because many other structures make expressions longer. D2 measures tree depth, where a model must resolve deeply nested partial results and propagate them back to the root. D8 measures sequential chain length, where only one live state is carried forward but each step depends on the previous result. D2 and D8 can both create long expression spines. Nevertheless, they stress different computations: recursive return through a tree versus repeated state update.

Operators and numeric magnitude. D3 isolates operator identity by holding depth, branch count, counting load, and operand size fixed while moving across an ordered operator ladder. This setup tests whether an error comes from the symbolic operation itself rather than from surrounding structure. D9 fixes the operator to multiplication and increases digit count, targeting partial-product accumulation and representation effects. D9 should therefore not be interpreted simply as a test of large numbers. It tests large operands under an operator whose cross-digit interactions grow rapidly.

Parallel memory and branch-local composition. D4 and D5 are the two most important structural dimensions to separate. D4 increases the number of independent products that must coexist before final summation. The arithmetic inside each branch is simple, so the intended burden is horizontal storage of many intermediate values. D5 keeps the number of branches fixed at three and instead increases the number of operations inside each branch. The intended burden is local composition before each branch resolves. A model can pass one of these dimensions and fail the other, which is why both are necessary.

Strategy choice and counting. D6 changes the algebraic problem type, from canonical equations to forms with several plausible solution routes. It tests strategy selection and form familiarity more than raw computation. D7 repeats an identical operand K times and asks the model to enumerate the repetitions exactly. This dimension is intentionally simple arithmetically: the point is to test cardinality tracking over repeated tokens. D7 is especially important because its failures are often not predicted by depth, branch count, or model scale.

This decomposition makes the benchmark diagnostic rather than merely difficult. A D4 failure indicates weakness in simultaneous intermediate-state management. A D7 failure indicates weakness in repeated-token cardinality. A D2 or D8 failure indicates difficulty with recursive or sequential state tracking. The appendix provides generated examples at multiple levels for every dimension ([Appendix A \(example items\)](#)).

4 Isolation, Scaling, and Verification

Each battery is generated under a one-active-variable protocol with two stages. First, we define the minimum viable template for the target dimension. Operands are single-digit when possible, operators are low-hardness unless D3 is active, branch count is minimal unless D4 is active, and sequential dependence is avoided unless D8 is active. Second, after this controlled baseline is fixed, we raise only the target variable. This design turns the benchmark into a set of controlled interventions on algebraic structure, rather than a collection of increasingly complicated problems.

Isolation examples. In D4, the generator creates a sum of independent single-digit products. Raising the level from small branch counts into the extended hundreds- and thousands-branch regime does not introduce harder operators or larger operands. It only increases the number of branch outputs that must be retained before the final sum. This distinction matters because the results show a standard-model transition band near moderate branch counts, while the reasoning-model extensions require much larger D4 levels before the next collapse appears. In D5, the generator fixes exactly three branches joined by two root operators and increases the operations inside each branch. In D7, the operand and arithmetic operation are fixed while the repeated-term count K grows. In D8, each level adds a dependent update step while keeping the live state count at one. These

Dim.	Name	What Is Varied	Failure Mechanism Isolated
D1	Syntactic length	Prefix-token count in simple expression chains.	Length extrapolation when arithmetic stays easy.
D2	Tree depth	Longest root-to-leaf path in a nested expression.	Recursive partial-result propagation.
D3	Operator hardness	Operator identity from basic arithmetic to powers and trigonometric forms.	Operator-specific symbolic competence.
D4	Working memory	Number of independent products summed after parallel evaluation.	Simultaneous storage of many intermediate values.
D5	Branch complexity	Operations inside each of three fixed branches.	Branch-local composition before branch outputs combine.
D6	Solution ambiguity	Algebraic problem type and number of plausible solution routes.	Strategy selection rather than raw arithmetic.
D7	Counting load	Number of identical repeated terms K .	Cardinality tracking over repeated tokens.
D8	Sequential chain length	Number of strictly dependent update steps.	Cascading error along a one-state chain.
D9	Numeric magnitude	Digit count in multiplication operands.	Cross-digit partial-product accumulation.

Table 1: The nine algebraic complexity dimensions used in the benchmark. Full definitions, examples, level schedules, and literature grounding are provided in [Appendix A \(dimension guide\)](#) and [Appendix A \(example items\)](#).

choices make a drop in accuracy interpretable.

Systematic scaling. The same generator can raise a dimension ceiling without redesigning the benchmark. D4 is scaled through 2, 4, 8, 12, 20, 25, 30, 40, 50, 60, 75, 100, 150, 200, 500, 1000, and 3000 branches. D5 is scaled from 0 to 100 operations per branch. D7 is scaled from $K = 5$ to $K = 300$ repeated terms. D8 is scaled from 1 to 500 dependent steps. Smaller increments are placed around expected transition regions, while wider jumps probe extrapolation after standard models have already collapsed. The complete schedules are listed in [Appendix A \(level schedules\)](#).

Problems are represented internally in Polish prefix notation because token count, tree depth, branching, and chain length are directly measurable from the sequence. Before prompting, every item is converted to standard ASCII infix notation. Each generated candidate is submitted to a computer algebra system (CAS). Candidates are rejected if they violate dimension controls, produce ambiguous solver output, exceed the CAS timeout, raise a solver exception, or yield a non-finite answer. This rejection-sampling loop allows the benchmark to be generated automatically while retaining deterministic, finite ground truth. The same automation makes the benchmark extensible: when a model saturates a level, the researcher can raise a generator parameter rather than manually authoring a new test set.

The dimensions are isolated as generator controls, not claimed to be mathematically independent in every surface statistic. Some co-movement is unavoidable: increasing D4 also lengthens the expression, D5 can increase local depth, and D8 produces both a long chain and a deep expression spine. We therefore analyse directional couplings separately in [Appendix D \(couplings\)](#). This lets the main analysis distinguish a target mechanism from

downstream surfaces such as length.

5 Experimental Setup

We evaluate eight models: GPT-4o Mini, Claude 3.5 Haiku, Qwen3 235B, DeepSeek V3, Gemma 3 12B, Llama 3 8B, DeepSeek R1 0528, and Qwen QwQ. The first six are standard instruction-tuned models that span hosted proprietary systems, open 8B/12B models, and large mixture-of-experts models. The final two are reasoning-specialised models with extended reasoning traces.

[Table 3](#) summarises the model coverage. The model set is designed for diagnostic comparison rather than leaderboard ranking. First, the open 8B/12B models and Qwen3 235B test whether failure thresholds shift smoothly with published parameter scale, while hosted proprietary systems provide contemporary API baselines without relying on undisclosed parameter counts. Second, decoder-only and mixture-of-experts models test whether architectural family changes the curve shape. Third, DeepSeek R1 0528 and Qwen QwQ test whether extended reasoning traces eliminate a failure mode or instead push it to larger generated levels. For this reason, the D4 and D5 reasoning extensions are essential. Without raising the ceiling beyond standard-model collapse, reasoning models would appear to solve the dimension.

Each model is evaluated on the same 50 CAS-verified problems for every evaluated dimension level, generated with seed 42. This sample size reflects a practical constraint rather than a theoretical optimum. The evaluation becomes costly at high levels because the prompt size and reasoning burden grow with the active dimension. For example, large D4 settings require a model to process hundreds or thousands of simultaneous products before producing the final sum, which increases input cost, output

Dim.	Scaling Intervention	What Is Held Fixed	Interpretation of Failure
D2	Increase nesting depth from shallow expressions to depth 50 extensions.	Operator family, operand size, and branch width.	Failure indicates nested partial-result propagation limits.
D4	Increase parallel products from $K = 2$ to $K = 3000$.	Each branch remains a trivial single-digit product.	Failure indicates inability to retain many simultaneous intermediate values.
D5	Increase branch-internal operations from 0 to 100.	Exactly three branches and two root operators.	Failure indicates local compositional breakdown before branch outputs combine.
D7	Increase identical repeated terms from $K = 5$ to $K = 300$.	Repeated operand and final arithmetic form.	Failure indicates cardinality/enumeration difficulty rather than arithmetic difficulty.
D8	Increase dependent chain length from 1 to 500 steps.	One live state, small operands, and no parallel branches.	Failure indicates cascading sequential-error accumulation.
D9	Increase multiplication operands from 1 to 15 digits.	Multiplication operator and minimal expression structure.	Failure indicates partial-product and numeric-representation limits.

Table 2: Representative scaling interventions in the one-active-variable protocol. The full nine-dimension schedule is reported in [Appendix A \(level schedules\)](#).

Model	Scale	Type	Architecture	Diagnostic Role
GPT-4o Mini	Undisclosed	Standard	Undisclosed	Hosted small instruction model.
Claude 3.5 Haiku	Undisclosed	Standard	Undisclosed	Hosted small instruction model with strong D7 behaviour.
Qwen3 235B	235B total 22B active	Standard	MoE decoder	Large standard model for scale comparison.
DeepSeek V3	671B total 37B active	Standard	MoE decoder	Large MoE standard model.
Gemma 3 12B	12B	Standard	Decoder-only	Mid-size open-weight model.
Llama 3 8B	8B	Standard	Decoder-only	Small open-weight baseline.
DeepSeek R1 0528	671B total 37B active	Reasoning	MoE + CoT	Reasoning-specialised model with extended traces.
Qwen QwQ	32.5B	Reasoning	Decoder + CoT	Reasoning-specialised model with extended traces.

Table 3: Model coverage in the cross-model evaluation. The selection is designed to compare scale, architecture family, and reasoning-style inference under the same dimension-controlled batteries. Proprietary hosted model sizes are not publicly disclosed; for MoE models, total and active denote total parameters and active parameters per forward pass.

length, and failure-analysis time. We therefore use a fixed 50-item protocol to keep the comparison controlled across models while retaining enough instances to expose threshold behaviour.

Models are queried at temperature zero with no self-consistency, majority voting, tool use, or human correction. The prompt asks for step-by-step reasoning and a final line in the form ANSWER: value. Correctness is computed against the CAS answer using the more permissive of $\pm 0.5\%$ relative error or ± 0.05 absolute error. API errors and CAS-generation errors are tracked separately and excluded from accuracy denominators. Prompt templates and example generated items are given in [Appendix B.2 \(prompt template\)](#) and [Appendix A \(example items\)](#).

6 Results

[Figure 1](#) summarises the first zero-accuracy point for D4, D5, and D7. This view makes the main threshold structure visible without compressing full heatmaps into the main text. D4 has a tightly clustered standard-model collapse, D5 shows branch-local degradation even with fixed branch count, and D7 reveals a different counting profile in which Claude 3.5 Haiku does not reach a zero-accuracy point within the evaluated range. Because a first-zero plot necessarily compresses non-monotonic recoveries, [Table 5](#) states those recoveries explicitly and the appendix provides the full heatmaps.

[Table 5](#) complements the threshold plot by reporting the exact first-zero levels and the observed D7 recoveries. D4 thresholds cluster tightly for standard models even though the models differ substantially in scale and architecture. D5 thresholds

Dim.	Diagnostic Role	Standard-Model Pattern	Reasoning-Model Pattern
D2	Vertical nesting.	Accuracy remains high through shallow depths and collapses for most standard models by depths 7–8.	Threshold moves outward; DeepSeek R1 0528 still degrades at extreme depths.
D4	Horizontal working memory.	Every standard model reaches at least one zero-accuracy point by 40 branches, despite isolated recoveries.	Qwen QwQ fails at 500 branches; DeepSeek R1 0528 fails at 3000.
D5	Branch-local composition.	Standard models separate by 3–8 operations per branch and all reach 0% by 50.	Qwen QwQ fails at 50 operations per branch; DeepSeek R1 0528 fails at 70.
D7	Repeated-token counting.	Highly model-specific: Claude stays perfect, Llama collapses early, and several models recover non-monotonically.	DeepSeek R1 0528 stays perfect through $K = 200$ but fails at $K = 300$; Qwen QwQ fails from $K = 60$ with a partial recovery at $K = 100$.
D8	Sequential dependence.	Chains become destructive after the 10-step region for most standard models.	DeepSeek R1 0528 remains perfect; Qwen QwQ eventually fails at 300 and 500 steps.

Table 4: Main diagnostic findings for the dimensions that most clearly separate model behaviour. Full heatmaps, all nine dimensions, and detailed tables are in [Appendix C \(heatmaps\)](#) and [Appendix C \(full tables\)](#).

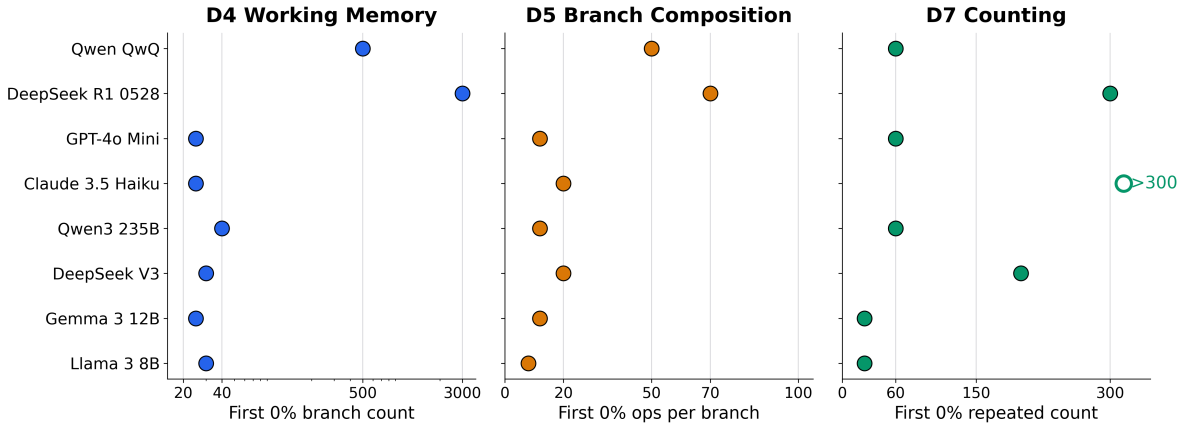


Figure 1: Failure-threshold summary for the three main diagnostic dimensions. Each dot marks the first zero-accuracy level observed for a model: D4 uses the number of parallel branches, D5 uses operations per branch, and D7 uses repeated-term count K . Larger x-axis values indicate a later collapse. The open D7 marker for Claude 3.5 Haiku indicates that no zero-accuracy point appears through $K = 300$. Full heatmaps and detailed tables are provided in [Appendix C \(heatmaps\)](#) and [Appendix C \(full tables\)](#).

are lower for smaller models but remain finite for every standard model. D7 follows a different pattern. The strongest standard model on D7, Claude 3.5 Haiku, is perfect throughout, while several larger or reasoning-specialised systems fail and partially recover. The D7 profile therefore motivates treating counting as a distinct diagnostic axis rather than a side effect of length.

D4 exposes a scale-limited working-memory bottleneck. D4 is the clearest test of simultaneous intermediate-state storage. The threshold plot in [Figure 1](#) shows that the first zero-accuracy D4 point for every standard model lies between 25 and 40 branches. The full D4 heatmap in [Figure C.2](#) gives the surrounding curve: all standard models are perfect through 8 branches, but by 12 branches Gemma 3 12B and Llama 3 8B have slipped to 75%, and the shared failure band appears from 20 to 40 branches. GPT-4o Mini reaches 0% at 25 branches, Qwen3

235B reaches 0% at 40, DeepSeek V3 reaches 0% at 30, and every standard model has at least one zero point by 40. Some models later recover at isolated levels, such as Claude 3.5 Haiku at 40, Qwen3 235B at 50, and DeepSeek V3 at 40, 50, 60, or 150. However, none sustains high-branch reliability. The $29\times$ parameter gap between Qwen3 235B and Llama 3 8B therefore does not translate into a proportionally shifted D4 threshold.

Reasoning-specialised models change the curve but do not remove the cliff. Qwen QwQ remains at 100% through 200 branches except for a 75% point at 40, then falls to 0% at 500, 1000, and 3000 branches. DeepSeek R1 0528 remains perfect through 1000 branches and collapses at 3000. This supports a threshold-shift interpretation: extended reasoning traces act like an external scratchpad, but sufficiently large parallel state still defeats the model.

Model	D4 First 0%	D5 First 0%	D7 Counting Pattern
GPT-4o Mini	25 branches	12 ops/branch	0% at $K = 60$, recovery at $K = 100$, then 0% from $K = 150$.
Claude 3.5 Haiku	25 branches	20 ops/branch	No collapse through $K = 300$; remains 100% at every D7 level.
Qwen3 235B	40 branches	12 ops/branch	0% at $K = 60$, partial recovery at $K = 100$, then 0% from $K = 150$.
DeepSeek V3	30 branches	20 ops/branch	Strong through $K = 150$, then 0% at $K = 200$ and $K = 300$.
Gemma 3 12B	25 branches	12 ops/branch	0% at $K = 25$, recovery at $K = 60$, then 0% from $K = 100$.
Llama 3 8B	30 branches	8 ops/branch	Earliest sustained D7 collapse; 0% from $K = 25$ onward.
DeepSeek R1 0528	3000 branches	70 ops/branch	Perfect through $K = 200$, then 0% at $K = 300$.
Qwen QwQ	500 branches	50 ops/branch	0% at $K = 60$, partial recovery at $K = 100$, then 0% from $K = 150$.

Table 5: Selected failure thresholds from the updated D4, D5, and D7 batteries. The table reports the first zero-accuracy point; non-monotonic recoveries are stated explicitly rather than hidden by a single threshold.

D5 is not reducible to D4. D5 fixes the number of branches at three and varies only the operations inside each branch. This construction produces a failure profile that differs from D4. Llama 3 8B drops to 17% at 3 operations per branch and reaches 0% at 8. GPT-4o Mini and Claude 3.5 Haiku remain perfect through 3 operations but fall sharply at 5–8. Qwen3 235B and DeepSeek V3 survive the shallow levels but reach zero once branch-local composition grows. All standard models are at 0% by 50 operations per branch. Reasoning models again shift the ceiling. Qwen QwQ is non-zero through 30 operations per branch but fails at 50, 70, and 100. DeepSeek R1 0528 remains perfect through 30, drops to 75% at 50, and fails at 70 and 100.

D7 reveals counting failures that structure alone misses. D7 is the least predictable dimension from model size or structural profiles. Claude 3.5 Haiku remains at 100% from $K = 5$ through $K = 300$, while Llama 3 8B is already unstable by $K = 10$ and is zero from $K = 25$ onward. DeepSeek V3 remains strong through $K = 150$ but fails at $K = 200$ and $K = 300$. DeepSeek R1 0528 is perfect through $K = 200$ yet fails at $K = 300$, showing that extended reasoning does not guarantee exact enumeration at long repetition lengths. Qwen QwQ, GPT-4o Mini, Qwen3 235B, and Gemma 3 12B show non-monotonic recoveries. These results are consistent with counting being affected by tokenisation, context position, and output policy rather than by a smooth arithmetic difficulty curve.

D2 and D8 separate nested from sequential dependence. D2 and D8 are structurally coupled because both can create long expression spines, but

they test different computations. D2 requires nested partial results to be returned through an expression tree; D8 requires a single live state to be updated repeatedly. Standard models remain strong on shallow D2 and short D8 chains, then degrade sharply as depth or step count grows. DeepSeek R1 0528 is robust on D8 through 500 steps but still degrades on extreme D2, indicating that explicit reasoning traces are more effective for linear state updates than for deeply nested return structure.

Figure 2 makes the D2/D8 contrast explicit at the level of scaling curves. In D2, most models remain at ceiling through shallow depths and then fall sharply once nested partial-result propagation becomes difficult. In D8, the x-axis is chain length rather than tree depth, and the failure pattern is different: DeepSeek R1 0528 remains at ceiling through 500 steps, while Qwen QwQ fails only at the longest levels and several standard models collapse much earlier. This matters because D2 and D8 can both create long expression spines. Their different curves show that length alone is not the mechanism. The overall nine-dimension heatmap in Figure C.1 is useful for broad model profiling, but the main analysis relies on per-dimension trajectories and thresholds because aggregate views can hide abrupt or mechanism-specific failures.

The remaining dimensions are informative but secondary in the current battery. D1 is mostly a proxy because long expressions often encode depth, counting, or sequential dependence. It remains useful for detecting prompt-length extrapolation, but it rarely identifies a unique algebraic mechanism by itself. D3 remains near ceiling through the present twelve operator levels, so future work should extend it to calculus-level and Diophantine operators.

Accuracy falls differently under nesting depth and sequential chain length

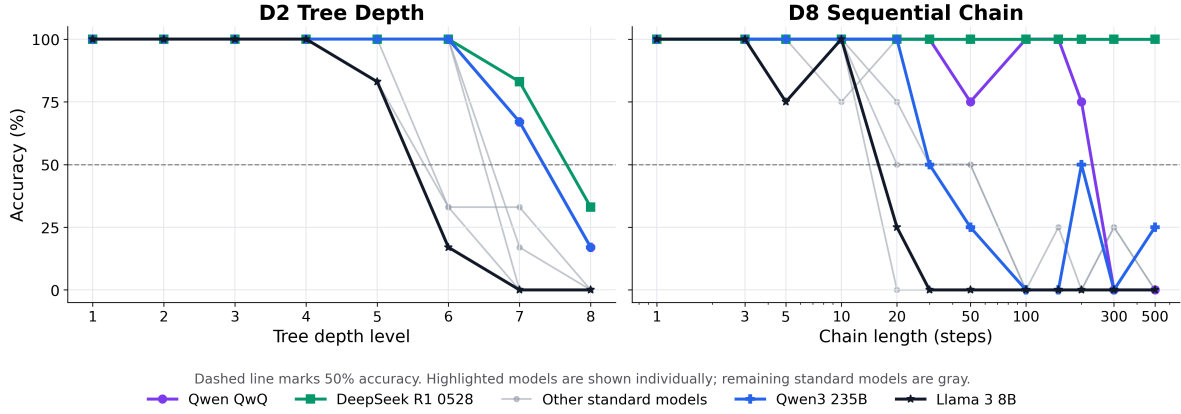


Figure 2: Level-wise accuracy curves for D2 tree depth and D8 sequential chain length. The dashed line marks 50% accuracy. DeepSeek R1 0528, Qwen QwQ, Qwen3 235B, and Llama 3 8B are highlighted individually; the remaining standard models are shown in grey. The curves show that both dimensions degrade with scale, but their trajectories differ despite related surface length. Full heatmaps, including the overall nine-dimension accuracy heatmap, are provided in [Appendix C \(heatmaps\)](#).

D6 is non-monotonic: quadratic factoring and cubic roots are difficult, but identity rewriting is easy despite nominal ambiguity, suggesting that familiar algebraic form matters more than the count of possible solution strategies. D9 amplifies D3 under multiplication because partial products grow quadratically with digit count (Dziri et al., 2023; Yuan et al., 2023); it is most informative when interpreted as an operator–magnitude interaction. These dimensions remain part of the full benchmark because they test mechanisms that are not captured by a single aggregate score.

Reasoning traces shift thresholds, not mechanisms. The reasoning-specialised models are strongest on D1, D4, D5, D8, and D9 because extended traces can externalise intermediate computations. However, their failures at extended D4 and D5 levels show that the generator can still identify the next threshold. This result demonstrates the importance of automatic scaling. If a model reaches ceiling at a current level, the benchmark does not need to be replaced. The same dimension definition can be extended until the failure mode reappears or is convincingly absent.

7 Conclusion

We presented a nine-dimension algebraic complexity framework and an automated CAS-verified generator that scales each dimension systematically. We also reported a cross-model evaluation over six standard instruction-tuned models and two

reasoning-specialised models. The results show that aggregate accuracy hides distinct structural failures. Standard models share a D4 working-memory collapse in the 20–40 branch region, D5 exposes branch-local composition failure, D7 reveals repeated-token counting instability, and D2 and D8 separate nested from sequential dependence. Reasoning-specialised models substantially shift several thresholds, but the extended D4, D5, and D7 batteries show that the underlying limits are displaced rather than removed.

The three contributions are mutually reinforcing. The dimensions provide the explanatory vocabulary by specifying which algebraic pressure is being tested. The automatic generator provides the scaling mechanism by raising one pressure at a time while preserving the controls needed for attribution. The cross-model evaluation shows why both components are necessary. Without controlled dimensions, the shared D4 collapse, the D5/D4 separation, and the unusual D7 counting patterns would be hidden inside a single score. Conversely, without the generator, any observed threshold would be tied to a fixed dataset and would quickly become obsolete as models improve.

The broader lesson is that algebraic reasoning should be evaluated as a structured set of capacities rather than as a monolithic skill. A model may handle shallow syntax but fail nested propagation. It may solve local branches but fail horizontal storage. It may reason through long chains

but miscount repeated tokens. A useful benchmark should make these distinctions visible. The nine-dimension framework is a step toward that goal. It turns algebraic reasoning failure into a phenomenon that can be named, scaled, compared, and tested again as models change.

Limitations

The current D3 battery does not include integration, differentiation, or Diophantine equations, which are likely necessary to expose a harder operator cliff. The highest D4, D5, and D7 levels create long prompts, so future extensions should control context length and output budgets explicitly. The evaluation uses a single generated seed and temperature-zero decoding; this supports deterministic diagnosis but does not estimate variation across generated problem sets. Non-monotonic points in D4, D5, D7, and D8 should therefore be re-run with additional seeds before being interpreted as robust recovery. Finally, the working-memory interpretation is supported by controlled curves and prior theory, but this paper does not directly intervene on model internals.

Acknowledgements

We thank the reviewers for their constructive feedback and our colleagues for discussions during the development of this evaluation framework. The authors acknowledge the use of AI-assisted writing tools, including Claude and Gemini, solely to improve clarity, grammar, and presentation. All experimental designs, generated problems, model evaluations, analyses, findings, interpretations, and conclusions are the authors' own responsibility.

References

- Alibaba Cloud. 2025. QwQ: A thinking model by Qwen. Technical report, Alibaba Cloud.
- Cem Anil, Yuhuai Wu, Anders Andreassen, Aitor Lewkowycz, Vedant Misra, Vinay Ramasesh, Ambrose Slone, Guy Gur-Ari, Ethan Dyer, and Behnam Neyshabur. 2022. [Exploring length generalization in large language models](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 38546–38556. Curran Associates, Inc.
- Yingshan Chang and Yonatan Bisk. 2024. Language models need inductive biases to count inductively. *arXiv preprint arXiv:2405.20131*.
- Shubh Chapra, Dhruv Kumar, Murari Mandal, and Yash Sinha. 2026. [The complexity ceiling benchmark: A multi-domain evaluation of sequential reasoning under depth scaling](#). In *ICML 2026 Workshop on Combining Theory and Benchmarks*.
- Sheng Chen, Chao Tian, Bo Hu, Kang Chen, Zheng Liu, Zheng Zhang, and Ji Zhou. 2025. Arrows of math reasoning data synthesis for large language models: Diversity, complexity and correctness. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 4665–4669.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.
- DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. Technical report, DeepSeek-AI.
- Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Sean Welleck, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena Hwang, Soumya Sanyal, Xiang Ren, Allyson Ettinger, Zaid Harchaoui, and Yejin Choi. 2023. Faith and fate: Limits of transformers on compositionality. In *Advances in Neural Information Processing Systems*, volume 36.
- Jingxuan Fan, Sarah Martinson, Erik Y. Wang, Kaylie Hausknecht, Jonah Brenner, Danxian Liu, Nianli Peng, Corey Wang, and Michael P. Brenner. 2024. HARDMath: A benchmark dataset for challenging problems in applied mathematics. *arXiv preprint arXiv:2410.09988*.
- Dongyu Gong and Hantao Zhang. 2024. Self-attention limits working memory capacity of transformer-based models. *arXiv preprint arXiv:2409.10715*.
- Aryan Gulati, Brando Miranda, Eric Chen, Emily Xia, Kai Fronsdal, Bruno De Moraes Dumont, and Sanmi Koyejo. 2025. [Putnam-AXIOM: A functional & static benchmark for measuring higher level mathematical reasoning in LLMs](#). In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 20723–20747. PMLR.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the MATH dataset. *arXiv preprint arXiv:2103.03874*.
- Arian Hosseini, Alessandro Sordoni, Daniel Toyama, Aaron Courville, and Rishabh Agarwal. 2024. Not all LLM reasoners are created equal. *arXiv preprint arXiv:2410.01748*.
- Kanav Kapoor, Dhruv Kumar, Jagat Sesh Challa, Murari Mandal, and Yash Sinha. 2026. [GraphStateEval: A step-by-step evaluation framework for graph algorithm execution in large language models via intermediate state tracing](#). In *ICML 2026 Workshop on Combining Theory and Benchmarks*.
- Guillaume Lample and François Charton. 2020. Deep learning for symbolic mathematics. In *International Conference on Learning Representations (ICLR)*.

- Chaitanya Manem, Pratik Prabhanjan Brahma, Prakamya Mishra, Zicheng Liu, and Emad Barsoum. 2025. SAND-Math: Using LLMs to generate novel, difficult and useful mathematics questions and answers. In *NeurIPS 2025 Workshop on MATH-AI*.
- Larisa Markeeva, Sean McLeish, Borja Ibarz, Wilfried Bounsi, Olga Kozlova, Alex Vitvitskyi, Charles Blundell, Tom Goldstein, Avi Schwarzschild, and Petar Veličković. 2024. The CLRS-Text algorithmic reasoning language benchmark. *arXiv preprint arXiv:2406.04229*.
- William Merrill and Ashish Sabharwal. 2023. The expressive power of transformers with chain of thought. *arXiv preprint arXiv:2310.07923*.
- Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. 2025. [GSM-Symbolic: Understanding the limitations of mathematical reasoning in large language models](#). In *International Conference on Learning Representations*.
- Mohammad Ramezanali, Mo Vazifeh, and Paolo Santi. 2025. [seqBench: A tunable benchmark to quantify sequential reasoning limits of LLMs](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33771–33792, Suzhou, China. Association for Computational Linguistics.
- David Saxton, Edward Grefenstette, Felix Hill, and Pushmeet Kohli. 2019. Analysing mathematical reasoning abilities of neural models. In *International Conference on Learning Representations (ICLR)*.
- Parshin Shojaee, Iman Mirzadeh, Keivan Alizadeh, Maxwell Horton, Samy Bengio, and Mehrdad Farajtabar. 2025. [The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity](#). *arXiv preprint arXiv:2506.06941*.
- Aaditya K. Singh and DJ Strouse. 2024. [Tokenization counts: the impact of tokenization on arithmetic in frontier LLMs](#). *arXiv preprint arXiv:2402.14903*.
- Rishabh Singh, Sumit Gulwani, and Sriram K. Rajamani. 2012. [Automatically generating algebra problems](#). In *Proceedings of the 26th AAAI Conference on Artificial Intelligence*, pages 1620–1628.
- Peiyang Song, Pengrui Han, and Noah Goodman. 2026. Large language model reasoning failures. *arXiv preprint arXiv:2602.06176*.
- Zhengyang Tang, Xingxing Zhang, Benyou Wang, and Furu Wei. 2024. MathScale: Scaling instruction tuning for mathematical reasoning. *arXiv preprint arXiv:2403.02884*.
- Yi Xu, Roger Smeets, and Rafael Bidarra. 2021. [Procedural generation of problems for elementary math education](#). *International Journal of Serious Games*, 8(2).
- Gilad Yehudai, Haim Kaplan, Guy Dar, Royi Rassin, Asma Ghandeharioun, Mor Geva, and Amir Globerson. 2024. [When can transformers count to n?](#) *arXiv preprint arXiv:2407.15160*.
- Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, and Songfang Huang. 2023. How well do large language models perform in arithmetic tasks? *arXiv preprint arXiv:2304.02015*.
- Xiang Zhang, Juntao Cao, and Chenyu You. 2024. Counting ability of large language models and impact of tokenization. *arXiv preprint arXiv:2410.19730*.
- Jun Zhao, Jingqi Tong, Yurong Mou, Ming Zhang, Qi Zhang, and Xuanjing Huang. 2024. Exploring the compositional deficiency of large language models in mathematical reasoning through trap problems. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 16361–16376.

Appendix Roadmap

Section	Contents
A. Benchmark Specification	Dimension definitions, fixed controls, level schedules, and example generated items.
B. Evaluation and Reproducibility	Models, prompt template, metrics, and evaluation rules.
C. Detailed Results	Overall accuracy summary and dimension-specific tables and heatmaps.
D. Cross-Dimensional Couplings	Directional interaction matrix and generator-level explanations.

A Benchmark Specification

The examples below are illustrative infix renderings of the generator families. The actual benchmark samples 50 CAS-verified instances per level using seed 42. Ellipses indicate repeated generator structure rather than omitted benchmark evidence.

A.1 D1: Syntactic length

D1 measures the prefix-token count of the generated expression. The active failure mechanism is length extrapolation: a model may know the local arithmetic pattern but fail when the same pattern is stretched beyond familiar sequence lengths (Anil et al., 2022; Markeeva et al., 2024). The D1 battery therefore uses simple right-growing addition chains with single-digit operands, ranging from 5 to 751 tokens.

Active variable	Prefix token count
Fixed controls	Single-digit operands, addition chains, no branching beyond the chain template.
Level schedule	5, 11, 21, 51, 75, 101, 201, 501, 751 tokens.
Low level	3+2+6
Middle level	5+8+1+7+4+...
High level	2+3+4+9+1+... with hundreds of prefix tokens.

A.2 D2: Tree depth

D2 measures the longest root-to-leaf path in the expression tree. Increasing depth raises the amount of nested partial-result propagation required before the root can resolve. The isolated battery controls width, operand size, and operator difficulty so that the active pressure is vertical nesting rather than hard arithmetic. This follows prior evidence that neural models lose accuracy as nesting depth increases (Saxton et al., 2019; Dziri et al., 2023).

Active variable	Maximum tree depth
Fixed controls	Simple operators, small operands, no wide branch accumulation.
Level schedule	Depth 1–8 for standard models; extended to 10, 20, and 50 for reasoning models.
Low level	3+2
Middle level	3+7*(6+5)
High level	3+7*(6+5*(2+5*(3+3*4))).

A.3 D3: Operator hardness

D3 measures the identity and intrinsic difficulty of the operator under test. The current battery orders twelve operator levels from low-hardness arithmetic through roots, exponentials, trigonometric operators, and powers while holding structural dimensions at minimum values. The present results remain near ceiling, which suggests that future D3 extensions should include calculus-level and Diophantine operators.

Active variable	Operator identity
Fixed controls	Depth, branch count, operand size, counting load, and chain length fixed.
Level schedule	Twelve ordered levels from negation through pow-hard.
Low level	-7
Middle level	sqrt(49)
High level	pow(7,5) or trigonometric forms.

A.4 D4: Working-memory load

D4 measures the number of independent intermediate results that must coexist before combination. Each problem is a sum of K independent single-digit products. The products are arithmetically simple. The intended challenge is to compute many branch results, retain them, and sum them correctly. This targets the absence of a register-like working memory in standard self-attention (Gong and Zhang, 2024) and connects to thresholded reasoning-collapse results (Shojaee et al., 2025).

Active variable	Number of parallel branches K
Fixed controls	Each branch is a trivial single-digit product.
Level schedule	2, 4, 8, 12, 20, 25, 30, 40, 50, 60, 75, 100, 150, 200, 500, 1000, 3000.
Low level	$3*2+6*5$
Middle level	$3*2+6*5+7*4+...$
High level	A sum of hundreds or thousands of independent products.

A.5 D5: Compositional branching

D5 measures the number of operations inside each branch before that branch resolves to a scalar. The generator fixes three branches joined by two root operators, then increases branch-internal operations from 0 to 100. D5 is separated from D4 by construction. D4 varies how many simple branch outputs must be stored, whereas D5 varies how difficult each branch is before storage (Zhao et al., 2024; Hosseini et al., 2024).

Active variable	Operations per branch
Fixed controls	Exactly three branches joined by two root operators.
Level schedule	0, 1, 2, 3, 5, 8, 12, 20, 30, 50, 70, 100.
Low level	$3+2+6$
Middle level	$(3+2*4)+(8-1*5)+(6+7)$
High level	Three fixed branches, each with tens of internal operations.

A.6 D6: Solution ambiguity

D6 measures the number of structurally distinct valid solution strategies. A linear equation such as $3x+6=-9$ has a nearly canonical path because isolating x is the dominant procedure. A quadratic such as $x^2-5x+6=0$ can be solved by factoring, by applying the quadratic formula, or by testing candidate roots, and these routes expose different intermediate representations. A rational identity such as $(a^2-b^2)/(a+b)$ can be approached by expansion, cancellation after recognising a difference of squares, or substitution-based checking. These examples are not difficult because the final arithmetic is large. They are difficult because the model must select a useful algebraic form before computation becomes straightforward.

The D6 battery therefore treats ambiguity as a strategy-selection burden rather than as a count of final answers. Root-extraction, factoring, rational simplification, and identity-rewrite items can all have compact solutions, but the compact solution is available only after the expression is represented in the right form. This connects D6 to symbolic-mathematics settings where search, form recognition, and route selection interact with model decoding behaviour (Lample and Charton, 2020).

Active variable	Problem type
Fixed controls	Numeric size and expression structure kept small inside each type.
Level schedule	Eight types from linear equations to identity rewrites.
Low level	$3x+6=-9$
Middle level	$x^2-5x+6=0$
High level	$(8^2-2^2)/(8+2)$ or identity-rewrite instances.

A.7 D7: Counting load

D7 measures the repeated-term count K for identical operands. The repeated value and operation are intentionally simple, and paired controls remove the counting demand. Failures are therefore attributed to enumeration and cardinality tracking rather than final arithmetic. This dimension is grounded in work on counting, tokenisation, positional encodings, and context effects (Chang and Bisk, 2024; Zhang et al., 2024; Yehudai et al., 2024).

Active variable	Repeated-term count K
Fixed controls	Repeated operand value and arithmetic operation fixed.
Level schedule	$K = 5, 7, 10, 12, 18, 25, 60, 100, 150, 200, 300$.
Low level	$7+7+7+7+7$
Middle level	7 repeated 60 times
High level	7 repeated 300 times.

A.8 D8: Sequential chain length

D8 measures the longest strictly dependent step path. Each intermediate value becomes the input to the next operation, so an early error contaminates all later steps. Unlike D4, only one live state must be carried forward. This isolates cascading sequential error (Merrill and Sabharwal, 2023; Ramezanali et al., 2025).

Active variable	Dependent chain length
Fixed controls	One live intermediate value, small operands, no branch accumulation.
Level schedule	1, 3, 5, 10, 20, 30, 50, 100, 150, 200, 300, 500 steps.
Low level	<code>start=4; +3</code>
Middle level	<code>((4+3)*2)-5)+...</code>
High level	A 500-step dependent update chain.

A.9 D9: Numeric magnitude

D9 measures the maximum digit count in multiplication operands. The generator fixes the operator to multiplication and increases operand size from 1 to 15 digits. This tests partial-product accumulation: multiplication creates many cross-digit interactions, unlike addition-like carry propagation (Dziri et al., 2023; Yuan et al., 2023). Tokenisation can further change arithmetic behaviour for equivalent numeric values (Singh and Strouse, 2024).

Active variable	Operand digit count
Fixed controls	Operator fixed to multiplication; minimal depth and branch count.
Level schedule	1, 2, 4, 6, 8, and 15 digits.
Low level	$3*2$
Middle level	$4829*7316$
High level	A 15-digit by 15-digit multiplication.

B Evaluation and Reproducibility

B.1 Models

Model	Params	Type	Architecture	Role in Evaluation
GPT-4o Mini	Undisclosed	Standard	Undisclosed	Frontier small instruction model.
Claude 3.5 Haiku	Undisclosed	Standard	Undisclosed	Frontier small instruction model.
Qwen3 235B	235B total 22B active	Standard	MoE decoder	Largest standard model in the evaluation.
DeepSeek V3	671B total 37B active	Standard	MoE decoder	Large mixture-of-experts standard model.
Gemma 3 12B	12B	Standard	Decoder-only	Mid-size open-weight instruction model.
Llama 3 8B	8B	Standard	Decoder-only	Small open-weight baseline.
DeepSeek R1 0528	671B total 37B active	Reasoning	MoE + CoT	Extended reasoning trace model.
Qwen QwQ	32.5B	Reasoning	Decoder + CoT	Extended reasoning trace model.

Table B.1: The eight evaluated models. Proprietary hosted model sizes are not publicly disclosed. For mixture-of-experts models, total and active denote total parameters and parameters active per forward pass.

B.2 Prompt Template

All models received the same instruction template at temperature zero:

Solve the following algebraic expression or equation step by step. Give the final result on a separate line using exactly the format `ANSWER: value`. Do not include any other text after the answer line.

The scorer extracts the final answer line, normalises simple numeric formatting, and compares it to the CAS result. A response with no parseable final answer is counted as a model failure. No model is given tool access, self-consistency sampling, majority voting, or human correction.

B.3 Metrics and Error Accounting

Metric	Definition
Accuracy	Percentage of valid responses within CAS tolerance.
Failure threshold	First level below 50%; later recoveries are reported separately.
First zero-accuracy level	First evaluated level at 0% accuracy, as reported in Table 5 .
Failure pattern	Qualitative curve type: collapse, model-discriminating, or non-monotonic.
Error accounting	API and CAS errors are excluded from accuracy denominators.

Table B.2: Evaluation metrics and error-accounting rules.

C Detailed Results

In these heatmaps, rows correspond to models and columns correspond either to dimensions or to increasing levels within one dimension. Darker green indicates higher accuracy, and red indicates collapse. The captions identify the interpretation most relevant to the main paper.

C.1 Overall Accuracy

Dim.	Dimension	GPT-4o	Claude	Qwen3	DS V3	Gemma	Llama	R1	QwQ
D1	Syntactic Length	60.78%	58.89%	73.33%	65.22%	35.56%	48.22%	100%	82.22%
D2	Tree Depth	75.00%	68.62%	85.50%	77.12%	66.62%	62.50%	89.50%	85.50%
D3	Operator Score	100.00%	100.00%	97.92%	100.00%	97.92%	95.83%	100.00%	100.00%
D4	Working Memory	27.94%	27.94%	29.88%	42.59%	23.53%	32.35%	94.12%	80.88%
D5	Comp. Branching	42.25%	42.25%	45.00%	46.67%	39.17%	26.33%	81.25%	68.33%
D6	Sol. Ambiguity	71.43%	76.14%	88.00%	78.57%	78.43%	48.00%	81.62%	83.38%
D7	Counting Load	61.82%	100%	56.82%	77.73%	54.55%	21.82%	90.91%	61.36%
D8	Seq. Chain Length	33.33%	47.92%	54.17%	52.08%	45.83%	33.33%	100%	79.17%
D9	Numeric Magnitude	70.83%	79.17%	62.50%	62.33%	41.67%	43.67%	100%	87.50%

Table C.1: Per-dimension mean accuracy summary across all nine dimensions.

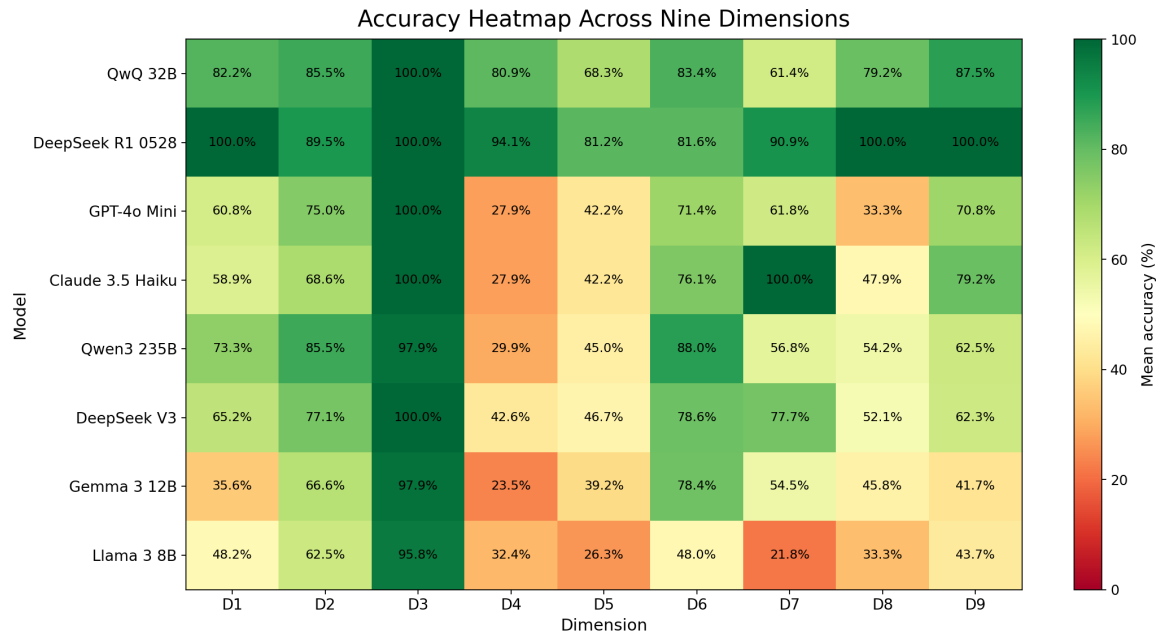


Figure C.1: Overall nine-dimension mean-accuracy heatmap. This appendix view is useful for broad model profiling, but it should be read alongside the per-dimension curves and heatmaps because mean accuracy can hide abrupt threshold failures on individual dimensions.

C.2 D4: Working-memory load

Model	12	20	25	30	40	50	200	500	1000	3000	Type
GPT-4o Mini	100%	75%	0%	0%	0%	0%	0%	0%	0%	0%	Standard
Claude 3.5 Haiku	100%	50%	0%	0%	25%	0%	0%	0%	0%	0%	Standard
Qwen3 235B	100%	25%	25%	25%	0%	33%	0%	0%	0%	0%	Standard
DeepSeek V3	100%	100%	75%	0%	50%	33%	0%	0%	0%	0%	Standard
Gemma 3 12B	75%	25%	0%	0%	0%	0%	0%	0%	0%	0%	Standard
Llama 3 8B	75%	75%	100%	0%	0%	0%	0%	0%	0%	0%	Standard
DeepSeek R1 0528	100%	100%	100%	100%	100%	100%	100%	100%	100%	0%	Reasoning
Qwen QwQ	100%	100%	100%	100%	75%	100%	100%	0%	0%	0%	Reasoning

Table C.2: D4 working-memory accuracy at selected branch counts. Columns denote parallel branches.

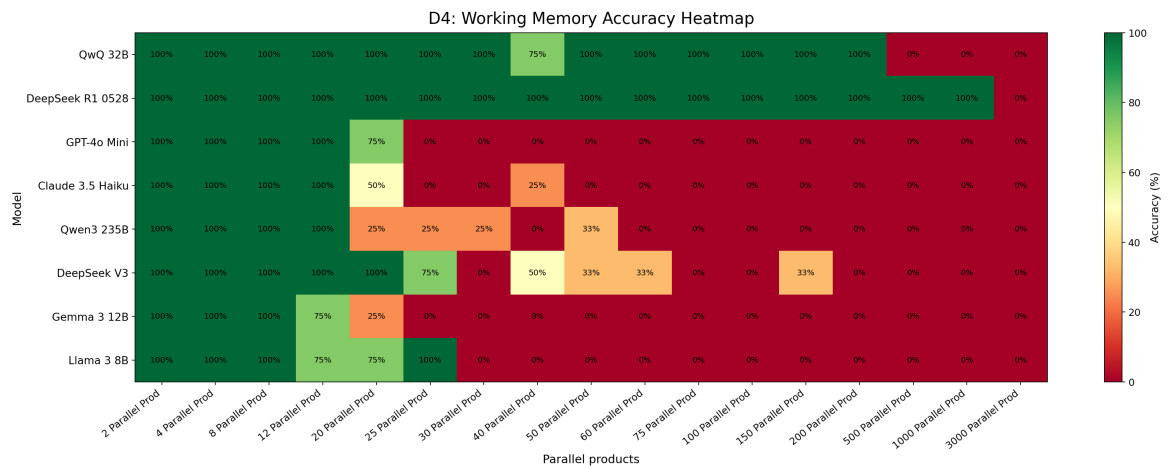


Figure C.2: D4, working-memory load. Standard models remain strong at small branch counts but enter a failure band from roughly 20 to 40 branches; every standard model reaches at least one zero-accuracy point by 40 branches. Qwen QwQ fails at 500, 1000, and 3000 branches, while DeepSeek R1 0528 fails at 3000.

C.3 D5: Compositional branching

Model	0	1	2	3	5	8	12	20	30	50	70	100
GPT-4o Mini	100%	100%	100%	100%	67%	40%	0%	0%	0%	0%	0%	0%
Claude 3.5 Haiku	100%	100%	100%	100%	67%	20%	20%	0%	0%	0%	0%	0%
Qwen3 235B	100%	100%	100%	100%	100%	20%	0%	0%	20%	0%	0%	0%
DeepSeek V3	100%	100%	100%	100%	100%	40%	20%	0%	0%	0%	0%	0%
Gemma 3 12B	100%	100%	100%	83%	67%	20%	0%	0%	0%	0%	0%	0%
Llama 3 8B	100%	83%	83%	17%	33%	0%	0%	0%	0%	0%	0%	0%
DeepSeek R1 0528	100%	100%	100%	100%	100%	100%	100%	100%	100%	75%	0%	0%
Qwen QwQ	100%	100%	100%	100%	100%	80%	100%	80%	60%	0%	0%	0%

Table C.3: D5 branch-complexity accuracy by operations per branch.

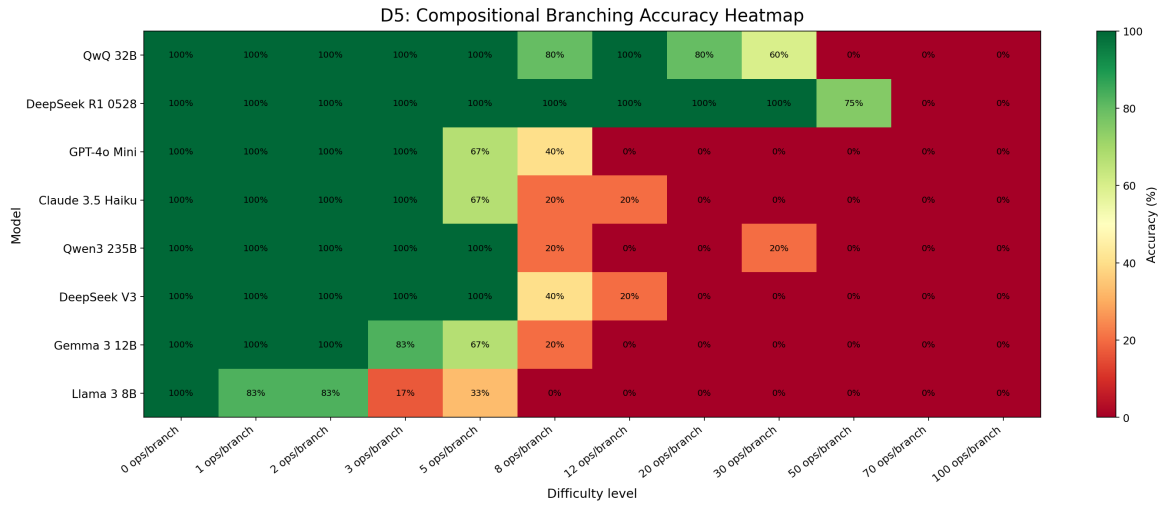


Figure C.3: D5, branch sub-expression complexity. The number of branches is fixed at three; only the operations inside each branch increase. All standard models reach 0% by 50 operations per branch. Qwen QwQ fails at 50, 70, and 100, while DeepSeek R1 0528 fails at 70 and 100.

C.4 D7: Counting load

Model	5	7	10	12	18	25	60	100	150	200	300
GPT-4o Mini	100%	100%	100%	100%	100%	80%	0%	100%	0%	0%	0%
Claude 3.5 Haiku	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
Qwen3 235B	100%	100%	100%	100%	100%	100%	0%	25%	0%	0%	0%
DeepSeek V3	100%	100%	100%	80%	100%	100%	100%	100%	75%	0%	0%
Gemma 3 12B	100%	100%	100%	100%	100%	0%	100%	0%	0%	0%	0%
Llama 3 8B	80%	100%	20%	20%	20%	0%	0%	0%	0%	0%	0%
DeepSeek R1 0528	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	0%
Qwen QwQ	100%	100%	100%	100%	100%	100%	0%	75%	0%	0%	0%

Table C.4: D7 counting-load accuracy by repeated-term count K .

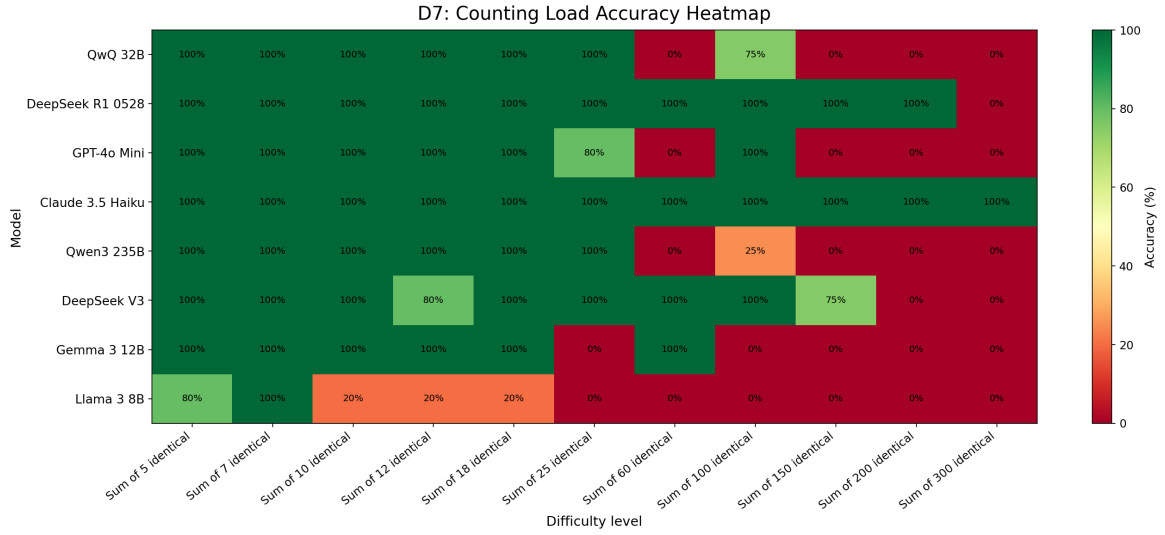


Figure C.4: D7, counting load. Claude 3.5 Haiku stays at 100% through $K = 300$. DeepSeek R1 0528 remains perfect through $K = 200$ but fails at $K = 300$. Several other models show non-monotonic recoveries, confirming that counting is not a smooth function of expression length alone.

C.5 D8: Sequential chain length

Model	1	5	10	20	30	50	100	500
GPT-4o Mini	100%	100%	100%	0%	0%	0%	0%	0%
Claude 3.5 Haiku	100%	100%	100%	50%	50%	50%	0%	0%
Qwen3 235B	100%	100%	100%	100%	50%	25%	0%	25%
DeepSeek V3	100%	100%	100%	75%	50%	50%	0%	0%
Gemma 3 12B	100%	100%	75%	100%	50%	25%	0%	0%
Llama 3 8B	100%	75%	100%	25%	0%	0%	0%	0%
DeepSeek R1 0528	100%	100%	100%	100%	100%	100%	100%	100%
Qwen QwQ	100%	100%	100%	100%	100%	75%	100%	0%

Table C.5: D8 sequential-chain accuracy at selected step counts.

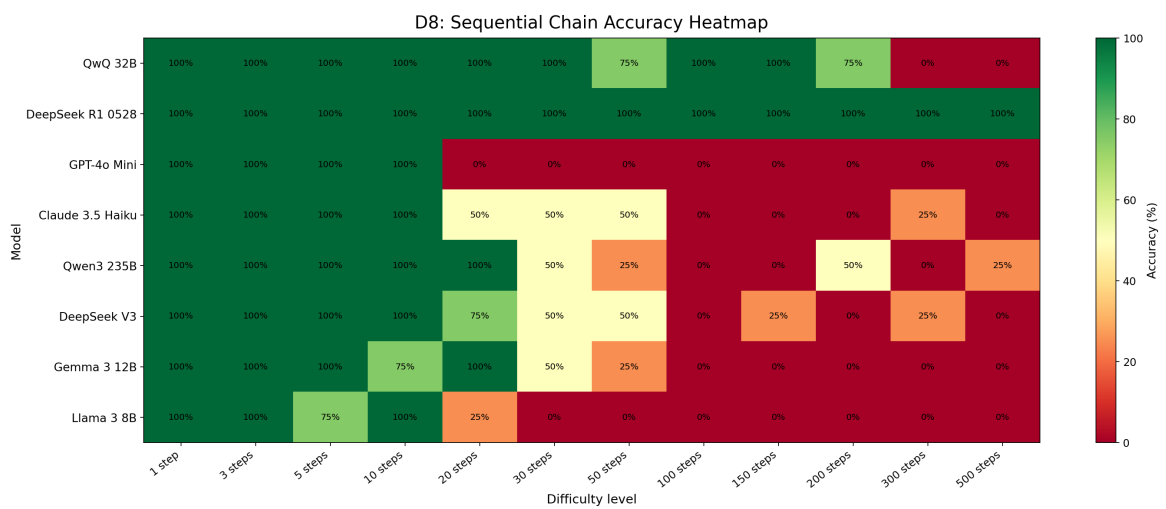


Figure C.5: D8, sequential chain length. Each added step depends on the previous state, so errors cascade through the chain. DeepSeek R1 0528 remains at ceiling, while Qwen QwQ eventually fails at the longest levels.

C.6 D1: Syntactic length

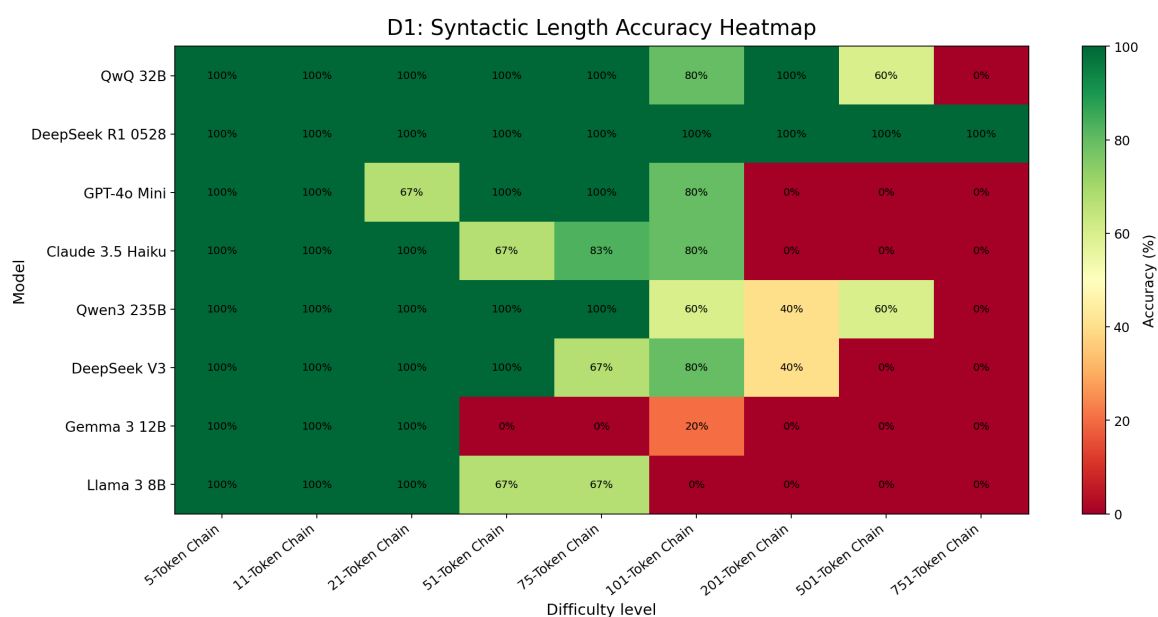


Figure C.6: D1, syntactic length. Long simple chains expose length extrapolation, but D1 is best read as a proxy because long expressions often arise from depth, counting, or sequential structure.

C.7 D2: Tree depth

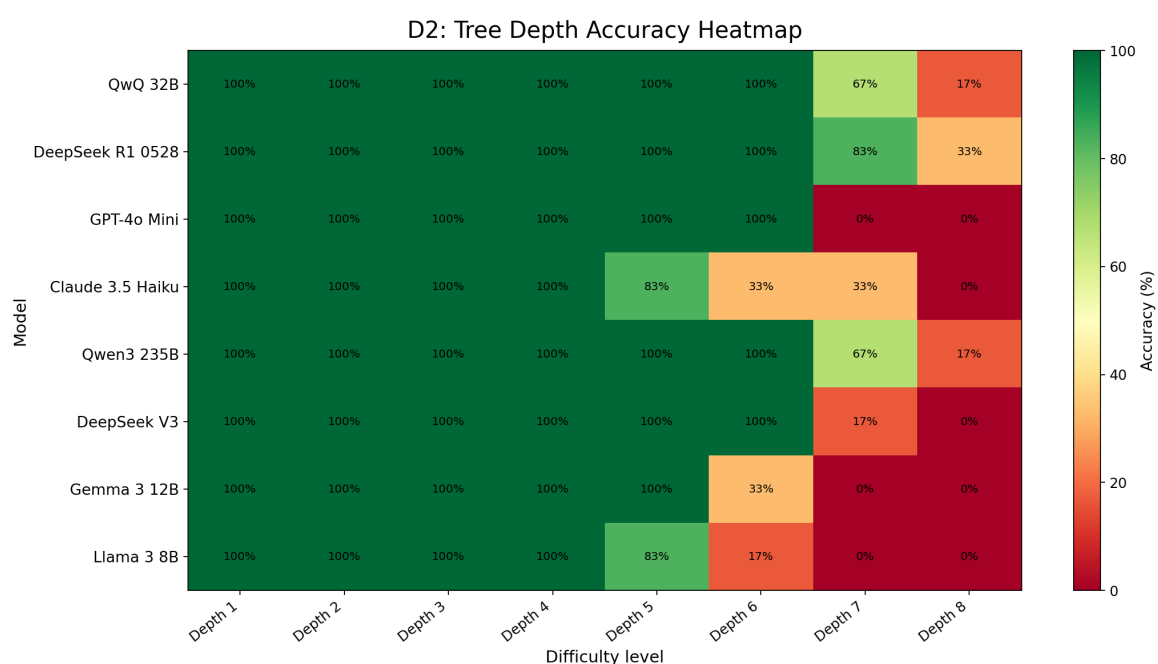


Figure C.7: D2, nesting depth. Accuracy is stable at shallow depths, separates around depths 5–6, and collapses for most standard models by depths 7–8.

C.8 D3: Operator hardness

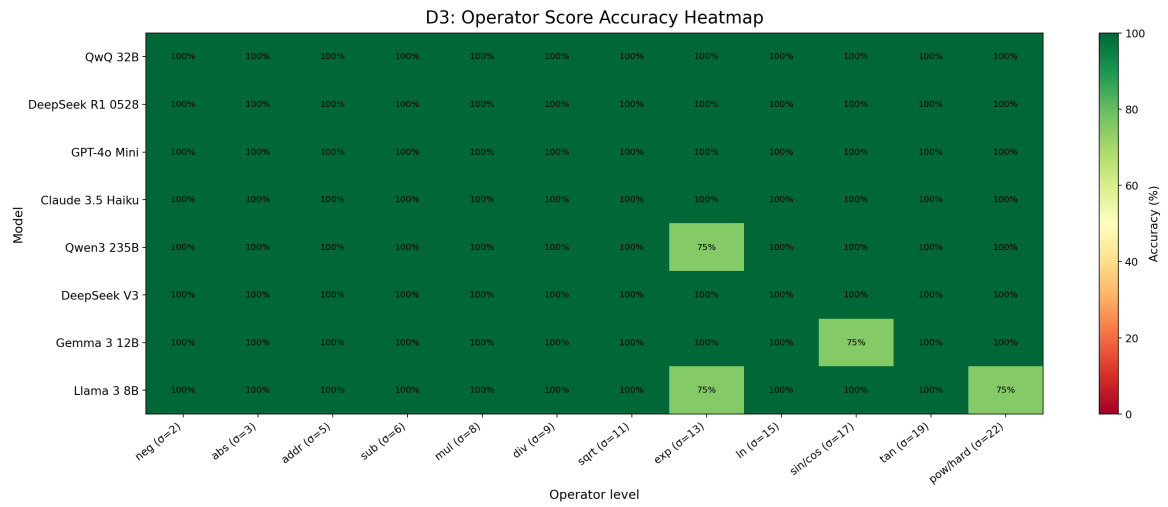


Figure C.8: D3, operator score. Most models remain near ceiling through the current operator ladder, indicating that the next operator cliff likely lies beyond the present battery.

C.9 D6: Solution ambiguity

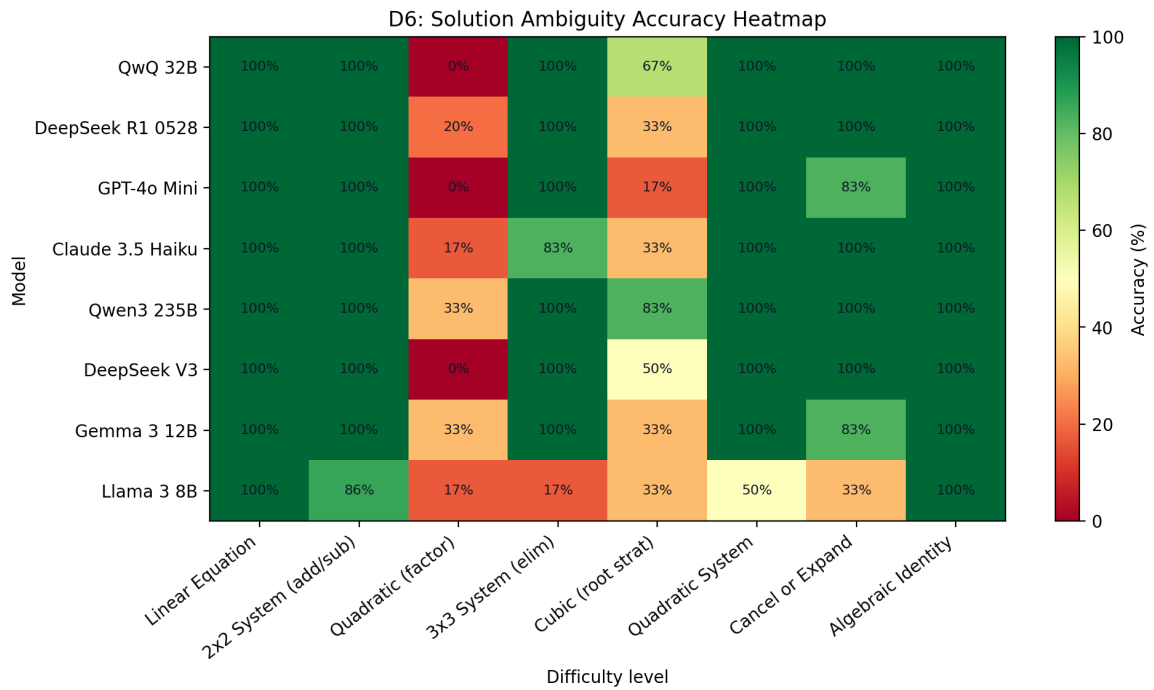


Figure C.9: D6, solution ambiguity. The pattern is non-monotonic: quadratic factoring and cubic-root forms are difficult, while identity rewriting is easy for all models.

C.10 D9: Numeric magnitude

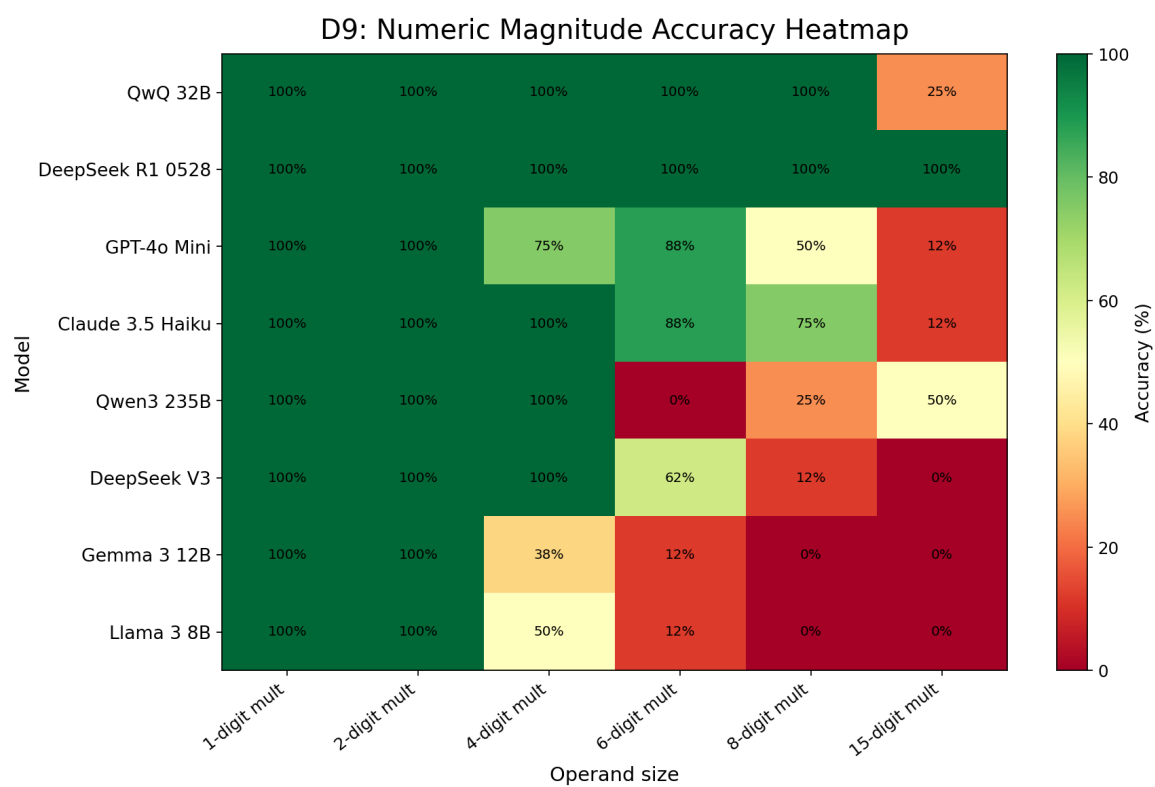


Figure C.10: D9, numeric magnitude. Large operands become difficult mainly under multiplication because the number of partial products grows quadratically with digit count.

D Cross-Dimensional Couplings

Generator-level isolation means one control variable is intentionally varied, but it does not imply that all measured surfaces remain fixed. Table D.1 reports directional co-movement. **S** means structurally guaranteed, **M** means template-induced or high-probability, **W** means weak finite-sample bleed, and **N** means fixed by construction.

Increased ↓ / Affected →	D1	D2	D3	D4	D5	D6	D7	D8	D9
D1 Syntactic Length	—	W	N	N	N	N	S	M	N
D2 Tree Depth	M	—	W	W	N	N	N	S	N
D3 Operator Score	W	N	—	N	N	N	N	N	M
D4 Working Memory	S	S	N	—	N	N	W	N	N
D5 Branch Complexity	S	S	N	N	—	N	N	M	N
D6 Solution Ambiguity	W	N	M	N	N	—	N	N	W
D7 Counting Load	S	M	N	N	N	N	—	W	N
D8 Sequential Chain	S	S	W	N	N	N	N	—	N
D9 Numeric Magnitude	W	N	M	N	N	N	N	N	—

Table D.1: Directional cross-dimensional interaction matrix. Entry (i, j) records how increasing D_i affects D_j .

D.1 D4 to D1 and D2

The D4 generator constructs a right-nested sum of N independent products of the form $a \cdot b$. Each branch contributes three prefix tokens and each joining $+$ contributes one operator, giving $4N - 1$ tokens. At $N = 200$, the expression has 799 tokens; at $N = 3000$, it has 11,999 tokens. The right-nested construction also increases tree depth, so D4-to-D1 and D4-to-D2 couplings are structurally guaranteed.

D.2 D5 to D1, D2, and D8

D5 fixes three branches and increases branch-internal operations. If `sub_ops` denotes operations per branch, token count grows approximately as $3(2 \cdot \text{sub_ops} + 1) + 2$; at 100 operations per branch, this is about 605 prefix tokens. The maximum tree depth grows with branch-internal depth, and a right-spine branch creates a local sequential dependency path, explaining the moderate D5-to-D8 coupling.

D.3 D7 and D8 to Length and Depth

D7 produces a repeated sum with K operands and $K - 1$ joining operators, so the prefix string has $2K - 1$ tokens. When right-nested, it also creates a long expression spine, but the intended mechanism remains enumeration. D8 appends one operator–operand pair per dependent step, giving $2 \cdot \text{steps} + 1$ tokens and a deep expression spine. For this reason, D2 and D8 should be interpreted jointly without being collapsed into one dimension.