ON COMPOSITIONAL LEARNING BEHAVIOURS IN FORMAL MATHEMATICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Language model-based AI agents capable of conquering the Olympiad-level tail of formal mathematics require Compositional Learning Behaviours (CLBs)—the capacity to ground and recombine novel symbolic structures in context, beyond mere recombination of prelearned atoms. We propose **S2B-LM**, an adaptation of the CLB-evaluating Symbolic Behaviour Benchmark that removes numerical processing as a confound and adds chain-of-thought scaffolding to elicit rather than merely probe latent CLB competency. Cross-evaluating ten Lean 4 theorem provers on CLB competency in S2B-LM and miniF2F whole-proof performance, we find correlational and causal evidence of our claim: First, a necessary-condition analysis via quadrant test yields $p = 0.004$, with model scale being ruled out as a confound. Second, extracting a CLB-encoding activation direction from *DeepSeek-Prover-V2-7B* using S2B-LM traces via Contrastive Activation Addition and applying it during miniF2F whole-proof generation on the AIME subset, CLB suppression collapses solve rate from 32.3% to 2.9%, without loss of coherence, while suppressing a random activation direction of equal magnitude leaves it at 31.9%. Together, these results show that CLB competency is *necessary but not sufficient* for the hard tail of formal mathematical verification.

1 INTRODUCTION

Formal Verification and Self-Evolving Agents Foundational to the development of self-evolving scientific agents is the reliance on formal verification loops, where automated theorem provers act as execution environments to provide unyielding reward signals for iterative improvement. Driven by this paradigm, frontier language models have achieved significant performance milestones (Xin et al., 2025; Lin et al., 2025; Ren et al., 2025; Lin et al., 2026) on whole-proof generation tasks by coupling deep neural architectures with intensive proof-tree search algorithms across verification environments such as Lean 4 (Zheng et al., 2022; Yang et al., 2023; Tsoukalas et al., 2024).

The Diagnostic Gap However, standard benchmarking practices typically collapse multidimensional model traits into a monolithic task-accuracy metric. This aggregation masks a critical distinction between shallow statistical pattern matching and genuine systematic abstraction. True autonomous self-evolution requires the acquisition of Compositional Learning Behaviours (CLBs; Denamganaï et al. (2022))—the capacity of an agent to ground never-before-seen atomic elements from in-context evidence and recombine them systematically to resolve novel structural configurations within the same episode, going beyond the mere recombination of prelearned atoms (see also Kim & Linzen (2020)). Without explicit diagnostic benchmarks separating across skills, such as CLBs, or structural compliance, being separated from search capacity, it remains ambiguous whether state-of-the-art provers are acquiring authentic symbolic processing capabilities or are merely leveraging expansive compute budgets to overfit to localized tactical distributions that can be seen in formal proving benchmarks. Following Yang et al. (2025c)'s open challenges and future directions, the aim of this paper is to provide a "controlled stud[y] on synthetic benchmarks for diagnosing reasoning failures" and highlight a way to "improve the model architecture for mathematical reasoning". We acknowledge the Symbolic Behaviour Benchmark (S2B - Denamganaï et al. (2022)) as a crucial diagnostic tool to evaluate the capability of state-of-the-art formal theorem proving LMs to perform CLBs. However, we note that the original version of the benchmark comes with many confounders when used to evaluate LMs. Firstly, the S2B relies heavily on numerical values (cf Section 3.1), whereas LMs notoriously lack numerical understanding and processing abilities (NUPA - Yang et al.

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.

(2025b)). Secondly, the S2B does not account for the competence-performance distinction (Firestone (2020) - cf. Section 2.3), which means that a zero-shot evaluation of CLBs without an adequate reasoning scaffold risks conflating a genuine absence of CLB competency with a mere performance failure—an important confound when interpreting ZSCT scores for LMs.

S2B-LM and Evaluation Protocol To address these confounders, we propose **S2B-LM**, an extension of the S2B that (i) replaces the continuous stimulus domain with a categorical one, removing floating-point segregation as a confound on measured CLB competency, and (ii) introduces a rule-based verbalizer that generates chain-of-thought exemplars from the episode's sync rounds, scaffolding the elicitation of latent CLB competency rather than merely probing it. We report the *adj-ZSCT* (adjusted Zero-Shot Compositional Test score), which maps the raw ZSCT onto an ability-above-chance scale in $[0, 100]$. We cross-evaluate ten state-of-the-art formal theorem proving models, projecting each onto adj-ZSCT alongside a formal verification performance metric. While several benchmarks provide valuable complementary perspectives—including LeanDojo (Yang et al., 2023) and Putnam-Bench (Tsoukalas et al., 2024)—we focus on miniF2F whole-proof test-split pass@32 (Zheng et al., 2022) as it is the evaluation setting we found to be the most consistently reported across papers; notably, 80 of its 244 test-split problems (32.8%) are drawn from competitive or Olympiad tiers (AMC, AIME, IMO), providing sufficient hard-tail coverage to expose the structural bottleneck we investigate. We submit the evaluation pool to a necessary-condition quadrant test, and complement this correlational analysis with a causal intervention via Contrastive Activation Addition on S2B-LM traces applied during AIME proof search.

Findings We find that CLB competency is a *necessary but not sufficient* condition for elite formal mathematics performance. We perform a necessary-condition analysis using a quadrant test, which dichotomises each axis at its empirical median and tests for depletion of a forbidden cell (high miniF2F, low adj-ZSCT) over all $\binom{10}{5} = 252$ label placements. This test finds the forbidden cell **empty** ($p = 0.004$): no model reaches the Olympiad-level tier without measurable CLB competency, while CLB competency alone does not determine ranking within that tier. Model scale is ruled out as a confound: *DeepSeek-Prover-V2-7B* (7B parameters) clears the Olympiad threshold at 75.6% miniF2F, demonstrating that scale is not a necessary condition for the hard tail. Beyond this correlational evidence, we seek to establish causality by extracting a CLB-encoding direction from DSP-V2-7B via Contrastive Activation Addition on S2B-LM traces and apply it during miniF2F whole-proof generation on the AIME subset: CLB suppression collapses solve rate from 32.3% to 2.9% while a random direction of equal magnitude leaves it at 31.9%, providing direct causal evidence that the CLB representational direction is necessary for Olympiad-tier theorem proving.

Contributions. We provide empirical findings that the capability to adapt to novel compositional structures (i.e. receptivity aspects of performing CLBs) is necessary, but not sufficient, for unlocking the hard tail of formal mathematical verification. We propose the S2B-LM as an extension of the Symbolic Behaviour Benchmark that more precisely evaluates the capabilities of LMs to perform CLBs, by introducing two features to control for distinct confounders: a categorical stimulus domain to remove NUPA as a confound, and a rule-based verbalizer providing chain-of-thought scaffolding to address the competence–performance distinction in LMs.

2 Background

2.1 Compositional Behaviours vs Compositional Learning Behaviours

Definitions. Denamganaï et al. (2022) define *Compositional Behaviours* (CBs) as the capacity to generalise from combinations of *trained-on* atomic components to novel re-combinations of those same atoms: the entire vocabulary is prelearned, and generalisation consists of applying known compositional rules to known elements in unseen arrangements. *Compositional Learning Behaviours* (CLBs) extend this to an *online* regime: the agent must generalise from a handful of in-context examples of *never-before-seen* atomic components to novel re-combinations of those atoms, all within a single episode. The within-episode grounding requirement is what distinguishes CLBs from standard compositional generalisation benchmarks, which test training-time generalisation over a prelearned vocabulary.

Code generation vs. Lean proof generation: CB and CLB in practice. Standard code generation is predominantly a CB task: a model trained on Python generalises by composing known identifiers—

`sorted()`, list comprehensions, lambda expressions—into novel arrangements, recombining a fully prelearned vocabulary according to known rules. Lean 4 proof generation introduces a qualitative shift. Lean's reflexive type-checking kernel acts as an unyielding execution environment: every proof step is machine-verified, leaving no room for syntactic hallucination. When a self-evolving prover or a human collaborator introduces a new tactic, a freshly-defined auxiliary lemma, or a novel intermediate proposition—analogous to a mathematician coining a fresh binary relation $\lesssim$ and illustrating its semantics across two or three in-context examples—that object is absent from the model's pretraining weights and cannot be recovered by memorisation. Grounding it from in-context evidence and deploying it correctly in a novel proof context requires decoding the new convention (receptivity) and recombining it with prelearned primitives in unseen arrangements (compositionality)—the full CLB requirement. Thus, while standard code generation and low-to-medium-difficulty theorem proving are accessible via CB capacity alone, advancing into the hard tail of formal mathematical verification demands within-episode symbolic acquisition that current training paradigms do not explicitly cultivate.

CLB as symbolic behaviour. Santoro et al. (2021) argue that symbolic behaviours do not inhere in agents intrinsically but exist only in relation to an interpreter: a system exhibits symbolic behaviour when another agent treats its outputs as symbols. Among the different aspects of symbolic behaviours, this work focuses on two (Denamganaï et al., 2022): *receptivity*—the capacity to decode a novel symbolic convention introduced by a speaker agent—and *constructivity*—the capacity to produce a novel convention that a listener agent can decode. The S2B operationalises both by swapping the role assigned to the tested system (cf. Section 2.4); the current work focuses solely on receptivity, as we justify in Appendix A.1 why it is necessary for downstream constructivity, thus making receptivity a necessary condition on prover competence rather than a complete diagnostic of it; constructivity is an independent dimension that we leave to future works.

2.2 Formal Mathematical Reasoning & Formal Theorem Proving

Learning-based automated theorem proving in Lean 4 has advanced rapidly: frontier whole-proof generation models—DSP-V1.5 (Xin et al., 2025), Goedel-Prover (Lin et al., 2025), DSP-V2 (Ren et al., 2025), and Goedel-Prover-V2 (Lin et al., 2026)—push miniF2F pass@32 from 48% to 88% via SFT on large formal corpora combined with Monte Carlo tree search or RL-based subgoal decomposition. Despite these gains, all operate over a fixed, prelearned tactic vocabulary: they recombine known atomic proof steps (CB), rather than acquiring new symbolic conventions in context (CLB). We evaluate on miniF2F-test whole-proof pass@32 (Zheng et al., 2022) (Table 2); while LeanDojo (Yang et al., 2023) and PutnamBench (Tsoukalas et al., 2024) offer complementary perspectives, our choice is guided by the consistent availability of pass@32 results across all ten evaluated models, its single-pass generation regime comparable to the S2B-LM listener evaluation, and its difficulty span—MATH levels 1–5 to IMO, with 80 of 244 test problems from AMC/AIME/IMO—providing the hard-tail coverage needed to expose a CLB bottleneck.

2.3 How to compare human and AI capabilities ?

A fundamental challenge in AI evaluation is distinguishing a *lack of capability* from a *failure to demonstrate* it under specific test conditions—the *competence–performance* distinction (Firestone, 2020), familiar from comparative psychology where subjects cannot be given explicit instructions. Applied to CLB evaluation, an LM may possess the underlying receptive CLB competency yet score near chance if the prompting context does not adequately scaffold its elicitation. Following Lampinen (2024), we address this by adapting the prompting context to a few-shot chain-of-thought regime that *elicits* rather than merely probes latent competency (cf. Section 3.2); Section 3.3 provides an ablation study confirming this design choice.

2.4 Symbolic Behaviour Benchmark

Santoro et al. (2021) argue that symbolic behaviours do not reside in agents intrinsically but exist only in relation to an interpreter: a system exhibits symbolic behaviour when another agent treats its outputs as symbols. The Symbolic Behaviour Benchmark (S2B; Denamganaï et al. 2022) operationalises this by requiring a *speaker* agent and a *listener* agent to acquire and deploy a shared symbolic code within

Meta-Referential Games (Meta-RGs), a meta-learning extension of referential games in which the latent symbolic structure and the communication vocabulary are freshly randomised at every episode.

Receptivity and constructivity. The S2B evaluates two aspects of CLBs by swapping the role assigned to the tested system. Placing it in the *listener* role—paired with a fixed posdis-compositional rule-based speaker (Chaabouni et al., 2020)—isolates *receptivity*: the capacity to decode an out-of-distribution symbolic convention. Placing it in the *speaker* role—paired with a rule-based listener—isolates *constructivity*: the capacity to produce a novel symbolic convention. In this work we focus exclusively on the receptivity axis.

Meta-RG structure and ZSCT. Each Meta-RG episode is governed by a latent symbolic structure $(d(1), \ldots, d(N_{\text{dim}}))$, where $d(i)$ is the number of possible values on latent dimension i, and by a freshly sampled random vocabulary permutation that bijectively re-indexes the speaker's tokens before they reach the listener, preventing the listener from exploiting an episode-invariant code. The posdis-speaker encodes latent value l_i at message position i as token $t_i = l_i + 1$ (an offset-1 positional code); the permutation then scrambles these indices, so the listener must rediscover the token-to-value mapping from scratch at every episode. Each game in an episode consists of N communication rounds followed by the listener's decision; after each game a dedicated sync round exposes the speaker's exact target stimulus to the listener unconditionally, providing a ground-truth token-to-value update regardless of whether the listener's decision was correct. Episodes unfold in two sequential phases. In the *supporting phase*, games are played until every value on every dimension has appeared in at least S full N_{dim}-dimensional stimulus vectors. The *querying/ZSCT phase* then presents a held-out set of stimulus combinations—multi-dimensional arrangements of those same values that were withheld as targets during the supporting phase—and accuracy on these games defines the Zero-Shot Compositional Test (ZSCT) metric. Success requires the agent to generalise systematically to novel combinatorial arrangements it has never encountered as targets.

Binding problems and the SCS representation. Stimuli are encoded using the Symbolic Continuous Stimulus (SCS) scheme: each dimension i partitions $[-1, +1]$ into $d(i)$ sections, and a stimulus entry is a real number sampled from the Gaussian $\mathcal{N}(\mu_{l(i)}, \sigma^2_{l(i)})$ associated with latent value $l(i)$. The *shape invariance property* ensures that every SCS stimulus is a vector in $[-1, +1]^{N_{\text{dim}}}$ regardless of the $d(i)$ values, so the structure of the current episode cannot be read off from any single observation: the agent must segregate floating-point coordinates over multiple consecutive supporting-phase games to infer both (i) how many distinct values are active per dimension and (ii) which Gaussian cluster each coordinate belongs to. Sub-problem (i) is the less difficult component of the binding problem; sub-problem (ii) is the harder, numerically demanding component. While this domain-agnostic design is appropriate for Deep RL agents, it introduces a systematic confounder for LMs, whose poor numerical understanding and processing abilities (Yang et al., 2025b) directly suppress ZSCT scores independently of any CLB competency.

3 Method: Adapting the Symbolic Behaviour Benchmark

3.1 Adapting the Domain

The original S2B represents stimuli using the Symbolic Continuous Stimulus (SCS) scheme (Denamganaï et al., 2022). Given a latent symbolic structure with N_{dim} dimensions described by the tuple $(d(1), \ldots, d(N_{\text{dim}}))$, a stimulus is a vector in $[-1, +1]^{N_{\text{dim}}}$: the i-th entry is a real number sampled from the Gaussian $\mathcal{N}(\mu_{l(i)}, \sigma^2_{l(i)})$ associated with the currently-selected latent value $l(i) \in \{1, \ldots, d(i)\}$ on dimension i. Thanks to the shape invariance property (cf. Section 2.4), this scheme instantiates a domain-agnostic binding problem that is appropriate for Deep RL agents. However, for LMs, the Gaussian-sampled real values are a systematic confounder: numerical understand-

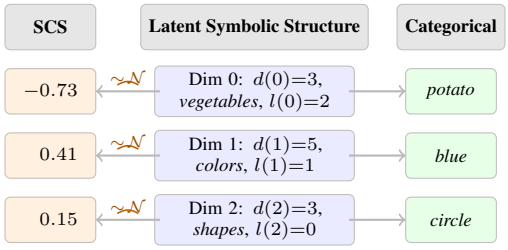

Figure 1: SCS vs. categorical encoding of the same latent stimulus ($l(0)$=2, $l(1)$=1, $l(2)$=0). The SCS scheme maps each latent value to a Gaussian-sampled float ($\sim\mathcal{N}$), requiring floating-point segregation as a prerequisite to CLB evaluation. The categorical scheme replaces floats with named tokens, removing that confound while preserving the binding problem.

ing is not required for compositional reasoning, yet low NUPA (Yang et al., 2025b) directly suppresses measured ZSCT scores, breaking the internal validity of the benchmark as a diagnostic of CLBs.

The Categorical Domain. We replace the SCS representation with a *categorical domain* in which every latent value is a human-readable word. A fixed registry of ten concept classes is defined—*vegetables, fruits, colors, shapes, animals, countries, metals, planets, sports, instruments*—each containing up to ten named items (e.g. *colors* $\mapsto$ {red, blue, green, yellow, ...}). At the start of each episode, N_{dim} categories are sampled without replacement from the registry, and for each selected category i, exactly $d(i) \in [V_{\text{min}}, V_{\text{max}}]$ items are drawn uniformly, forming the episode's latent value set for that dimension. A stimulus on dimension i with latent value $l(i)$ is then presented as the corresponding item name rather than a continuous float. The full N_{dim}-dimensional stimulus is an ordered tuple of natural language tokens—e.g. (*carrot*, *blue*, *circle*) for a latent symbolic structure with three dimensions using categories *vegetables*, *colors*, *shapes*.

Preservation of the Binding Problem. The categorical substitution preserves the binding problem structure. The set of active items per dimension is unknown to the tested model *a priori*: although each item name is semantically familiar, the mapping from category items to latent positions is randomised at every episode. The model must therefore integrate evidence across the supporting-phase stimuli to infer which items are active on each dimension and in which combinatorial arrangement—the same latent-structure discovery challenge as in the SCS domain, now expressed entirely in natural language tokens. The train/test combinatorial split strategy and ZSCT evaluation protocol are identical to the original S2B (cf. Section 2.4).

3.2 Providing Few-Shot Exemplars via a Rule-Based Listener Verbalizer

A second key design objective of S2B-LM is to address the competence–performance distinction (cf. Section 2.3): a tested LM may possess the underlying CLB competency yet score near chance if the prompting context does not adequately scaffold its elicitation. We therefore provide the tested LM with few-shot chain-of-thought exemplars that supply the structured reasoning context needed to *elicit* rather than merely probe its latent CLB competency. We construct these exemplars via a hypothesis-tracking rule-based listener verbalizer, a dedicated agent that monitors the episode and generates human-readable reasoning traces in real time from the sync rounds of the S2B (cf. Section 2.4). The sync round is the game mechanism that unconditionally reveals the speaker's exact target stimulus after every game, making it a natural and game-internal source of ground-truth evidence; the verbalizer converts this evidence into interpretable reasoning steps that are appended to the LM's context before each game decision (see Algorithm 1 in Appendix B for the full episode loop; Figure 2 for a compact illustrative game exchange; and Figure 8 in Appendix C for a complete supporting-phase conversation).

S2B-LM: one supporting-phase game ($N_{\text{dim}} = 3$, *instruments, sports, vegetables*)

[User — Sync & Game #1]
Sync: speaker observed `[piano, swimming, eggplant]` at game #0. Your decision was *similar*—**correct**.

At game #1, you observe `[piano, golf, pepper]`. Speaker's message: `[8, 11, 13]`.
Same latent meaning? Answer 0 (Yes) or 1 (No).

[Listener — Game #1]
Let's think step by step and leverage past games.

From the last game syncing, we can learn that: symbol 8 at pos 0 → *piano* ; symbol 5 at pos 1 → *swimming* ; symbol 6 at pos 2 → *eggplant*.

In the current game, if the speaker were observing a similar stimulus as ours, [*piano*, *golf*, *pepper*], then: at pos 0, *piano* → symbol 8 (from game #0) ; at pos 1, *golf* has not been observed yet ; at pos 2, *pepper* has not been observed yet.

Since the speaker's message is `[8, 11, 13]`, yield 1/3 matches, they are likely observing a *different* stimulus. **Answer: 1**

Figure 2: Excerpt of one S2B-LM supporting-phase game illustrating the verbalizer's sync-summary → inverse-prediction → match-comparison reasoning chain. Full conversation in Figure 8 (Appendix C).

In each supporting-phase game, the tested LM—acting as listener—receives two inputs: its own *stimulus*, an ordered tuple of natural-language category items such as `[piano, swimming, eggplant]`, and the posdis-speaker's *message*, an integer-token sequence such as `[8, 5, 6]`. The LM must decide whether both agents are observing stimuli with the same latent meaning. The sync round that follows unconditionally reveals the speaker's exact target stimulus, providing ground-truth evidence on the token-to-value correspondence at every message position regardless of whether the LM's decision was correct.

The rule-based verbalizer maintains a *value map* $\mathcal{V} : (\text{pos}, \text{tok}) \mapsto \{\text{val} \mapsto \text{count}\}$ that accumulates this sync-derived evidence across all preceding games in the episode. Before each game, the current state of $\mathcal{V}$ is converted into a natural-language reasoning trace appended to the LM's context, providing chain-of-thought exemplars (Wei et al., 2022b) that prime the LM to replicate the verbalizer's compositional inference on the unseen querying-phase stimuli.

Each verbalized trace follows a three-step template: (i) *sync summary*—what the most recent sync round revealed about token-to-value correspondences; (ii) *inverse prediction*—what message the speaker would have sent had it observed the same stimulus as the listener, derived by inverting $\mathcal{V}$; (iii) *match comparison*—how many positions of the actual message agree with the prediction, yielding the same/different decision. By exposing both the accumulated mapping evidence and the compositional reasoning chain used to reach each answer, these traces prime the tested LM to replicate the same inference pattern on the unseen querying-phase stimuli.

3.3 Validation via Ablation Study

We show via an ablation study that our extended version of the S2B has increased internal validity towards evaluating CLBs in tested LMs.

Because the raw ZSCT operates with an intrinsic 50% random-guessing floor, we report the *adjusted ZSCT* (adj-ZSCT), which maps ZSCT onto a pure ability-above-chance scale in $[0, 100]$:

$$\text{adj-ZSCT} = \max\left(0,\ \frac{\text{ZSCT} - 50}{100 - 50}\right) \times 100 \tag{1}$$

Table 1: Ablation results for DSP-V2-7B across benchmark configurations. ZSCT scores reported as mean $\pm$ std. err. over 4 random seeds. The original S2B uses the SCS domain with no CoT scaffold; S2B-LM uses the categorical domain with 10-shot chain-of-thought exemplars.

Configuration	ZSCT (%)	Adj. ZSCT
S2B (SCS, 0-shot)	57.9 ± 5.7	15.8
Categorical + 0-shot	40.0 ± 8.5	0.0
S2B-LM (Categorical + 10-shot)	87.0 ± 4.1	74.0

Table 1 reports ZSCT and adj-ZSCT scores for DeepSeek-Prover-V2-7B under three benchmark configurations, evaluated over 4 random seeds, as an illustration of successful elicitation.

The three configurations isolate the contribution of each design choice. First, the original S2B (SCS, 0-shot) yields only 57.9%—marginally above the 50% guessing floor (adj-ZSCT $= 15.8$)—consistent with NUPA (Yang et al., 2025b) suppressing measured CLB scores independently of any compositional competency. Second, switching to the categorical domain while retaining zero-shot prompting (Categorical + 0-shot) drops performance *below* chance (40.0%, adj-ZSCT $= 0$): without a CoT scaffold to elicit the latent reasoning strategy, the model fails to exploit the now-interpretable stimuli, illustrating the competence–performance distinction. Third, the full S2B-LM design (categorical + 10-shot chain-of-thought) recovers strongly to 87.0% (adj-ZSCT $= 74.0$), demonstrating that the verbalizer-generated exemplars successfully surface the model's underlying CLB competency. Together, these results confirm that both design choices are necessary.

4 Experimental Results

Table 2: Evaluation results: adj-ZSCT (S2B-LM receptive listener CLB competency, ability above chance) and miniF2F whole-proof test-split accuracy (Pass@32). ZSCT scores reported as mean $\pm$ std. err. over 8 random seeds. †: model accessed via API.

Model	Size	Base Model	ZSCT (%)	Adj. ZSCT ($[0-100]$)	miniF2F test Pass@32 (%)
DeepSeek-Prover-V1 (Xin et al., 2024)	7B	DeepSeekMath-7B (Shao et al., 2024)	47.6 ± 6.7	0.0	46.1
DeepSeek-Prover-V1.5-SFT (Xin et al., 2025)	7B	DSP-V1.5-Base (Xin et al., 2025)	56.0 ± 1.6	12.0	48.2
DeepSeek-Prover-V1.5-RL (Xin et al., 2025)	7B	DSP-V1.5-SFT (Xin et al., 2025)	56.9 ± 3.3	13.8	50.0
Goedel-Prover-SFT (Lin et al., 2025)	7B	DSP-V1.5-Base (Xin et al., 2025)	58.1 ± 3.0	16.3	57.6
Goedel-Prover-DPO (Lin et al., 2025)	7B	Goedel-Prover-SFT (Lin et al., 2025)	55.3 ± 3.4	10.6	60.3
Goedel-Prover-V2-8B (Lin et al., 2026)	8B	Qwen3-8B (Yang et al., 2025a)	75.0 ± 7.5	50.0	84.6
Goedel-Prover-V2-32B† (Lin et al., 2026)	32B	Qwen3-32B (Yang et al., 2025a)	65.2 ± 6.4	30.4	88.1
DeepSeek-Prover-V2-7B (Ren et al., 2025)	7B	DSP-V1.5-Base-7B (Xin et al., 2025)	89.8 ± 2.7	79.6	75.6
DeepSeek-Prover-V2-671B† (Ren et al., 2025)	671B	DeepSeek-V3-Base (DeepSeek-AI et al., 2025)	100.0 ± 0.0	100.0	82.4
Kimina-Prover† (Wang et al., 2025b;a)	72B	Qwen2.5-72B (Qwen et al., 2025)	97.4 ± 1.6	94.8	84.0

Experimental setup. miniF2F results report whole-proof test-split accuracy under Pass@32. S2B-LM evaluations use $N_{\text{dim}} = 3$ latent dimensions with $V_{\min} = 3$ and $V_{\max} = 5$ values per latent dimension, number of object-centric samples $O = 1$ (the categorical domain does not support object-centric sampling), and number of shots $S = 1$ (every atomic element appears in at least one target

Correct verbalization (positive)

Let's think step by step and leverage past games.

From the last game syncing, we can learn that: **symbol 3 at pos 0 → violin ; symbol 5 at pos 0 → trumpet ; symbol 7 at pos 1 → football.**

In the current game, if the speaker were observing a similar stimulus as ours, [**violin**, **football**], then: at pos 0, violin → symbol 3 (from game #2) ; at pos 1, football → symbol 7 (from game #1).

Since the speaker's message is [3, 7], yield **2/2** matches, they are likely observing a **similar** stimulus.

Answer: **0**

Corrupted verbalization (negative)

Let's think step by step and leverage past games.

From the last game syncing, we can learn that: **symbol 3 at pos 0 → trumpet ; symbol 5 at pos 0 → violin ; symbol 7 at pos 1 → football.**

In the current game, if the speaker were observing a similar stimulus as ours, [**violin**, **football**], then: at pos 0, violin → symbol 5 (from game #1) ; at pos 1, football → symbol 7 (from game #1).

Since the speaker's message is [3, 7], yield **1/2** matches, they are likely observing a **different** stimulus.

Answer: **1**

Figure 4: Synthetic example ($N_{\text{dim}} = 2$) of a matched teacher-forced pair used to extract $\mathbf{v}_{\text{CLB}}$. Both verbalizations share the identical game structure (message `[3, 7]`, listener stimulus `[violin, football]`). The **correct** trace (left) uses the true hypothesis map and predicts a 2/2 match, yielding decision 0. The **corrupted** trace (right) uses a cyclically shifted binding map at pos 0 (`symbol 3 → trumpet`, `symbol 5 → violin`) — the text is surface-coherent but the inverse prediction at pos 0 is wrong, reducing to 1/2 matches and flipping the decision to 1. The per-layer mean residual over each reasoning span is subtracted (correct − corrupted) to obtain the CLB activation direction.

stimulus before the querying phase begins). All reported statistics are computed over 8 random seeds per model. Adj-ZSCT scores are computed via Equation 1. Table 2 logs the evaluation results across the evaluated model suite. Figure 3 visualizes the empirical scatter of adj-ZSCT against miniF2F accuracy.

4.1 Correlational Evidence

We perform a necessary-condition analysis via a *quadrant test*: each axis is dichotomised at its empirical median (adj-ZSCT median = 23.3; miniF2F median = 67.95%), and we test whether the *forbidden cell*—high miniF2F with low adj-ZSCT—is depleted relative to random pairings. Significance is assessed via a one-sided Fisher's exact test by exact enumeration of all $\binom{10}{5} = 252$ high-performance label placements (margins fixed). Figure 3 displays these median thresholds as dashed reference lines.

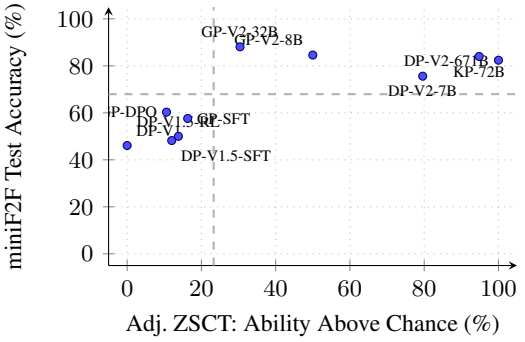

Figure 3: Empirical scatter of adj-ZSCT against miniF2F accuracy. The adj-ZSCT axis captures structured receptive CLB capability adjusted above the 50% random-guessing baseline.

In the observed data, the forbidden cell is **empty**: no model achieves high miniF2F with low adj-ZSCT. With a resulting p-value of $p = 0.004$, this result is establishing CLB competency as a necessary condition for crossing the Olympiad-performance threshold. We report on further complementary correlational evidence in Appendix D.

Limitations. Throughout, we rely exclusively on exact permutation methods rather than parametric or asymptotic estimators, as the small evaluation pool ($N = 10$) renders standard asymptotic tools structurally invalid. We provide further statistical rationale in Appendix E. A competing explanation is model scale: larger models may jointly drive both adj-ZSCT and miniF2F performances. Crucially, however, *DeepSeek-Prover-V2-7B* (7B parameters) achieves 75.6% miniF2F—placing it in the high-performance half—while remaining a small model. This shows that scale is not a necessary condition for the hard tail. However, the current result is correlational only. Thus, in the next section, we turn to causal evidence to strengthen the CLB account beyond correlation.

4.2 Causal Evidence

CLB-encoding direction via CAA. To move beyond correlation, we extract a CLB-encoding activation direction from *DeepSeek-Prover-V2-7B* using Contrastive Activation Addition (CAA), following Turner et al. (2023) and Rimsky et al. (2024). For each querying-phase test game we

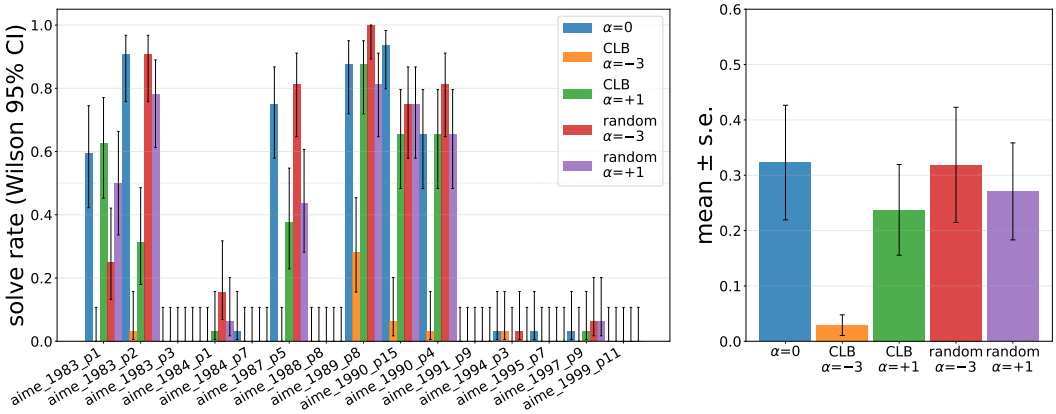

Figure 5: Per-problem (left) and aggregate (right) solve rates (pass@32) under CLB and random CAA steering applied to DSP-V2-7B. CLB suppression ($\alpha = -3$) collapses solve rate from 32.3% to 2.9%; random suppression leaves it at 31.9%.

teacher-force the model with two verbalizations over the *identical* game structure (same symbolic stimulus and speaker message): a *positive* trace produced by the correct rule-based hypothesis tracker (with the true value↔symbol binding map accumulated from sync rounds), and a *negative* trace produced by a corrupted tracker whose binding map has been deranged by a cyclic shift—the trace is surface-coherent but reasons from wrong hypotheses, producing wrong inverse predictions and ultimately a wrong listener decision (see Figure 4 for an illustration). Because the prompt context is identical across the pair it cancels in the difference. Only the CLB reasoning is contrasted between the two traces. The per-layer mean residual-stream activation over each reasoning span is computed in a matched forward pass, and $\mathbf{v}_{\text{CLB}}$ is the weighted mean of (correct − corrupted) differences over all matched pairs across episodes. We record it only over the last quarter of the model's layers (i.e. the 7 last layers of DeepSeek-Prover-V2-7B). At inference time, $\alpha \cdot \mathbf{v}_{\text{CLB}}$ is added to the residual stream at each token, where $\alpha > 0$ amplifies and $\alpha < 0$ suppresses the CLB direction.

Validation on S2B-LM. Figure 6 shows the ZSCT dosage curve as α is swept from -4 to $+1$. CLB steering produces a monotone causal response: suppression at $\alpha = -4$ reduces ZSCT from 92.6% (adj-ZSCT $= 85.2$) at baseline to 57.2% (adj-ZSCT $= 14.4$), near the 50% guessing floor, while mild amplification at $\alpha = +1$ yields 95.2% (adj-ZSCT $= 90.4$). Two control conditions—an *orthogonal* direction (a random vector projected to be orthogonal to $\mathbf{v}_{\text{CLB}}$) and a *random* direction—show flat responses across the same α range, establishing the specificity of the CLB direction. We measure *output coherence* as the fraction of generated proofs that are well-formed Lean 4—either proven correct or syntactically valid but unproven (i.e. free of parse or elaboration errors)—as opposed to outputs containing syntax errors or empty extractions. The coherence barplot (Figure 7) confirms that CLB suppression does not degrade output coherence: the fraction of well-formed outputs remains stable across all steering magnitudes, ruling out the confound that performance drops stem from garbled generations rather than from suppressed CLB competency.

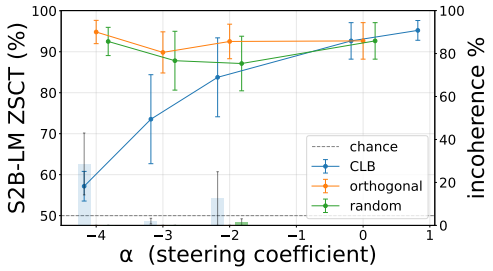

Figure 6: S2B-LM dosage curve for DSP-V2-7B under CLB CAA steering (α on the x-axis), with $n = 3$ random seeds per setting. Left y-axis: ZSCT (%); right y-axis: mean output coherence. CLB suppression ($\alpha < 0$) monotonically drives ZSCT toward the 50% guessing floor while coherence remains marginally intact, while control directions are flat across the same range.

Transfer to AIME theorem proving. Having validated that $\mathbf{v}_{\text{CLB}}$ selectively modulates CLB competency in S2B-LM, we apply the *same* CLB-encoding CAA vector during miniF2F proof generation, restricting to the 15 AIME problems in the Olympiad-level, tail subset to maximise the expected CLB signal. Figures 5 and 7 show per-problem solve rates (pass@32) and aggregate

performance under baseline (α=0), CLB suppression (α=-3) and amplification (α=$+1$), and random controls ($\alpha \in \{-3, +1\}$).

CLB suppression at $\alpha = -3$ collapses aggregate solve rate from 32.3% at baseline to 2.9%, without creating incoherent Lean proof candidates. This coherent collapse amounts to an 91% relative drop, which is spread nearly uniformly across all 15 problems. By contrast, the random control at the same magnitude ($\alpha = -3$) leaves performance virtually unchanged at 31.9%, confirming that the collapse is specific to the CLB direction and not an artefact of any residual-stream perturbation. CLB amplification at $\alpha = +1$ yields 23.8%—below baseline. We tentatively attribute this to near-saturation of the model's CLB representations at $\alpha = 0$ (ZSCT 92.6%, close to ceiling), such that further amplification introduces representational noise rather than additional benefit; however, we acknowledge that this explanation is post-hoc and that the amplification direction may simply be poorly calibrated for proof generation, leaving the asymmetry between suppression and amplification as a limitation for future work. Together, these results establish a causal role for the CLB direction in AIME-level theorem proving: selectively suppressing it is sufficient to dismantle performance, while the random perturbation of the same norm leaves performance intact.

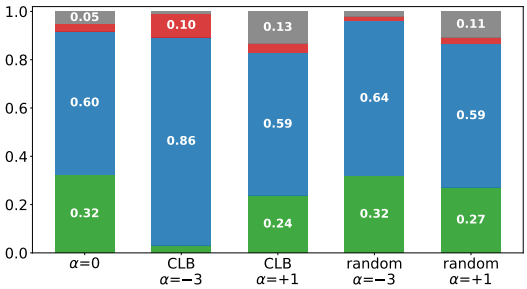

Figure 7: Mean output coherence fractions per condition under CAA steering applied to DSP-V2-7B on the AIME subset. Each bar shows the fraction of generated proofs falling into one of four classes: **solved** (correct proof), **valid Lean, unproven** (well-formed but incomplete), **invalid Lean** (parse/elaboration error), and **system/timeout** (Lean compiler timeout or system error). Coherence (solved + valid Lean) remains stable across all conditions, ruling out garbled-output confounds.

5 Discussion & Future Work

We investigated the role of Compositional Learning Behaviours (CLBs) in formal theorem proving by cross-evaluating ten state-of-the-art provers on S2B-LM—an adaptation of the S2B that substitutes a categorical stimulus domain for the continuous one and adds chain-of-thought scaffolding, controlling respectively for NUPA confounders and the competence–performance distinction. Two complementary lines of evidence converge on the same conclusion. **Correlationally**, a quadrant test dichotomising each axis at its empirical median finds the forbidden cell (high miniF2F, low adj-ZSCT) **empty** under exact enumeration of all $\binom{10}{5} = 252$ label placements ($p = 0.004$): no model attains Olympiad-level performance without measurable CLB competency, yet CLB competency does not determine ranking *within* the hard tail. Model scale is ruled out as the driving confound as *DeepSeek-Prover-V2-7B* clears the Olympiad threshold of miniF2F with only 7B parameters. **Causally**, a CLB-encoding direction extracted from the same model via Contrastive Activation Addition on matched teacher-forced S2B-LM pairs transfers to proof generation: suppression ($\alpha = -3$) collapses the AIME solve rate from 32.3% to 2.9% while preserving Lean syntactic coherence, whereas a random direction of equal norm leaves it unchanged. Together, these results make CLB competency *necessary-but-not-sufficient* for Olympiad-level formal mathematics.

We argue in Appendix A.1 that evaluating provers in S2B-LM in the listener role is a deliberate diagnostic focus motivated by the assumption that constructive synthesis is structurally downstream of receptive grounding.

Moreover, while CB capacity and e.g. proof-tree search can help provers navigate the tractable bulk of formal mathematics, we argue that the within-episode generalisation that CLBs afford has consequences for the stability of self-improvement loops, which we develop in Appendix A.2.

Future work. Two directions follow directly: (1) repurposing S2B-LM to generate synthetic training data that explicitly instils CLB competency into formal provers via supervised learning; and (2) evaluating the *constructivity* axis of CLBs by placing provers in the speaker role of S2B-LM.

AI USE STATEMENT

In this work, we used generative AI tools for (i) refining hypotheses, (ii) providing feedback on research methodology and experiments, (iii) implementing methods, and (iv) interpreting results. We have not used generative AI tools for (i) helping develop theoretical models or conceptual frameworks, (ii) supporting qualitative and thematic data analysis. Furthermore, (i) generating synthetic data sets, (ii) formulating mathematical claims, (iii) providing critical ingredients for proving mathematical claims, (iv) assisting in the writing of proofs, (v) assisting with translation, (vi) cleaning and reformatting dataset are not applicable to this work.

Additionally, we used generative AI tools for (i) creating and modifying scientific figures or images, (ii) creating and editing software code, (iii) drafting parts of a research paper, (iv) brainstorming, sourcing/searching for information, and (v) editing a research paper to improve readability. We have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness and internal validity. LLM-aided refinement of parts of the paper were performed using a strict human-in-the-loop workflow, guaranteeing that all refinement candidates are verified for correctness and accuracy as well as expanded upon before making it into the paper draft. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Rahma Chaabouni, Eugene Kharitonov, Diane Bouchacourt, Emmanuel Dupoux, and Marco Baroni. Compositionality and generalization in emergent languages. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 4427–4442, 2020.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo,

Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. Deepseek-v3 technical report, 2025. URL https://arxiv.org/abs/2412.19437.

Kevin Denamganaï, Sondess Missaoui, and James Alfred Walker. Meta-referential games to learn compositional learning behaviours. *arXiv preprint arXiv:2207.08012*, 2022.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, et al. A survey on in-context learning, 2024. *URL https://arxiv.org/abs/2301.00234*, 19, 2022.

Jan Dul. Necessary condition analysis (nca) logic and methodology of "necessary but not sufficient" causality. *Organizational research methods*, 19(1):10–52, 2016.

Chaz Firestone. Performance vs. competence in human–machine comparisons. *Proceedings of the National Academy of Sciences*, 117(43):26562–26571, 2020.

Phillip Good. *Permutation, parametric and bootstrap tests of hypotheses*. Springer, 2005.

Najoung Kim and Tal Linzen. COGS: A compositional generalization challenge based on semantic interpretation. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9087–9105, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.731. URL https://aclanthology.org/2020.emnlp-main.731/.

Andrew Lampinen. Can language models handle recursively nested grammatical structures? a case study on comparing models and humans. *Computational Linguistics*, 50(4):1441–1476, December 2024. doi: 10.1162/coli_a_00525. URL https://aclanthology.org/2024.cl-4.8/.

Yong Lin, Shange Tang, Bohan Lyu, Jiayun Wu, Hongzhou Lin, Kaiyu Yang, Jia LI, Mengzhou Xia, Danqi Chen, Sanjeev Arora, and Chi Jin. Goedel-prover: A frontier model for open-source automated theorem proving. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=x2y9i2HDjD.

Yong Lin, Shange Tang, Bohan Lyu, Ziran Yang, Jui-Hui Chung, Haoyu Zhao, Lai Jiang, Yihan Geng, Jiawei Ge, Jingruo Sun, Jiayun Wu, Jiri Gesi, Ximing Lu, David Acuna, Kaiyu Yang, Hongzhou Lin, Yejin Choi, Danqi Chen, Sanjeev Arora, and Chi Jin. Goedel-prover-v2: Scaling formal theorem proving with scaffolded data synthesis and self-correction. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=j4C0nALrgK.

Fortunato Pesarin and Luigi Salmaso. *Permutation tests for complex data: theory, applications and software*. John Wiley & Sons, 2010.

Edwin James George Pitman. Significance tests which may be applied to samples from any populations. ii. the correlation coefficient test. *Supplement to the Journal of the Royal Statistical Society*, 4(2):225–232, 1937.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Z. Z. Ren, Zhihong Shao, Junxiao Song, Huajian Xin, Haocheng Wang, Wanjia Zhao, Liyue Zhang, Zhe Fu, Qihao Zhu, Dejian Yang, Z. F. Wu, Zhibin Gou, Shirong Ma, Hongxuan Tang, Yuxuan Liu, Wenjun Gao, Daya Guo, and Chong Ruan. Deepseek-prover-v2: Advancing formal mathematical reasoning via reinforcement learning for subgoal decomposition, 2025. URL https://arxiv.org/abs/2504.21801.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.828. URL https://aclanthology.org/2024.acl-long.828/.

Adam Santoro, Andrew Lampinen, Kory Mathewson, Timothy Lillicrap, and David Raposo. Symbolic behaviour in artificial intelligence. *arXiv preprint arXiv:2102.03406*, 2021.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300, 2024. URL https://doi.org/10.48550/arXiv.2402.03300.

George Tsoukalas, Jasper Lee, John Jennings, Jimmy Xin, Michelle Ding, Michael Jennings, Amitayush Thakur, and Swarat Chaudhuri. Putnambench: Evaluating neural theorem-provers on the putnam mathematical competition. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 11545–11569. Curran Associates, Inc., 2024. doi: 10.52202/079017-0368. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/1582eaf9e0cf349e1e5a6ee453100aa1-Paper-Datasets_and_Benchmarks_Track.pdf.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.

Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272, 2020.

Quang H Vuong. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica: journal of the Econometric Society*, pp. 307–333, 1989.

Haiming Wang, Mert Unsal, Xiaohan Lin, Mantas Baksys, Junqi Liu, Marco Dos Santos, Flood Sung, et al. Kimina-Prover: Applying test-time RL search on large formal reasoning models. https://huggingface.co/blog/AI-MO/kimina-prover, 2025a.

Haiming Wang, Mert Unsal, Xiaohan Lin, Mantas Baksys, Junqi Liu, Marco Dos Santos, Flood Sung, Marina Vinyes, Zhenzhe Ying, Zekai Zhu, Jianqiao Lu, Hugues de Saxcé, Bolton Bailey, Chendong Song, Chenjun Xiao, Dehao Zhang, Ebony Zhang, Frederick Pu, Han Zhu, Jiawei Liu, Jonas Bayer, Julien Michel, Longhui Yu, Léo Dreyfus-Schmidt, Lewis Tunstall, Luigi Pagani, Moreira Machado, Pauline Bourigault, Ran Wang, Stanislas Polu, Thibaut Barroyer, Wen-Ding Li, Yazhe Niu, Yann Fleureau, Yangyang Hu, Zhouliang Yu, Zihan Wang, Zhilin Yang, Zhengying Liu, and Jia Li. Kimina-prover preview: Towards large formal reasoning models with reinforcement learning, 2025b. URL https://arxiv.org/abs/2504.11354.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022a. ISSN 2835-8856. URL https://openreview.net/forum?id=yzkSU5zdwD. Survey Certification.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022b. URL https://openreview.net/forum?id=_VjQlMeSB_J.

Huajian Xin, Daya Guo, Zhihong Shao, Zhizhou Ren, Qihao Zhu, Bo Liu, Chong Ruan, Wenda Li, and Xiaodan Liang. Deepseek-prover: Advancing theorem proving in llms through large-scale synthetic data. *arXiv preprint arXiv:2405.14333*, 2024.

Huajian Xin, Z.Z. Ren, Junxiao Song, Zhihong Shao, Wanjia Zhao, Haocheng Wang, Bo Liu, Liyue Zhang, Xuan Lu, Qiushi Du, Wenjun Gao, Haowei Zhang, Qihao Zhu, Dejian Yang, Zhibin Gou, Z.F. Wu, Fuli Luo, and Chong Ruan. Deepseek-prover-v1.5: Harnessing proof assistant feedback for reinforcement learning and monte-carlo tree search. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=I4YAIwrsXa.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025a. URL https://arxiv.org/abs/2505.09388.

Haotong Yang, Yi Hu, Shijia Kang, Zhouchen Lin, and Muhan Zhang. Number cookbook: Number understanding of language models and how to improve it. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=BWS5gVjgeY.

Kaiyu Yang, Aidan M Swope, Alex Gu, Rahul Chalamala, Peiyang Song, Shixing Yu, Saad Godil, Ryan Prenger, and Anima Anandkumar. Leandojo: Theorem proving with retrieval-augmented language models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=g7OX2sOJtn.

Kaiyu Yang, Gabriel Poesia, Jingxuan He, Wenda Li, Kristin E. Lauter, Swarat Chaudhuri, and Dawn Song. Position: Formal mathematical reasoning—a new frontier in AI. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025c. URL https://openreview.net/forum?id=HuvAM5x2xG.

Kunhao Zheng, Jesse Michael Han, and Stanislas Polu. minif2f: a cross-system benchmark for formal olympiad-level mathematics. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=9ZPegFuFTFv.

A Further Discussions

A.1 On the Asymmetry of Receptivity vs Constructivity in CLBs: Receptivity is a Structural Gatekeeper

A potential critique is that a linguistic listener-only benchmark like S2B-LM introduces a task-modality mismatch when evaluating formal theorem provers, which are inherently constructive sequence generators. However, this argument overlooks the structural asymmetry between the two aspects of CLBs established in Section 2.1. In formal proof synthesis, constructive aspects are structurally downstream of receptive ones. Even in the whole-proof generation regime, a prover does not synthesise proof steps from a fixed prelearned vocabulary alone: Olympiad-level theorems require the chain-of-thought to decode, bind, and recombine novel intermediate propositions, unfamiliar tactic compositions, and out-of-distribution mathematical objects that cannot be retrieved from pretraining weights by memorisation—the same within-episode grounding requirement that defines the CLB regime (Section 2.1).

If an agent lacks the capacity to systematically process novel, out-of-distribution symbolic structures encountered in context (receptivity), then any subsequent constructive generation degrades. Confronted with mathematical configurations beyond its compositional horizon, the agent with low receptive CLB competency cannot decompose and recombine the relevant symbolic structures systematically, causing its constructive search to collapse into stochastic matching over its prelearned tactic distribution.

Thus, isolating the receptive axis in S2B-LM is a deliberate diagnostic focus. Receptive systematicity serves as a strict structural gatekeeper. While low-to-medium difficulty theorems of miniF2F (Zheng et al. (2022); $164/244$ MATH-level problems, i.e. miniF2F $\leq 67.2\%$) can be proved purely without measurable CLB competency, breaking into the Olympiad-level tail (miniF2F $> 67.2\%$) requires clearing this foundational bottleneck.

A.2 On Self-Evolving Agents requiring CLBs.

The stability of a self-evolving scientific agent's improvement loop depends critically on CLB capacity. While CB capacity and proof-tree search suffice for the tractable bulk of formal mathematics, a genuinely self-evolving agent must progressively ground proof structures outside its fixed prelearned vocabulary. Without robust CLB competency, an agent cannot isolate abstract principles from successful proof trajectories. Indeed, its constructive competency would collapse into stochastic matching over prelearned tactics. Systematic receptivity thus stabilises the exploration space, enabling the policy to ground newly-encountered symbolic objects and recombine them with prelearned primitives as difficulty increases.

B RULE-BASED LISTENER VERBALIZER ALGORITHM

Algorithm 1 Hypothesis-tracking rule-based listener verbalizer (one episode)

Require: Posdis-speaker π^S, N_{dim}, supporting-phase games G_{sup}

$\mathcal{V} \leftarrow \{\}$ {value map: $(\text{pos}, \text{tok}) \rightarrow \{\text{val} : \text{count}\}$}
$\mathbf{m}_{\text{last}} \leftarrow \{\}$
for game $g = 1, \ldots, G_{\text{sup}}$ **do**
  Receive speaker message $\mathbf{m} = [t_0, \ldots, t_{N_{\text{dim}}-1}]$ and own stimulus $\mathbf{s} = [s_0, \ldots, s_{N_{\text{dim}}-1}]$
  // Sync round: update $\mathcal{V}$ from speaker's target and message m_{last} at previous game
  if $m_{\text{last}} \neq \{\}$ **then**
    Receive speaker target at previous game, from listener feedback/sync step $\mathbf{f}_{\text{last}} = [f_0^{\text{last}}, \ldots, f_{N_{\text{dim}}-1}^{\text{last}}]$
    for $i = 0, \ldots, N_{\text{dim}} - 1$ **do**
      if $t_i^{\text{last}} \neq$ EOS **then**
        $\mathcal{V}[(i,\, t_i^{\text{last}})][f_i^{\text{last}}] \mathrel{+}= 1$
      end if
    end for
  end if
  // Bookkeeping: records the current game's speaker message for the next sync round
  $\mathbf{m}_{\text{last}} = [t_0^{\text{last}}, \ldots, t_{N_{\text{dim}}-1}^{\text{last}}] \leftarrow \mathbf{m}$
  // Invert $\mathcal{V}$: predict the message the speaker would send for $\mathbf{s}$
  for $i = 0, \ldots, N_{\text{dim}} - 1$ **do**
    if $(i,\, s_i)$ has evidence in $\mathcal{V}$ (as a value) **then**
      $\hat{t}_i \leftarrow \arg\max_t \; \mathcal{V}[(i,\, t)][s_i]$
    else
      $\hat{t}_i \leftarrow$ UNKNOWN
    end if
  end for
  // Decision: count matches between actual and predicted message
  $n_{\text{match}} \leftarrow |\{i : \hat{t}_i = t_i,\ \hat{t}_i \neq \text{UNKNOWN}\}|$
  $d \leftarrow 0$ (same class) if $n_{\text{match}} \geq N_{\text{dim}}$, else 1 (different class)
  EMIT(VERBALIZE($\mathcal{V}$, $\mathbf{s}$, $\mathbf{m}$, $\hat{\mathbf{t}}$, n_{match}, d))
end for

C CONVERSATION EXAMPLE

S2B-LM supporting-phase conversation: 6 games, $N_{\text{dim}} = 3$ (*instruments, sports, vegetables*), vocab size 16, max sentence length 3

[System]
You are a listener agent in a referential game played over multiple rounds. Use the conversation history to learn the speaker's private code across games, then decide whether the current stimulus shares the same latent meaning. Think step by step, then end your response with **Answer: 0** (same latent meaning) or **Answer: 1** (different).
You and your partner play a sequence of referential games; you are the listener. In the first phase you will get acquainted with the atomic components of the possible observations; then you will be tested with new observations combining the same atomic components in novel ways. At each game each of you observes a stimulus representing a latent meaning, and your common goal is to figure out whether you are observing different or similar latent meanings. Your partner can send you messages via a communication channel of 16 symbols combined into sentences of maximum length 3. Symbol 0 is the end-of-message symbol: any symbol following it is ignored and regularised to 0.

[User — Game #0]
At game #0, you are observing stimulus: `[['piano', 'swimming', 'eggplant']]`. Your partner has sent you the following message: `[8, 5, 6]`.
Question #1: At game #0, do you think you are observing a stimulus with the same latent meaning as your partner? Answer 0: Yes or 1: No. End your response with your decision as a single integer.

[Listener — Game #0]
Let's think step by step and leverage past games. No sync step data yet — cannot predict expected symbols. **Answer: 0**

[User — Sync & Game #1]
At the end of game #0, sync step: the exact stimulus your partner observed was `[['piano', 'swimming', 'eggplant']]`. You decided: *similar* latent meanings. This was **correct** — you have won game #0.
At game #1, you are observing stimulus: `[['piano', 'golf', 'pepper']]`. Your partner has sent you the following message: `[8, 11, 13]`.
Question #1: At game #1, do you think you are observing a stimulus with the same latent meaning as your partner? Answer 0: Yes or 1: No. End your response with your decision as a single integer.

[Listener — Game #1 (first full verbalized trace)]
Let's think step by step and leverage past games. From the last game syncing, we can learn that: symbol 8 at pos 0 → *piano* ; symbol 5 at pos 1 → *swimming* ; symbol 6 at pos 2 → *eggplant*. In the current game, if the speaker were observing a similar stimulus as ours, [*piano*, *golf*, *pepper*], then: at pos 0, *piano* → symbol 8 (from game #0) ; at pos 1, *golf* has not been observed yet ; at pos 2, *pepper* has not been observed yet. Since the speaker's message is `[8, 11, 13]`, yield 1/3 matches, they are likely observing a *different* stimulus. **Answer: 1**

[User — Sync & Game #2]
At the end of game #1, sync step: the exact stimulus your partner observed was `[['piano', 'golf', 'pepper']]`. You decided: *different* latent meanings. This was **incorrect** — you have lost game #1.
At game #2, you are observing stimulus: `[['oboe', 'rugby', 'broccoli']]`. Your partner has sent you the following message: `[3, 9, 2]`.
Question #1: At game #2, do you think you are observing a stimulus with the same latent meaning as your partner? Answer 0: Yes or 1: No. End your response with your decision as a single integer.

[Listener — Game #2]
Let's think step by step and leverage past games. From the last game syncing, we can learn that: symbol 8 at pos 0 → *piano* ; symbol 11 at pos 1 → *golf* ; symbol 13 at pos 2 → *pepper*. In the current game, if the speaker were observing a similar stimulus as ours, [*oboe*, *rugby*, *broccoli*], then: at pos 0, *oboe* has not been observed yet ; at pos 1, *rugby* has not been observed yet ; at pos 2, *broccoli* has not been observed yet. Since the speaker's message is `[3, 9, 2]`, yield 0/3 matches, they are likely observing a *different* stimulus. **Answer: 1**

[User — Sync & Game #3]
At the end of game #2, sync step: the exact stimulus your partner observed was `[['oboe', 'rugby', 'broccoli']]`. You decided: *different* latent meanings. This was **incorrect** — you have lost game #2.
At game #3, you are observing stimulus: `[['oboe', 'rugby', 'broccoli']]`. Your partner has sent you the following message: `[3, 15, 6]`.
Question #1: At game #3, do you think you are observing a stimulus with the same latent meaning as your partner? Answer 0: Yes or 1: No. End your response with your decision as a single integer.

[Listener — Game #3]
Let's think step by step and leverage past games. From the last game syncing, we can learn that: symbol 3 at pos 0 → *oboe* ; symbol 9 at pos 1 → *rugby* ; symbol 2 at pos 2 → *broccoli*. In the current game, if the speaker were observing a similar stimulus as ours, [*oboe*, *rugby*, *broccoli*], then: at pos 0, *oboe* → symbol 3 (from game #2) ; at pos 1, *rugby* → symbol 9 (from game #2) ; at pos 2, *broccoli* → symbol 2 (from game #2). Since the speaker's message is `[3, 15, 6]`, yield 1/3 matches, they are likely observing a *different* stimulus. **Answer: 1**

[User — Sync & Game #4]
At the end of game #3, sync step: the exact stimulus your partner observed was `[['oboe', 'skiing', 'eggplant']]`. You decided: *different* latent meanings. This was **correct** — you have won game #3.
At game #4, you are observing stimulus: `[['drums', 'golf', 'pepper']]`. Your partner has sent you the following message: `[12, 11, 13]`.
Question #1: At game #4, do you think you are observing a stimulus with the same latent meaning as your partner? Answer 0: Yes or 1: No. End your response with your decision as a single integer.

[Listener — Game #4]
Let's think step by step and leverage past games. From the last game syncing, we can learn that: symbol 3 at pos 0 → *oboe* ; symbol 15 at pos 1 → *skiing* ; symbol 6 at pos 2 → *eggplant*. In the current game, if the speaker were observing a similar stimulus as ours, [*drums*, *golf*, *pepper*], then: at pos 0, *drums* has not been observed yet ; at pos 1, *golf* → symbol 11 (from game #1) ; at pos 2, *pepper* → symbol 13 (from game #1). Since the speaker's message is `[12, 11, 13]`, yield 2/3 matches, they are likely observing a *different* stimulus.
Answer: 1

[User — Sync & Game #5]
At the end of game #4, sync step: the exact stimulus your partner observed was `[['drums', 'golf', 'pepper']]`. You decided: *different* latent meanings. This was **incorrect** — you have lost game #4.
At game #5, you are observing stimulus: `[['guitar', 'rugby', 'eggplant']]`. Your partner has sent you the following message: `[4, 9, 6]`.
Question #1: At game #5, do you think you are observing a stimulus with the same latent meaning as your partner? Answer 0: Yes or 1: No. End your response with your decision as a single integer.

[Listener — Game #5]
Let's think step by step and leverage past games. From the last game syncing, we can learn that: symbol 12 at pos 0 → *drums* ; symbol 11 at pos 1 → *golf* ; symbol 13 at pos 2 → *pepper*. In the current game, if the speaker were observing a similar stimulus as ours, [*guitar*, *rugby*, *eggplant*], then: at pos 0, *guitar* has not been observed yet ; at pos 1, *rugby* → symbol 9 (from game #2) ; at pos 2, *eggplant* → symbol 6 (from game #3). Since the speaker's message is `[4, 9, 6]`, yield 2/3 matches, they are likely observing a *different* stimulus.
Answer: 1

Figure 8: S2B-LM conversation examples over 6 supporting-phase games illustrating the hypothesis-tracking rule-based listener verbalizer. **System** sets the task context. **User** messages present each game's stimulus and message, followed by the sync-step revelation of the speaker's target. **Listener** responses are produced by VERBALIZE (Algorithm 1): game #0 defaults to *same* (no sync data yet); subsequent games accumulate sync-derived token-to-value evidence in $\mathcal{V}$ and apply the sync-summary → inverse-prediction → match-comparison chain; games #3–#5 illustrate how the evidence base grows across episodes, with partial matches (1/3, 2/3) correctly classified as *different* even as the mapping fills in.

D FURTHER CORRELATIONAL EVIDENCE

We cross-evaluate the ten provers of Table 2 against two complementary non-parametric tests to determine whether CLB competency—as measured by adj-ZSCT—constitutes a necessary prerequisite for high miniF2F performance. We begin with a global ceiling-zone analysis over all $10! = 3{,}628{,}800$ pairings (Section D.1). Seeing that this global test falls just short of conventional significance ($p = 0.052$), we consider the likely confound introduced by the heavily right-skewed difficulty distribution of miniF2F (Zheng et al., 2022): search-tractable theorems in the lower range can be solved without any measurable CLB competency, diluting any global boundary signal. We therefore apply a structurally targeted tail partition test (Section D.2) that isolates the Olympiad-level regime (miniF2F $> 67.2\%$, which is the structural undergraduate-Olympiad boundary of the benchmark) and asks whether the models that penetrate it cluster exclusively at high adj-ZSCT. To rule out model scale as a competing explanation, we embed both predictors in an identical permutation space and compare their significance footprints (Section D.3). Throughout, we rely exclusively on exact permutation methods rather than parametric or asymptotic estimators, as the small evaluation pool ($N = 10$) renders standard asymptotic tools structurally invalid; the full statistical rationale is given in Appendix E.

D.1 GLOBAL CONTINUOUS BOTTLENECK TEST

We first apply a global, continuous bottleneck check using non-parametric ceiling zone analysis (Dul, 2016). This framework operationalises the prerequisite hypothesis geometrically: under a true bottleneck, the upper-left region of the (adj-ZSCT, miniF2F) scatter—high miniF2F, low adj-ZSCT—should be systematically vacant.

Limitation: slope dependency. The vacancy penalty statistic requires a boundary line $Y = s \cdot X$ that separates the forbidden upper-left region from the permitted lower-right region. Because adj-ZSCT and miniF2F are expressed on the same $[0, 100]$ percentage scale but measure incommensurate quantities, the slope s cannot be derived from first principles — it is a free parameter. We set it in a data-determined, scale-free way as $s = \max(\text{miniF2F}) / \max(\text{adj-ZSCT}) = 88.1/100.0 = 0.881$, which normalises each axis by its observed maximum so that the boundary passes through the top-right corner of the empirical scatter. This is the most principled choice available, but the p-value remains sensitive to it (e.g. $p = 0.050$ at $s = 0.88$, rising to $p = 0.054$ at the bare diagonal $s = 1.0$). The slope-free quadrant test in Section 4.1 avoids this limitation entirely.

Null Hypothesis ($\mathcal{H}_0^{\text{CLB,global}}$). The joint distribution of (adj-ZSCT, miniF2F) pairs exhibits no directional vacancy in the upper-left quadrant; the observed triangular layout is consistent with a random re-pairing of the empirical adj-ZSCT and miniF2F marginal distributions.

Alternative Hypothesis ($\mathcal{H}_1^{\text{CLB,global}}$). The upper-left quadrant of the (adj-ZSCT, miniF2F) space is systematically empty, indicating that low CLB competency constrains the maximum achievable formal verification accuracy.

We define the test statistic T as the negative upper-left vacancy penalty:

$$T = -\sum_{i=1}^{N} \max\left(0,\ \frac{\text{miniF2F}_i}{100} - s \cdot \frac{\text{adj-ZSCT}_i}{100}\right) \tag{2}$$

$T = 0$ indicates no violations of the prerequisite ordering; increasingly negative values reflect models that achieve high miniF2F despite low adj-ZSCT. Under $\mathcal{H}_1^{\text{CLB,global}}$, the observed pairing should yield a T value significantly closer to zero than under random re-pairings.

An exact pairings permutation test evaluating all $10! = 3{,}628{,}800$ possible (adj-ZSCT, miniF2F) pairings, using the scale-normalised slope $s = 0.881$, yields an observed statistic of $T = -3.237$ and an empirical $p_{\text{CLB}}^{\text{global}} = \mathbf{0.050}$. This sits exactly at the conventional significance threshold ($\alpha = 0.05$) and should be interpreted with caution given its sensitivity to the slope choice noted above. The primary driver of the residual penalty is *Goedel-Prover-V2-32B*, which achieves miniF2F $= 88.1\%$ at a comparatively low adj-ZSCT $= 30.4$, placing a data point in the upper-left region and inflating the observed vacancy penalty. We attribute this anomaly to the competence–performance distinction (Section 3.2): the model may possess latent CLB competency that the 10-shot scaffold fails to elicit at this scale and model family, causing adj-ZSCT to underestimate its true CLB ability.

D.2 TAIL PARTITION PERMUTATION TEST

The failure of the global continuous test motivates a more targeted, structurally motivated hypothesis. The difficulty distribution of miniF2F is heavily right-skewed (Zheng et al., 2022): 164 of the

244 test-split problems (67.2%) are drawn from MATH levels 1–5 and custom sources, forming a low-to-medium difficulty bulk solvable via syntactic fluency and extensive proof-tree search without requiring systematic compositional generalisation, while the remaining 80 problems (32.8%) are drawn from AMC, AIME, and IMO—the competitive and Olympiad tiers that resist brute-force enumeration and should expose a genuine CLB bottleneck. Any global boundary test will therefore be confounded by the abundance of search-solvable theorems in the lower miniF2F range.

We formalise this observation as a *tail partition test* anchored at the structural undergraduate–Olympiad boundary of 67.2%: 164 of 244 test-split problems (67.2%) are MATH-level bulk; the remaining 80 (32.8%) are AMC/AIME/IMO Olympiad problems. The empirical median of miniF2F across our ten models (67.95%) corroborates this structural cut. The five models that surpass this threshold—*Goedel-Prover-V2-8B* (84.6%), *Goedel-Prover-V2-32B* (88.1%), *DeepSeek-Prover-V2-7B* (75.6%), *DeepSeek-Prover-V2-671B* (82.4%), and *Kimina-Prover* (84.0%)—constitute the *hard tail* partition $\mathcal{T}$.

Null Hypothesis (H_0^{tail}). The five models with the highest miniF2F scores are drawn uniformly at random from the full pool with respect to their adj-ZSCT scores; any apparent clustering at high adj-ZSCT is attributable to chance.

Alternative Hypothesis (H_1^{tail}). The five models that penetrate the Olympiad-level tail exhibit systematically higher CLB competency (adj-ZSCT) than would be expected under a random partition of the pool.

The test statistic is the aggregate adj-ZSCT sum over the hard tail:

$$S_{\mathcal{T}} = \sum_{i \in \mathcal{T}} \text{adj-ZSCT}_i \tag{3}$$

We enumerate all $\binom{10}{5} = 252$ possible ways to select five models from the pool and compute $S_{\mathcal{T}}$ for each partition. The observed value is $S_{\mathcal{T}} = 354.80$. Strikingly, this value corresponds to the *maximum achievable* adj-ZSCT sum across all 252 partitions: the five models in $\mathcal{T}$ are identically the five models with the highest adj-ZSCT scores in the full pool. Exactly **1** configuration out of 252 meets or exceeds this value, yielding an exact empirical p-value of $p_{\text{CLB}}^{\text{tail}} = \frac{1}{252} = \mathbf{0.00397}$. This result is highly statistically significant ($p < 0.01$), supporting H_1^{tail}. The set of models that conquer the Olympiad-level tail of formal mathematics is in perfect correspondence with the set of models that exhibit the strongest CLB competency—a coincidence with probability less than 0.4% under the null. Crucially, this finding is consistent with CLB competency being a *necessary* condition for elite formal verification performance: no model outside the top adj-ZSCT tier enters the hard tail, regardless of architecture scale or search budget.

Takeaway 1. CLB competency is not required for the tractable bulk ($p_{\text{CLB}}^{\text{global}} = 0.052$, non-significant) but is a strict structural prerequisite for the Olympiad-level tier: the top-5 models by miniF2F are precisely the top-5 by adj-ZSCT across all $\binom{10}{5} = 252$ partitions ($p_{\text{CLB}}^{\text{tail}} = 0.004$).

D.3 Decoupling CLB Competency from Model Scale

A compelling competing explanation for the observed capability dissociation is the *scale confound hypothesis* ($\mathcal{H}_{\text{scale}}$): larger models may dominate both the CLB diagnostic and miniF2F accuracy simultaneously, with raw parameter count—rather than systematic compositional generalisation—as the true latent predictor. To formally adjudicate between $\mathcal{H}_{\text{CLB}}$ and $\mathcal{H}_{\text{scale}}$, we deploy a competitive hypothesis testing tournament. This is motivated by the non-nested model selection logic of Vuong (1989) but operating entirely within an exact permutation framework (Pitman, 1937; Pesarin & Salmaso, 2010).

Choice of Test Statistic: Pearson r Under Permutation. We adopt Pearson's product-moment correlation coefficient r as the test statistic for the continuous boundary test. A common misconception holds that using r forces bivariate normality assumptions; this is only true when p-values are derived from asymptotic Student's t-distributions. When embedded in an exact permutation framework, r operates as a purely deterministic summary of directional coupling, with no distributional preconditions (Pitman, 1937). Relative to rank-based alternatives such as Spearman's ϱ, Pearson r better captures proportional metric gradients: a model that breaks a structural bottleneck and scales sharply on the miniF2F axis in proportion to its CLB gain produces a stronger r signal. Critically, by shifting the construction of the null distribution entirely to exact permutation enumeration, we bypass the standard requirements of homoscedasticity and normality that would otherwise render our small-sample analysis invalid (Pesarin & Salmaso, 2010).

Continuous Space Permutation Distributions. All analyses use $N = 10$ consistently. Under $\mathcal{H}_0^{\text{continuous}}$, miniF2F performance is structurally independent of the predictor; any pairing of a model's predictor value with another model's miniF2F score is equally likely. We fix the miniF2F vector and compute r across all $10! = 3{,}628{,}800$ permutations of each predictor, forming an exact empirical null distribution:

$$p = \frac{1}{N!} \sum_{i=1}^{N!} \mathbb{I}(r_{\pi_i} \geq r_{\text{obs}}) \tag{4}$$

For $\mathcal{H}_{\text{CLB}}$ (adj-ZSCT), the observed correlation $r = 0.7652$ yields an exact p-value of $\mathbf{p}_{\text{CLB}}^{\text{continuous}} = \mathbf{0.00692}$—highly significant. For $\mathcal{H}_{\text{scale}}$ (total parameters in billions), the observed correlation $r = 0.3597$ yields $p_{\text{scale}}^{\text{continuous}} = 0.17837$, providing no evidence of a continuous dependency between parameter count and miniF2F accuracy.

Scale Tail Partition Test (N=10, consistent). The most direct rebuttal to the scale hypothesis is *DeepSeek-Prover-V2-7B*, which achieves 75.6% miniF2F on only 7B parameters—placing it in the hard tail while remaining the smallest model in the suite. This single counterexample directly falsifies scale as a *necessary* condition for Olympiad-level performance. We corroborate this with an exact tail partition test using total parameter count as the predictor ($N = 10$ consistently). The five hard-tail models yield an aggregate parameter sum of 790B. Enumerating all $\binom{10}{5} = 252$ partitions, 6 meet or exceed this value ($p_{\text{scale}}^{\text{tail}} = 0.02381$). By contrast, the CLB tail test yields exactly 1 out of 252 partitions ($p_{\text{CLB}}^{\text{tail}} = 0.00397$).

Non-Parametric Competitive Verdict. Under the exact permutation framework (Pitman, 1937; Pesarin & Salmaso, 2010), CLB competency achieves significance on both the continuous ($p_{\text{CLB}}^{\text{continuous}} = 0.0069$) and partition ($p_{\text{CLB}}^{\text{tail}} = 0.0040$) axes. Scale establishes no continuous footprint ($p_{\text{scale}}^{\text{continuous}} = 0.178$) and, while marginally significant in the tail partition ($p_{\text{scale}}^{\text{tail}} = 0.024$), is directly refuted as a necessary condition by the DP-V2-7B counterexample. Systematic compositional generalisation is thus the dominant structural predictor of elite formal verification performance.

Takeaway 2. CLB competency achieves significance on both the continuous ($p_{\text{CLB}}^{\text{continuous}} = 0.0069$) and partition ($p_{\text{CLB}}^{\text{tail}} = 0.0040$) axes. Raw parameter count achieves no continuous signal ($p_{\text{scale}}^{\text{continuous}} = 0.178$) and is directly refuted as a necessary condition by DP-V2-7B (7B params, in the hard tail), ruling out scale as the driving confound.

E Methodological Limitations

Evaluating frontier foundational models introduces severe sample constraints: our pool ($N = 10$) is strictly bounded by the immense compute requirements of training and running state-of-the-art provers. Operating within a small-sample regime limits the use of traditional parametric or asymptotic statistical tools, rendering standard regression or continuous correlation metrics structurally invalid.

Inadequacy of Parametric and Asymptotic Estimators. Standard statistical methodologies depend on asymptotic behaviours, where normality and uniform variance are assumed as sample sizes approach infinity ($N \to \infty$). Applying these to an $N = 10$ evaluation space introduces critical vulnerabilities:

- **Parametric Inferences:** Testing linear regression models via Student's t or F-statistics requires normally distributed residuals. In a 10-point dataset, testing for normality using standard diagnostics lacks sufficient statistical power, making significance claims highly sensitive to subtle outliers.
- **Quantile Regression:** While boundary limits are traditionally evaluated via high-quantile curves (e.g., tracking the 90th percentile), these estimators require dense data along the margins of the distribution. At $N = 10$, estimating high-quantile paths lacks the necessary degrees of freedom, creating unstable, overfit boundaries that track individual peripheral data points.
- **Heteroscedastic Distortions:** Prerequisite capability boundaries inherently display non-uniform variance (heteroscedasticity). Low input capabilities restrict output variance tightly near a baseline floor, while high input scores unlock the full vertical range of execution. Standard Ordinary Least Squares (OLS) models underperform under these conditions, distorting standard errors and inflating the risk of false-positive significance claims.

The Rigor of Exact Permutation Tools. To circumvent these small-sample limitations, our pipeline uses exact permutation tests processed via verified open-source scientific computing engines Good (2005); Virtanen et al. (2020). Permutation methods offer an elegant, distribution-free framework that provides exact mathematical precision on small sample pools:

1. **Zero Distributional Dependencies:** Permutation techniques operate without making structural assumptions regarding the normality or variance profiles of the underlying data points. They bypass population parameter estimation by calculating significance directly from the empirical dataset's internal combinatorial structure.
2. **Deterministic Execution Limits:** When total possible dataset configurations remain low, numerical engines can directly process the entire factorial set of allocations ($10! = 3{,}628{,}800$ for the global pairing test, and $\binom{10}{5} = 252$ for the tail partition). By calculating outcomes across the entire permutation space, the resulting p-value represents an exact combinatorial probability under the null hypothesis rather than an asymptotic approximation.

By shifting from continuous parametric approximations to exact partition structures, we ensure that our confirmation of the tail-dependent bottleneck remains robust, reproducible, and completely free from small-sample estimation bias.

On the scaling of Few-shot CoT prompting. In our experiments, few-shot CoT prompting with $N = 10$ exemplars is shown sufficient in eliciting the CLB competences of a wide range of LMs, from 7B to 671B. However, in-context learning capabilities, and with it the efficiency of few-shot CoT prompting, are known to scale with LM size (Brown et al., 2020; Wei et al., 2022a; Dong et al., 2022) and to depend on model family. Thus, it is not impossible that $N = 10$ exemplars may simply not be enough for some of our tested models to perform as well as their competency allows. However, we argue that this is fairly unlikely given the fact that the competency has been successfully elicited in LMs of size 7B, 8B, 32B, 72B, and 671B, spanning the whole range of tested model sizes and accounting for most of the different base model families represented in our evaluation (see Table 2).