AN EXTREME-VALUE PERSPECTIVE ON REASONING FAILURE IN TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study extreme residual-stream changes as markers of reasoning failure in transformers. Under explicit assumptions on tail behavior, dependence, and coupling to logical validity, we derive conditional localization and overshoot laws, characterize a Kesten-type mechanism generating heavy tails, and analyze tail-index estimation accounting for within-trajectory dependence. On eligible failed Lean 4 proofs generated by DeepSeek-Prover-V2-7B, Goedel-Prover-V2-8B, and Kimina-Prover-Distill-8B, the largest whitened residual-stream increment identifies the first rejected step in 17.3–25.4% of cases on miniF2F and, for DeepSeek and Goedel, in 20.4–29.4% on ProofNet-Verified, compared with length-adjusted uniform baselines of 8.7–9.7% and 13.3–14.2%, respectively. In controlled deduction experiments with Qwen2.5-7B-Instruct, an exponential model of success probability as a function of reasoning length, fitted on calibration tasks, reduces the held-out Brier score by 9.2% relative to a constant-probability baseline at temperature 1.0.

1 INTRODUCTION

A formal prover emits a complete proof before Lean checks it, and for a failed proof semantic replay identifies the first tactic that Lean rejects. This gives exact ground truth for a question that matters wherever no step-level verifier exists: *does the model's own residual stream register the step at which its reasoning first goes wrong, and does it add information beyond position and token uncertainty?*

At a fixed decoder block, we read the residual state at each formal-step boundary and measure consecutive increments. Extreme-value theory supplies a conditional framework for this statistic. If the first rejection is coupled to crossing a fixed geometric tolerance, failure becomes an event about the maximum of the increments, and tail behavior and dependence govern where failures occur, how large the failing step is, and how reliability decays with length. Under coupling and extremal independence the framework predicts that the largest increment *is* the first failure (Theorem 3). A scalar stochastic recurrence shows how occasional expansion can produce heavy-tailed deviations, as Kesten recursions do for SGD iterates (Gürbüzbalaban et al., 2021).

The data show which of these conditions trained provers satisfy. Across three provers, two benchmarks and two sampling temperatures, the largest whitened increment points to the first rejection about twice as often as a length-adjusted uniform choice, and among the steps up to the rejection the rejected step is the largest in 35–43% of proofs. A fixed level of the increment, by contrast, does not identify failure: exceedances are more frequent in failed than in verified proofs of the same tasks, yet the first exceedance coincides with the rejected step in up to a third of threshold-exceeding failed proofs, thresholds calibrated on one benchmark shift on the other, and adding jump features to the online baseline yields no statistically significant improvement. Residual jumps therefore behave as a *within-proof relative marker* of failure rather than a fixed-threshold signal. In the language of the framework, the condition that fails is the absolute tolerance in the validity coupling, while extremality within a trace persists. Beyond cheap step features, the jump improves a learned localizer in one of three settings, and tail-separation, overshoot and grokking diagnostics do not support the heavy-tail mechanism.

Contributions.

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.

- Conditional laws connecting maxima to first rejection, with the required validity and dependence assumptions stated explicitly, and a Kesten-type surrogate and pooled-trace tail-index analysis (Section 3).
- Localization evidence across provers, benchmarks and temperatures, including positional baselines and a prefix-restricted analysis that excludes text after the rejection (Section 5).
- Empirical diagnostics of trace-level coupling, clustering of large steps and step-level coupling that characterize departures from the idealized localization limit (Section 6), complemented by online (Section 7) and learned (Section 8) uses of the jump.
- Tests of tail behavior and length-dependent success in formal proofs, controlled deduction and modular addition (Section 9).

The theory specifies what would have to hold for the largest step to be the failure; the experiments measure which of these conditions trained provers satisfy.

2 Measurement and evaluation target

Eligible failures and the first rejected step. For proof attempt i with extracted formal steps $s_{i1}, \dots, s_{iL_i}$, whole-proof verification in Lean 4 (de Moura & Ullrich, 2021) establishes the outcome, and semantic replay identifies

$$t_i^* = \min\{t : \ s_{it} \text{ is rejected during the aligned replay}\}. \tag{1}$$

We write $E_i = 1$ when an attempt is unsuccessful and has a reliable first-rejection label, token alignment and score extraction. All localization results condition on $E_i = 1$. Steps are extracted with Lean syntax and semantic replay, preserving compound tactics; parse errors, resource failures and terminal unsolved goals are not attributed to the last tactic (Appendix B).

Residual changes and surprisal. At a fixed layer, $X_{it} \in \mathbb{R}^{4096}$ is the decoder block output at the last token of s_{it}, before final normalization, and X_{i0} is taken at formal-proof entry. We define

$$\xi_{it} = X_{it} - X_{i,t-1}, \qquad w_{it} = \left\|\hat{\Sigma}^{-1/2}(\xi_{it} - \hat{\mu})\right\|_2, \tag{2}$$

with mean and covariance (isotropic shrinkage 0.1) fitted on verified proofs from a disjoint calibration fold. Whitening adjusts for anisotropy of increment coordinates, including massive-activation directions (Sun et al., 2024). Mean step surprisal on the same token sequence is

$$q_{it} = -\frac{1}{|I_{it}|} \sum_{k \in I_{it}} \log p_\theta(y_{ik} \mid \pi_i, y_{i,<k}), \tag{3}$$

where I_{it} are the token positions of step s_{it} and π_i is the prompt. The jump and surprisal localizers are

$$J_i = \min \arg\max_{1 \le t \le L_i} w_{it}, \qquad S_i = \min \arg\max_{1 \le t \le L_i} q_{it}. \tag{4}$$

Both are retrospective: all steps of the generated proof, including those after the first rejection, compete for the maximum.

Metrics and the role of the baseline. For N eligible attempts, the exact-localization rate and uniform reference are $\hat{a}_J = N^{-1} \sum_{i:E_i=1} \mathbf{1}\{J_i = t_i^*\}$ and $\hat{a}_U = N^{-1} \sum_{i:E_i=1} L_i^{-1}$. The uniform rule adjusts for length but not for concentration of failures or maxima at particular positions. Even if the two positions are conditionally independent,

$$\mathbb{P}(J = t^* \mid L, E) = \sum_{t=1}^{L} \mathbb{P}(J = t \mid L, E)\, \mathbb{P}(t^* = t \mid L, E), \tag{5}$$

which need not equal $1/L$. Exceeding the uniform reference therefore shows that the score carries positional information; whether it carries information beyond position is tested directly in Section 8. Uncertainty resamples whole tasks (exercise groups on ProofNet) with their attempts.

3 CONDITIONAL EXTREME-VALUE FRAMEWORK

The localizers above need no tail assumption. This section gives a separate idealized model for interpreting maxima. Let (Z_t) be nonnegative step deviations, $M_L = \max_{t \le L} Z_t$, and V_L indicate acceptance of all L steps. Until the first rejection, v_t records step validity; afterwards labels are absorbing, $v_s = 0$. Absorption is a measurement convention.

Hypothesis 1 (Validity coupling)**.** For a fixed read-out and deviation statistic there is $\tau > 0$ such that, up to and including the first rejection, $v_t = \mathbf{1}\{Z_t \le \tau\}$.

Under Hypothesis 1 and absorption, $\{V_L = 0\} = \{M_L > \tau\}$ and the first rejection is the first exceedance.

Write $\overline{F} = 1 - F$. Regular variation with index $-\alpha < 0$, $\overline{F} \in \mathrm{RV}_{-\alpha}$, means $\overline{F}(xy)/\overline{F}(x) \to y^{-\alpha}$ for all $y > 0$; rapid variation means the ratio tends to zero for $y > 1$. For a stationary regularly varying sequence the spectral tail process (Θ_t) satisfies $\Theta_0 = 1$ and

$$\mathcal{L}\big(Z_0/x, (Z_t/Z_0)_{|t| \le h} \mid Z_0 > x\big) \Rightarrow \mathcal{L}\big(Y, (\Theta_t)_{|t| \le h}\big), \tag{6}$$

with $\mathbb{P}(Y > y) = y^{-\alpha}$, $y \ge 1$, and Y independent of Θ (Basrak & Segers, 2009). Because a finite-context model has finitely many possible read-outs, tail statements describe an intermediate range of the observed distribution.

3.1 OVERSHOOTS AND LENGTH

Proposition 1 (Marginal overshoot)**.** *For $u < x_F$ let $Q_u(y) = \mathbb{P}(Z > uy \mid Z > u)$. If $\overline{F} \in \mathrm{RV}_{-\alpha}$, then $Q_u(y) \to y^{-\alpha}$ for $y \ge 1$. If $\overline{F}$ is rapidly varying, $Q_u(y) \to 0$ for $y > 1$; the same holds as $u \uparrow x_F$ when $0 < x_F < \infty$.*

This marginal result needs no independence; selecting the first rejected step in a dependent sequence can change the law (Theorem 3).

Proposition 2 (Conditional failure–length law)**.** *Assume Hypothesis 1, absorption, and a stationary sequence with continuous marginal F, $x_F = \infty$ and extremal index θ. If $\tau_L \to \infty$ and $L\overline{F}(\tau_L) \to \lambda \in (0, \infty)$, then $\mathbb{P}(V_L = 0) = 1 - e^{-\theta\lambda} + o(1)$, and $1 - e^{-\theta\lambda} = \theta\lambda + O(\lambda^2)$.*

The extremal-index assumption means the maximum probability converges to $e^{-\theta\lambda}$ (Leadbetter et al., 1983). This joint high-threshold, long-window limit motivates the exponential success–length predictor of Section 9, whose single coefficient absorbs θ and τ jointly.

3.2 LOCALIZATION WITH AND WITHOUT EXTREMAL CLUSTERING

Theorem 3 (Conditional localization)**.** *Assume Hypothesis 1 and absorption. Let $(Z_t)_{t \in \mathbb{Z}}$ be stationary with continuous marginal F and $x_F = \infty$. Fix L and let $\tau \to \infty$. Let T be the earliest argmax on $1, \dots, L$ and $N_\tau = \#\{t \le L : Z_t > \tau\}$.*

(i) If the sequence is regularly varying and $\mathbb{P}(\Theta_j = 1) = 0$ for $j \ne 0$, then

$$\mathbb{P}(T = t^* \mid V_L = 0) \to \pi_L(\Theta) = \frac{\sum_{t=1}^{L} \mathbb{P}(Y\Theta_{s-t} \le 1 \ \forall s < t,\ \Theta_{s-t} < 1 \ \forall s > t)}{\sum_{t=1}^{L} \mathbb{P}(Y\Theta_{s-t} \le 1 \ \forall s < t)}. \tag{7}$$

The limit lies in $[0, 1]$ and is positive if $\mathbb{P}(\Theta_j < 1,\ 1 \le j < L) > 0$. If Θ has atoms at one, strict and non-strict future inequalities bound the $\liminf$ *and* $\limsup$.

(ii) Under extremal independence, $\mathbb{P}(Z_s > x, Z_t > x) = o(\overline{F}(x))$ for $s \ne t$,

$$\mathbb{P}(N_\tau = 1 \mid V_L = 0) \to 1, \qquad \mathbb{P}(T = t^* \mid V_L = 0) \to 1, \qquad \mathbb{P}(t^* = t \mid V_L = 0) \to L^{-1}, \tag{8}$$

and $\mathbb{P}(Z_{t^*} > \tau y \mid V_L = 0)$ tends to $y^{-\alpha}$ under regular variation and to 0 for $y > 1$ under rapid variation.

Theorem 3 gives a yardstick for the data. Coupling with extremal independence yields near-perfect localization and uniform failure positions; clustering of large steps lowers the limit to $\pi_L(\Theta)$; and

without coupling, identity (5) shows how shared positional structure alone lifts localization above $1/L$. Each of these ingredients (coupling, clustering, positional nonstationarity) is estimable from saved traces, and Section 6 estimates them.

3.3 A KESTEN-TYPE SURROGATE

Write $X_t = X_t^\star + E_t$ with $X_t^\star$ a reference trajectory, and postulate

$$E_t = a_t E_{t-1} + B_t, \qquad a_t > 0, \qquad B_t \sim \mathcal{N}(0, \Sigma),\ \Sigma \succ 0, \tag{9}$$

with (a_t) and (B_t) independent i.i.d. sequences. Routing is exogenous, so a_t is independent of E_{t-1}. This is an illustrative stochastic recurrence (Kesten, 1973; Goldie, 1991; Buraczewski et al., 2016) that isolates one route to heavy tails.

Theorem 4 (Scalar multiplier mechanism)**.** *Assume* $\mathbb{E}|\log a| < \infty$ *and* $\mathbb{E}\log a < 0$ *in* (9)*. Then a unique stationary causal solution exists. Moreover:*

(i) If $\mathbb{P}(a > 1) > 0$*,* $\mathbb{E}a^\alpha = 1$ *for some* $\alpha > 0$*,* $\mathbb{E}a^{\alpha+\delta} < \infty$ *for some* $\delta > 0$*, and* $\log a$ *is non-arithmetic, then for some* $C > 0$*,*

$$\mathbb{P}(\|E_t\| > x) \sim C x^{-\alpha}, \qquad \mathbb{P}(\|E_t - E_{t-1}\| > x) \sim C\,\mathbb{E}|a-1|^\alpha x^{-\alpha}. \tag{10}$$

Adding uniformly bounded reference increments preserves this asymptotic for $\|X_t - X_{t-1}\|$*.*

(ii) If $a \le \rho < 1$ *a.s., then* $\|E_t - E_{t-1}\|$ *is rapidly varying with all moments and a Gaussian-type tail bound (Appendix A.3).*

For a two-point multiplier that expands with probability p, the tail index decreases in p (Appendix A.3), linking heavier tails to more frequent expansive updates within this scalar family.

3.4 TAIL ESTIMATION WITH DEPENDENT STEPS

Reasoning traces supply many short, internally dependent sequences. Appendix A.4 shows that the Hill estimator on steps pooled across i.i.d. traces is consistent under arbitrary within-trace dependence (Theorem 5), and gives a normal limit whose variance is inflated by a cluster term; empirical intervals therefore resample whole tasks.

4 EXPERIMENTAL SETUP

Models and read-out. We evaluate DeepSeek-Prover-V2-7B, Goedel-Prover-V2-8B and Kimina-Prover-Distill-8B (Ren et al., 2025; Lin et al., 2025; Wang et al., 2025), reading zero-based layers 14, 17 and 17. States and surprisal come from replaying the sampled token IDs (generated with vLLM; Kwon et al., 2023) through the pinned HuggingFace model.

Benchmarks and populations. On miniF2F (Zheng et al., 2022), 417 tasks per model receive eight attempts at each of temperatures 0.6 and 1.0, with calibration and evaluation separated by task. On ProofNet-Verified, a verified variant of ProofNet (Azerbayev et al., 2023), 197 exercise groups are used for evaluation with four attempts per problem at temperature 0.6 (DeepSeek and Goedel). The main analysis uses temperature 0.6, with 2,018 eligible failed attempts across five model–benchmark cells. Temperature-1.0 attempts (1,498 eligible failures) are evaluated with the transform, localizers and learned models frozen from the 0.6 analysis. Threshold diagnostics compare eligible failures with verified evaluation attempts from the same tasks or groups: 172, 231 and 217 on miniF2F and 35 and 40 on ProofNet. The learned comparison on ProofNet requires $L \ge 2$ and complete token boundaries, giving 452 DeepSeek attempts in 166 groups and 259 Goedel attempts in 133 groups (62.0% and 34.0% of non-verified attempts), with equal group weights and a whitening transform fitted on verified miniF2F calibration tasks.

Predictions and design. The framework yields five predictions: (P1) tail indices differ between failed and verified proofs; (P2) the largest increment marks the first rejected step; (P3) the increment at the first rejection has a Pareto overshoot; (P4) the tail index changes at the grokking transition; and (P5) success decays with assigned length. P2 is tested in Section 5 and P1, P3, P4 and P5 in

Table 1: Exact first-rejection localization on eligible failed proofs (%, equal attempt weights). N/G: attempts/tasks (exercise groups on ProofNet). Uniform is the average of $1/L$. Temperature-1.0 rows use rules frozen from the 0.6 analysis.

Dataset	Model	Temp.	N/G	Jump	Uniform	Surprisal
miniF2F	DeepSeek	0.6	528/122	24.4	9.7	17.6
miniF2F	Goedel	0.6	448/122	25.4	8.7	13.2
miniF2F	Kimina	0.6	324/94	17.3	8.7	12.0
ProofNet	DeepSeek	0.6	456/167	20.4	13.3	22.6
ProofNet	Goedel	0.6	262/134	29.4	14.2	22.5
miniF2F	DeepSeek	1.0	544/140	20.4	10.3	16.0
miniF2F	Goedel	1.0	456/124	20.2	9.1	14.0
miniF2F	Kimina	1.0	498/122	19.1	10.7	12.1

Table 2: Single-signal and learned localizers on the learned-comparison ProofNet population (%, equal exercise-group weights). ±1: $|\hat{t} - t^*| \leq 1$. M0/M1 predictions are out of fold (Section 8).

	DeepSeek (452/166)		Goedel (259/133)	
Localizer	Exact	±1	Exact	±1
Uniform	12.55	32.00	13.15	32.19
Argmax surprisal	21.40	36.81	21.27	43.75
Argmax jump	18.50	43.21	26.57	48.15
M0: position, span, tactic, surprisal	40.85	52.05	57.82	66.30
M1: M0 features plus jump	43.51	54.28	57.13	65.49

Section 9. The learned comparison uses DeepSeek as the primary cell, Goedel as the replication cell and $+3$ pp as the useful effect; the temperature-1.0 evaluation tests seven one-sided hypotheses with Holm correction. Cells with fewer than 30 exceedances or few verified groups have no intervals.

5 Localization within a proof

The largest increment points to the first rejection. The localization prediction (P2) holds. At temperature 0.6 the jump argmax matches the first rejection in 17.3–25.4% of eligible miniF2F failures and 20.4–29.4% of ProofNet failures, 1.5–2.9 times the uniform reference (Table 1). On miniF2F it exceeds the surprisal argmax in every cell; on ProofNet surprisal is higher for DeepSeek and lower for Goedel, and the equal-group contrasts, jump minus surprisal, are -2.58 pp $[-8.58, 3.29]$ and $+5.19$ pp $[-2.15, 12.53]$ (97.5% intervals). At temperature 1.0 the pattern holds for all three provers: 19.1–20.4% against 9.1–10.7% uniform (Holm-adjusted one-sided $p \leq 6.1 \times 10^{-4}$), again 4.4–7.0 points above surprisal. Within one step, the jump argmax reaches 40.1, 36.4 and 35.1% against 39.3, 34.2 and 24.9% for surprisal, a difference that is clear for Kimina (adjusted $p = 0.002$) and small for DeepSeek and Goedel (Appendix D).

Position and a one-step tolerance. Position is controlled by the learned localizer of Section 8, whose features include step index, relative position and first/last-step indicators. Within one step the jump argmax has the highest rate of the single signals, 43.21% and 48.15% (Table 2).

The rejected step stands out before any post-rejection text. The full-proof argmax lets text generated after the rejection compete. Restricting candidates to steps $1, \ldots, t^*$ removes that text: the rejected step then carries the largest increment in 35.3–42.9% of proofs, against 16.5–24.8% for a uniform choice (Table 3). Because t^* is always the last step of its prefix, a positional trend in w_t would also favor it, and it does in part: in failed attempts whose rejection occurs later, the last step of an error-free prefix of the same length is its maximum in 21.4–34.8% of cases. On lengths with control support, the rejected step exceeds this positional reference by 8.2–16.7 points in every cell, retaining 45–98% of the excess over uniform, also without the first increment and with equal group weights (Appendix D).

Table 3: Prefix-restricted localization for attempts with $t^* \geq 2$ (%). Prefix hit: the argmax over $t \leq t^*$ is t^*; Uniform: mean of $1/t^*$. Positional control: endpoint-maximum rate in error-free prefixes of the same length from attempts failing later. The positional comparison uses only lengths with control support (Appendix D; at most 0.7% excluded); its reported contrast Δ_{pos} is in pp with 95% task/group bootstrap intervals.

Dataset	Model	Prefix hit	Uniform	Positional control	Δ_{pos}
miniF2F	DeepSeek	38.52	16.97	25.89	$+12.71$ [6.81, 18.54]
miniF2F	Goedel	38.19	16.51	23.92	$+14.36$ [8.10, 20.80]
miniF2F	Kimina	35.27	16.71	21.40	$+14.12$ [7.05, 21.61]
ProofNet	DeepSeek	42.86	24.76	34.80	$+8.18$ [3.08, 13.35]
ProofNet	Goedel	41.34	24.40	24.93	$+16.65$ [9.96, 23.14]

Table 4: Threshold diagnostics at $q = 0.95$ (%, 95% task/group bootstrap intervals); (b) and (c) are conditional on $M > u$ in failed proofs. Verified proofs come from the same tasks or groups as the failed ones; ProofNet verified cells are descriptive (35 and 40 attempts). All quantiles in Table 13.

		(a) $\mathbb{P}(M > u)$		(b) $T = F$	(c) $F = t^*$
Dataset	Model	Failed	Verified	Failed	Failed
miniF2F	DeepSeek	44.1 [36.2, 52.4]	25.0 [16.1, 35.5]	48.9 [39.9, 57.6]	15.0 [7.1, 24.0]
miniF2F	Goedel	50.5 [42.1, 58.4]	29.0 [19.4, 40.2]	37.2 [28.6, 46.8]	17.7 [9.4, 26.9]
miniF2F	Kimina	57.7 [47.9, 67.4]	35.5 [24.0, 48.0]	38.0 [29.0, 47.6]	19.8 [11.7, 28.8]
ProofNet	DeepSeek	84.7 [79.7, 89.2]	62.9	25.9 [20.9, 31.3]	22.5 [17.1, 28.0]
ProofNet	Goedel	87.8 [83.0, 92.1]	75.0	23.5 [17.1, 30.3]	26.1 [18.2, 34.3]

6 Coupling and clustering in trained provers

Theorem 3 attributes perfect localization to coupling (the first rejection is the first exceedance of a fixed tolerance) and extremal independence (large steps do not cluster). For thresholds u_q, the q-quantiles of increments on verified calibration proofs, let $M = \max_t w_t$, $F = \min\{t : w_t > u_q\}$ and T the earliest argmax. We measure (a) trace-level coupling, $\mathbb{P}(M > u_q)$ in failed versus verified proofs of the same tasks; (b) clustering, $\mathbb{P}(T = F \mid M > u_q)$ in failed proofs; and (c) step-level coupling, $\mathbb{P}(F = t^* \mid M > u_q)$ in failed proofs.

A fixed level is associated with failure but does not identify it. At $q = 0.95$ on miniF2F, 44–58% of failed proofs exceed the threshold against 25–35% of verified proofs of the same tasks (Table 4), and the gap persists at every quantile (Figure 1). The association is real, but at no threshold does exceedance separate the two outcomes: failed proofs without an exceedance and verified proofs with one are both common.

Step-level coupling and extremal independence both depart from the idealized limit. When a failed proof exceeds the threshold, its first exceedance is the rejected step in 15–20% of miniF2F cases and 23–26% on ProofNet at $q = 0.95$, and below a third at every quantile. The clustering diagnostic shows that the maximum coincides with the first exceedance in 37–49% of miniF2F cases at $q = 0.95$ and 49–59% at $q = 0.99$, so a later and larger step often follows the first one. Under the theorem's conditions both quantities would tend to one.

The absolute scale shifts across benchmarks. With the whitening transform and thresholds calibrated on miniF2F, 85–88% of failed and 63–75% of verified ProofNet proofs exceed the $q = 0.95$ threshold. The same fixed transform places ProofNet increments far higher, so a tolerance calibrated on one benchmark does not carry over to another. Together with Section 5, these diagnostics locate the departure from Hypothesis 1 in its absolute tolerance: the rejected step is often extreme *relative to its own proof*, but not beyond a common level.

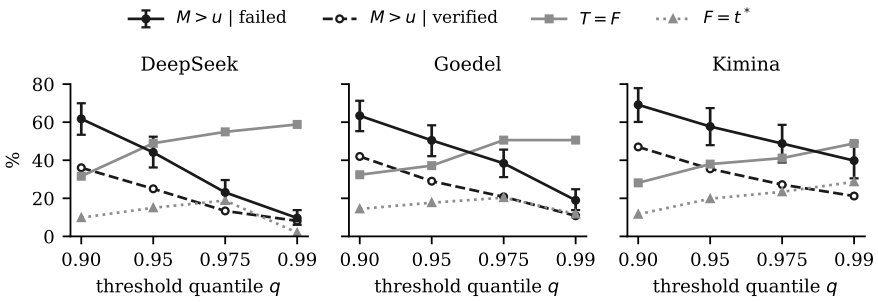

Figure 1: Threshold diagnostics on miniF2F across quantiles q of verified calibration increments. Failed proofs exceed the threshold more often than verified proofs of the same tasks (black); the first exceedance falls on the rejected step in 2–29% of cases (grey dotted), and the maximum coincides with the first exceedance in 28–59% (grey solid). Error bars: 95% task-bootstrap intervals.

Table 5: Online detection on ProofNet (%, equal group weights): exact detection at t^* with no earlier flag, and realized false-positive rate on verified evaluation attempts. Primary contrast: DeepSeek at 10%.

Model	Target FPR	Exact H0	Exact H1	Δ (pp)	95% interval	Realized FPR H0 / H1
DeepSeek	10	13.85	13.39	−0.45	[−2.51, 1.86]	15.6 / 15.6
DeepSeek	20	19.86	21.08	+1.22	[−0.80, 3.33]	28.1 / 34.4
Goedel	10	8.33	9.96	+1.63	[−1.50, 5.14]	4.4 / 4.4
Goedel	20	19.14	14.66	−4.47	[−8.67, −0.65]	14.9 / 12.3

7 Online detection

Online detection judges each step before the rest of the proof exists and therefore tests the jump as a prospective signal. We fit a discrete hazard model on ProofNet: for failed attempts the steps $1, \ldots, t^*$ enter with target $\mathbf{1}\{t = t^*\}$, and verified attempts of the same groups enter with all steps as survivors. The baseline H0 uses only past and present features (step index, $\log t$, span length, tactic category, surprisal and its running statistics); H1 adds $\log(1 + w_t)$, the running maximum of w and the ratio of w_t to that maximum. Both use the nested group cross-validation of Section 8. The detector flags the first step whose hazard exceeds a threshold set on inner out-of-fold verified attempts for a target false-positive rate.

Online exact localization reaches 8–21%, and jump features do not improve it: the change is -0.45 pp in the primary cell ($[-2.51, 1.86]$) and ranges from -4.47 to $+1.63$ pp elsewhere (Table 5); for Goedel at a 20% target the jump features lower it ($[-8.67, -0.65]$). Realized false-positive rates depart from their targets (15.6% at a 10% target for DeepSeek): a detection threshold calibrated on held-out verified proofs shifts on evaluation proofs, as the threshold diagnostics of Section 6 predict. The jump localizes the first rejection relative to the complete proof and adds no value to step-by-step detection.

8 Added value beyond cheap step features

Design. We compare two independently fitted ridge-regularized conditional-logit localizers (McFadden, 1974). For attempt i, each assigns a distribution over candidate steps,

$$p_{it}^{(m)} = \frac{\exp(r_{it}^{(m)})}{\sum_{s=1}^{L_i} \exp(r_{is}^{(m)})}, \qquad r_{it}^{(0)} = \theta_0^\top \phi(b_{it}), \qquad r_{it}^{(1)} = \theta_1^\top \phi(b_{it}) + \beta u_{it}. \tag{11}$$

Baseline features are step index t, relative position t/L_i, first/last-step indicators, $\log(1 + n_{it})$ for token-span length, AST tactic category and mean surprisal, with cubic B-splines for the four numeric features; u_{it} is standardized $\log(1+w_{it})$ and β is unconstrained in sign. The loss is the negative log-probability of t_i^* with equal total weight per exercise group, under five outer and three inner group

Table 6: Incremental effect of the jump on ProofNet (equal group weights). $\Delta H = H_1 - H_0$ in percentage points; $\Delta\mathrm{NLL} = \mathrm{NLL}_0 - \mathrm{NLL}_1$ in nats, positive favors M1. 95% paired group-bootstrap intervals.

Cell	H_0	H_1	ΔH	95% interval	ΔNLL	95% interval
DeepSeek (primary)	40.85	43.51	$+2.66$	$[0.45, 5.17]$	$+0.009$	$[-0.004, 0.022]$
Goedel (replication)	57.82	57.13	-0.69	$[-3.13, 1.57]$	-0.056	$[-0.092, -0.023]$

folds (Varma & Simon, 2006); all preprocessing and the ridge parameter are fitted within training splits, and paired intervals resample whole groups 20,000 times on the saved out-of-fold predictions (Appendix C).

Results. The learned localizer M0 is far stronger than any single signal (40.9 and 57.8% exact; Table 2). For DeepSeek, adding the jump raises exact localization to 43.51%, a gain of $+2.66$ pp ($[0.45, 5.17]$; 97.5% family interval $[0.25, 5.57]$), with 14 attempts corrected and 6 lost, concentrated in short proofs (Table 9). For Goedel the change is -0.69 pp ($[-3.13, 1.57]$) with worse out-of-fold NLL (Table 6). With M0 and M1 refitted on miniF2F at temperature 0.6 and frozen, the DeepSeek gain is not confirmed on temperature-1.0 samples (Holm-adjusted one-sided $p = 0.48$). The standardized jump coefficient is positive in every outer fold for both provers (0.11–0.42), so larger increments are associated with the first rejection conditional on position, span, tactic and surprisal, while the held-out gain from that association is small and model-dependent.

Agreement with surprisal. Predicting a region only when the jump and surprisal maxima fall on the same or adjacent steps covers 23–45% of eligible failures (Appendix E), and the region, at most two steps long, contains the first rejection in 45–60% of them, 1.4–2.0 times as often as a random window of the same size. Against the jump and the surprisal given the same number of candidates on proofs of the same length, all ten differences are positive ($+2$ to $+17$ points), and three exclude zero after Bonferroni correction over 40 contrasts (DeepSeek on miniF2F against both signals, Goedel on miniF2F against surprisal). Exact agreement gives 48.7–61.4% at 10–22% coverage for DeepSeek and Goedel; with a one-step margin, none of the adjusted contrasts against matched single signals excludes zero. Agreement selects regions enriched for first rejections within failed proofs. As a detector of failed proofs, it shows no statistically significant gain over the length-only reference in any cell.

9 Other predictions of the framework

Success versus length (P5, supported). Following Proposition 2, we test Qwen2.5-7B-Instruct (Qwen Team, 2024) on synthetic modus-ponens chains with distractors (Saparov & He, 2023), with assigned lengths $L = 3, \ldots, 12$, 1,000 calibration and 500 held-out tasks per temperature, and a symbolic checker. A one-parameter model $p_c(L) = \exp(-cL)$ fitted on calibration outcomes lowers held-out Brier score from 0.1398 to 0.1269 at temperature 1.0 ($c = 0.268$), an improvement of 9.2% (97.5% interval $[0.0024, 0.0242]$ in absolute terms), and by 0.0065 ($[-0.0040, 0.0178]$) at 0.6, so assigned length carries predictive information as the failure–length law predicts.

Tail behavior (P1, P3, P4; not supported). A Pareto law at the first rejection (P3) remains unconfirmed, consistent with the absence of a common threshold in Section 6. For P1, refuted-minus-verified moment-estimator contrasts have 95% intervals $[-1.333, 0.754]$, $[-0.239, 0.141]$ and $[-0.995, 0.353]$ for DeepSeek, Goedel and Kimina, establishing neither a difference nor equality (Appendix F). The peak-to-neighbor ratio of the maximum shows no statistically significant outcome effect after length and task adjustment. For P4, in five paired modular-addition trainings (Nanda et al., 2023), signed tail estimates are nonpositive before and after the generalization transition and for shuffled-label controls, so no transition to a positive tail-index estimate is observed.

10 Related work

Extreme values and recurrences. Tail processes describe dependence around large observations (Basrak & Segers, 2009; Kulik & Soulier, 2020), and the extremal index links clustering to limits of maxima (Leadbetter et al., 1983). Kesten–Goldie theory (Kesten, 1973; Goldie, 1991; Buraczewski et al., 2016) yields heavy tails of SGD iterates (Gürbüzbalaban et al., 2021; Hodgkinson & Mahoney, 2021); heavy tails also appear in gradient noise (Şimşekli et al., 2019) and weight spectra (Martin & Mahoney, 2021).

Error signals and formal reasoning. Hidden-state probes detect falsehoods and hallucinations (Azaria & Mitchell, 2023; Orgad et al., 2025), semantic entropy detects confabulation (Farquhar et al., 2024), and process supervision assesses individual reasoning steps (Lightman et al., 2024); our learned comparison tests one scalar geometric feature against cheap step features. The provers we study generate verifiable Lean proofs (Ren et al., 2025; Lin et al., 2025; Wang et al., 2025); compositionality studies model failure as compounding per-step error (Dziri et al., 2023; Saparov & He, 2023), and work on modular addition (Nanda et al., 2023) and heuristic arithmetic (Nikankin et al., 2025) motivates the algorithm/heuristic contrast behind the tail regimes.

11 Discussion and conclusion

Residual jumps provide a relative marker of the first rejection within a proof: the full-proof maximum finds it about twice as often as a length-adjusted uniform choice across provers, benchmarks and temperatures, and it is the largest of the steps up to the rejection in 35–43% of proofs. Fixed-threshold diagnostics show departures from both step-level coupling and extremal independence, while the increment scale shifts between benchmarks. Adding jump features yields no statistically significant improvement in the online comparisons. Controlled deduction also supports the predictive value of assigned reasoning length. These findings motivate relative statistics, such as a step's rank within its proof, as features for localization without a step verifier, and a trace-relative coupling in which rejection is tied to being extreme within a trajectory. We plan to evaluate whether residual jumps localize the first error in informal mathematical proofs, using independently annotated reasoning steps.

Limitations. Eligibility covers 62% (DeepSeek) and 34% (Goedel) of non-verified ProofNet attempts. The full-proof localizers use complete generations.

Reproducibility statement

Appendix A gives the proofs; Appendices B–G give read-out, segmentation, frozen revisions, the group-weighted objective, feature preprocessing, nested cross-validation and uncertainty procedures. The experimental archive fixes code and data by SHA256 before localizer fitting and records group splits, preprocessing parameters, coefficients, per-step out-of-fold predictions, per-group contrasts, bootstrap replicates and per-attempt exclusions. Reproduction additionally requires the prompts and the executable model and Lean environments.

AI use statement

We used generative AI tools to polish the language of the manuscript and to check consistency between manuscript claims and reported experimental results. We have reviewed all AI-assisted edits and take responsibility for the final content of this work.

References

Amos Azaria and Tom Mitchell. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023.

Zhangir Azerbayev, Bartosz Piotrowski, Hailey Schoelkopf, Edward W. Ayers, Dragomir Radev, and Jeremy Avigad. ProofNet: Autoformalizing and formally proving undergraduate-level mathematics. *arXiv preprint arXiv:2302.12433*, 2023.

Bojan Basrak and Johan Segers. Regularly varying multivariate time series. *Stochastic Processes and their Applications*, 119(4):1055–1080, 2009.

Leo Breiman. On some limit theorems similar to the arc-sin law. *Theory of Probability and its Applications*, 10(2):323–331, 1965.

Dariusz Buraczewski, Ewa Damek, and Thomas Mikosch. *Stochastic Models with Power-Law Tails: The Equation $X = AX + B$*. Springer, 2016.

Laurens de Haan and Ana Ferreira. *Extreme Value Theory: An Introduction*. Springer, 2006.

Leonardo de Moura and Sebastian Ullrich. The Lean 4 theorem prover and programming language. In *Automated Deduction – CADE 28*, 2021.

Arnold L. M. Dekkers, John H. J. Einmahl, and Laurens de Haan. A moment estimator for the index of an extreme-value distribution. *Annals of Statistics*, 17(4):1833–1855, 1989.

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Sean Welleck, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D. Hwang, Soumya Sanyal, Xiang Ren, Allyson Ettinger, Zaïd Harchaoui, and Yejin Choi. Faith and fate: Limits of transformers on compositionality. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630, 2024.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *PMLR*, pp. 2151–2159, 2019.

Charles M. Goldie. Implicit renewal theory and tails of solutions of random equations. *Annals of Applied Probability*, 1(1):126–166, 1991.

Mert Gürbüzbalaban, Umut Şimşekli, and Lingjiong Zhu. The heavy-tail phenomenon in SGD. In *International Conference on Machine Learning (ICML)*, 2021.

Liam Hodgkinson and Michael W. Mahoney. Multiplicative noise and heavy tails in stochastic optimization. In *International Conference on Machine Learning (ICML)*, 2021.

Harry Kesten. Random difference equations and renewal theory for products of random matrices. *Acta Mathematica*, 131:207–248, 1973.

Rafał Kulik and Philippe Soulier. *Heavy-Tailed Time Series*. Springer, 2020.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles (SOSP)*, 2023.

M. Ross Leadbetter, Georg Lindgren, and Holger Rootzén. *Extremes and Related Properties of Random Sequences and Processes*. Springer, 1983.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations (ICLR)*, 2024.

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, 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. *arXiv preprint arXiv:2508.03613*, 2025.

Charles H. Martin and Michael W. Mahoney. Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning. *Journal of Machine Learning Research*, 22(165):1–73, 2021.

Daniel McFadden. Conditional logit analysis of qualitative choice behavior. In P. Zarembka (ed.), *Frontiers in Econometrics*, pp. 105–142. Academic Press, 1974.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *International Conference on Learning Representations (ICLR)*, 2023.

Yaniv Nikankin, Anja Reusch, Aaron Mueller, and Yonatan Belinkov. Arithmetic without algorithms: Language models solve math with a bag of heuristics. In *International Conference on Learning Representations (ICLR)*, 2025.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. LLMs know more than they show: On the intrinsic representation of LLM hallucinations. In *International Conference on Learning Representations (ICLR)*, 2025.

James Pickands. Statistical inference using extreme order statistics. *Annals of Statistics*, 3(1):119–131, 1975.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

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. *arXiv preprint arXiv:2504.21801*, 2025.

Sidney I. Resnick. *Extreme Values, Regular Variation, and Point Processes*. Springer, 1987.

Sidney I. Resnick. *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*. Springer, 2007.

Abulhair Saparov and He He. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. In *International Conference on Learning Representations (ICLR)*, 2023.

Umut Şimşekli, Levent Sagun, and Mert Gürbüzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In *International Conference on Machine Learning (ICML)*, 2019.

Mingjie Sun, Xinlei Chen, J. Zico Kolter, and Zhuang Liu. Massive activations in large language models. In *Conference on Language Modeling (COLM)*, 2024.

Sudhir Varma and Richard Simon. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7:91, 2006.

Haiming Wang, Mert Unsal, Xiaohan Lin, Mantas Baksys, Junqi Liu, Marco Dos Santos, Flood Sung, Marina Vinyes, Zhenzhe Ying, Zekai Zhu, et al. Kimina-Prover Preview: Towards large formal reasoning models with reinforcement learning. *arXiv preprint arXiv:2504.11354*, 2025.

Kunhao Zheng, Jesse Michael Han, and Stanislas Polu. miniF2F: A cross-system benchmark for formal Olympiad-level mathematics. In *International Conference on Learning Representations (ICLR)*, 2022.

A Proofs and full estimation assumptions

A.1 Overshoots and the failure–length law

For Proposition 1, $Q_u(y) = \overline{F}(uy)/\overline{F}(u)$, so the claims for regularly and rapidly varying laws follow from the definitions. If $0 < x_F < \infty$, then $1 < Z/u \leq x_F/u$ given $Z > u$, and the upper bound tends to one. Unbounded Gumbel-domain laws are a standard rapidly varying case (de Haan & Ferreira, 2006; Resnick, 1987): their auxiliary function f satisfies $f(u)/u \to 0$ and

$\overline{F}(u + xf(u))/\overline{F}(u) \to e^{-x}$. For fixed $y > 1$ and $M > 0$, eventually $(y-1)u \geq Mf(u)$, so $\limsup \overline{F}(uy)/\overline{F}(u) \leq e^{-M}$; let $M \to \infty$.

For Proposition 2, coupling identifies $\{V_L = 0\}$ with $\{M_L > \tau_L\}$, and the extremal-index limit gives the formula; the small-λ statement is a Taylor expansion of the limit. The two approximations are logically separate: the joint high-threshold, long-window limit comes first, and only its limiting expression is expanded. Neither implies an exact fixed-threshold exponential law.

A.2 Proof of conditional localization

Let $A_t = \{Z_t > \tau,\ Z_s \leq \tau \text{ for } s < t\}$. These disjoint events partition $\{M_L > \tau\}$, and coupling gives $t^* = t$ on A_t. For a regularly varying sequence, stationarity and (6) give

$$\mathbb{P}(A_t)/\overline{F}(\tau) \to q_t = \mathbb{P}(Y\Theta_{s-t} \leq 1 \text{ for } s < t), \tag{12}$$

$$\mathbb{P}(A_t, T = t)/\overline{F}(\tau) \to r_t = \mathbb{P}(Y\Theta_{s-t} \leq 1 \text{ for } s < t,\ \Theta_{s-t} < 1 \text{ for } s > t). \tag{13}$$

The past boundaries have probability zero because Y is continuous and independent of Θ, and $\mathbb{P}(\Theta_j = 1) = 0$ removes the future ones. Thus the conditional probability tends to $\sum_t r_t / \sum_t q_t$. The denominator is at least $q_1 = 1$, $0 \leq r_t \leq q_t$, and $r_1 = \mathbb{P}(\Theta_j < 1,\ 1 \leq j < L)$ gives the positivity criterion. With atoms at one, the Portmanteau theorem applied to strict and non-strict future inequalities gives the bounds. This proves part (i) of Theorem 3.

Under extremal independence and fixed L, $\mathbb{P}(N_\tau \geq 2) \leq \sum_{s<t} \mathbb{P}(Z_s > \tau, Z_t > \tau) = o(\overline{F}(\tau))$. Also $\mathbb{P}(A_t) = \overline{F}(\tau)(1 + o(1))$ and $\mathbb{P}(M_L > \tau) = L\overline{F}(\tau)(1 + o(1))$. Conditional on failure there is exactly one exceedance with probability tending to one; it is both the first rejection and the global maximum, and each position has limiting probability $1/L$. For $y \geq 1$,

$$\mathbb{P}(Z_{t^*} > \tau y \mid V_L = 0) = \frac{L\overline{F}(\tau y) + o(\overline{F}(\tau))}{L\overline{F}(\tau) + o(\overline{F}(\tau))}, \tag{14}$$

and regular or rapid variation completes part (ii). The argument does not infer extremal independence or coupling from an observed localization rate.

A.3 Proof of the scalar multiplier mechanism

Put $\Pi_{t,0} = 1$ and $\Pi_{t,j} = a_t \cdots a_{t-j+1}$. The strong law gives $\Pi_{t,j}^{1/j} \to e^{\mathbb{E}\log a} < 1$ a.s., and Gaussian innovations grow subexponentially a.s. by Borel–Cantelli, so $E_t = \sum_{j\geq 0} \Pi_{t,j} B_{t-j}$ converges absolutely, solves the recurrence and is its stationary causal solution; uniqueness follows from contraction of the products (Buraczewski et al., 2016). Conditional on the multipliers, E_t is Gaussian with covariance $R_t\Sigma$, where $R_t = \sum_{j\geq 0} \Pi_{t,j}^2 = a_t^2 R_{t-1} + 1$.

Heavy tails. Goldie's renewal theorem (Goldie, 1991) with multiplier a^2 and exponent $\alpha/2$ gives $\mathbb{P}(R_t > x) \sim C_R x^{-\alpha/2}$, $C_R > 0$; the moment condition implies $\mathbb{E}a^\alpha \log^+ a < \infty$, and the renewal condition is assumed. Since $E_t \overset{d}{=} \sqrt{R_t}\, G$ with $G \sim \mathcal{N}(0, \Sigma)$ independent of R_t, Breiman's lemma (Breiman, 1965) gives $\mathbb{P}(\|E_t\| > x) \sim C_R \mathbb{E}\|G\|^\alpha x^{-\alpha}$. Next $E_t - E_{t-1} = (a_t - 1)E_{t-1} + B_t$ with a_t independent of E_{t-1}, and the same lemma applies because $\mathbb{E}|a-1|^{\alpha+\delta} < \infty$. Light perturbations do not change the asymptotic: if $\|V\|$ is regularly varying and $\mathbb{P}(\|W\| > \varepsilon x) = o(\mathbb{P}(\|V\| > x))$, the triangle inequality places $\mathbb{P}(\|V + W\| > x)/\mathbb{P}(\|V\| > x)$ between $(1+\varepsilon)^{-\alpha} + o(1)$ and $(1-\varepsilon)^{-\alpha} + o(1)$; letting $\varepsilon \downarrow 0$ covers the Gaussian term and bounded reference increments.

Light tails. If $a \leq \rho < 1$, then $R_{t-1} \leq (1-\rho^2)^{-1}$ and, conditional on the multipliers, the increment has covariance $s_t^2\Sigma$ with $s_t^2 = 1 + (a_t - 1)^2 R_{t-1} \in [1, (2-\rho^2)/(1-\rho^2)]$. Gaussian concentration for $\|G\|$, with $G \sim \mathcal{N}(0, \Sigma)$ and $s_\rho^2 = (2-\rho^2)/(1-\rho^2)$, gives $\mathbb{P}(\|E_t - E_{t-1}\| > x) \leq \exp\{-(x/s_\rho - \mathbb{E}\|G\|)_+^2/(2\|\Sigma\|_{\mathrm{op}})\}$ and all moments. For rapid variation, let $\psi(u) = \mathbb{P}(\|G\| > u)$; a one-coordinate Gaussian lower bound along a top eigenvector and the concentration upper bound give $\psi(uy)/\psi(u) \to 0$ for $y > 1$, hence $\mathbb{E}\psi(xy/s_t)/\mathbb{E}\psi(x/s_t) \leq \sup_{u \geq x/s_\rho} \psi(uy)/\psi(u) \to 0$.

Moment index and the two-point example. Suppose a has all positive moments. Since $\|E_t\| \overset{d}{=} \sqrt{R_t}\|G\|$, $\mathbb{E}\|E_t\|^s < \infty$ iff $\mathbb{E}R_t^{s/2} < \infty$. If $\mathbb{E}a^s < 1$, subadditivity ($s \leq 2$) or Minkowski's inequality ($s > 2$) bounds this moment by a geometric series; if $\mathbb{E}a^s > 1$, $R_t \geq \Pi_{t,j}^2$ gives $\mathbb{E}R_t^{s/2} \geq$

$(\mathbb{E}a^s)^j \to \infty$. When $\mathbb{P}(a > 1) > 0$ and $\mathbb{E}\log a < 0$, convexity of $s \mapsto \log \mathbb{E}a^s$ gives a unique positive root α, the moment index; if $a \le 1$ a.s., all moments exist. The increment has the same moment index by the triangle and reverse triangle inequalities. For the two-point law, $g_p(s) = (1-p)\rho^s + p\kappa^s - 1$ is strictly convex, vanishes at zero and has negative slope there exactly when $p < p_c$, so it has a unique positive root on that range. Since $\partial_p g_p(s) = \kappa^s - \rho^s > 0$, the root decreases strictly and continuously in p; as $p \downarrow 0$, $\alpha(p) \to \infty$ and $p\kappa^{\alpha(p)} \to 1$, so $\alpha(p) \sim \log(1/p)/\log\kappa$; as $p \uparrow p_c$, the slope at zero vanishes and strict convexity forces $\alpha(p) \downarrow 0$. These statements need no non-arithmetic condition, which is required only for the exact power-tail constant.

A.4 Pooled-trace estimation: consistency and a normal limit

Let $Z^{(1)}, \dots, Z^{(m)}$ be i.i.d. traces of fixed length L with arbitrary within-trace dependence, average marginal survival $\overline{F}(x) = L^{-1}\sum_{t\le L}\mathbb{P}(Z_t > x)$ and pooled sample size $n = mL$. The Hill estimator on the pooled descending order statistics is $\hat{\gamma}_k^H = k^{-1}\sum_{j\le k}\log(Z_{(j)}/Z_{(k+1)})$.

Theorem 5 (Pooled-trace consistency)**.** *If $\overline{F} \in \mathrm{RV}_{-\alpha}$, $k \to \infty$ and $k/n \to 0$, then $\hat{\gamma}_k^H \to 1/\alpha$ in probability for fixed L, without any condition on dependence within a trace.*

Proof of Theorem 5. Let $b = b(n/k)$ be the upper quantile of the average marginal law, so $(n/k)\overline{F}(b) \to 1$, and define $\nu_m = k^{-1}\sum_{j\le m}\sum_{t\le L}\delta_{Z_t^{(j)}/b}$. For bounded continuous f supported in $[y_0,\infty]$, $y_0 > 0$, the expectation of $\nu_m(f)$ tends to its integral under $\nu_\alpha(y,\infty] = y^{-\alpha}$. A trace contributes at most L points; if it contributes N, then $N^2 \le LN$, and independence across traces gives

$$\operatorname{Var}\nu_m(f) \le \|f\|_\infty^2 \,\frac{L}{k}\cdot\frac{n\mathbb{P}(Z \ge by_0)}{k} \to 0, \tag{15}$$

where Z has the average marginal law. The pooled tail empirical measure therefore converges in probability under arbitrary within-trace dependence. Quantile bracketing at $b(1\pm\varepsilon)$ gives $Z_{(k+1)}/b \to 1$, and $\hat{\gamma}_k^H = \int_1^\infty \nu_m(yZ_{(k+1)}/b,\infty]\,dy/y$. Monotonicity and continuity of the limit give uniform convergence on compact y-intervals; on the bracketing event a Potter bound with $0 < \eta < \alpha$ bounds the expected remainder beyond M by a constant times $\int_M^\infty y^{-\alpha+\eta-1}dy \to 0$. Hence $\hat{\gamma}_k^H \to \int_1^\infty y^{-\alpha-1}dy = 1/\alpha$ (Resnick, 2007).

Theorem 6 (Conditional normal limit for pooled Hill estimation)**.** *In the setting of Theorem 5, put $\gamma = 1/\alpha$ and assume the average marginal law is continuous in its upper tail. Suppose its upper quantile U satisfies second-order regular variation, $\{U(tx)/U(t) - x^\gamma\}/A(t) \to x^\gamma(x^{\rho_2}-1)/\rho_2$ with $\rho_2 \le 0$ (logarithmic limit at $\rho_2 = 0$), $A(t) \to 0$ and $\sqrt{k}A(n/k) \to \lambda \in \mathbb{R}$. Suppose*

$$\sigma_{\mathrm{cl}}^2 = \lim_{x\to\infty}\frac{\operatorname{Var}\big(\sum_{t\le L} g_x(Z_t)\big)}{\gamma^2 L\overline{F}(x)} \in (0,\infty), \qquad g_x(z) = \log^+(z/x) - \gamma\mathbf{1}\{z > x\}. \tag{16}$$

With b chosen so that $n\overline{F}(b) = k$, assume local stochastic equicontinuity at one of $W_m(y) = \sqrt{k}\{k^{-1}\sum_{j,t}\mathbf{1}\{Z_t^{(j)} > by\} - (n/k)\overline{F}(by)\}$: for every $\varepsilon > 0$, $\lim_{\delta\downarrow 0}\limsup_m \mathbb{P}(\sup_{|y-1|\le\delta}|W_m(y) - W_m(1)| > \varepsilon) = 0$. Then $\sqrt{k}(\hat{\gamma}_k^H - \gamma) \Rightarrow \mathcal{N}(\lambda/(1-\rho_2), \gamma^2\sigma_{\mathrm{cl}}^2)$. For independent steps within a trace, $\sigma_{\mathrm{cl}}^2 = 1$.

Proof. Let $Y_j = \sum_t g_b(Z_t^{(j)})$, let $\tilde{\gamma}_b$ be the ratio of pooled $\log^+(Z/b)$ to the pooled number of exceedances of b, and $N_m(y) = \nu_m(y,\infty]$. Then $\sqrt{k}(\tilde{\gamma}_b - \gamma) = k^{-1/2}\sum_j Y_j/N_m(1)$, and the denominator tends to one. The second-order condition gives the usual bias $\sqrt{k}(m/k)\mathbb{E}Y_1 \to \lambda/(1-\rho_2)$, since expectations depend only on the average marginal law (de Haan & Ferreira, 2006). The numerator variance tends to $\gamma^2\sigma_{\mathrm{cl}}^2$. For Lindeberg, $|Y_1| \le L(\log^+(\max_t Z_t/b)+\gamma)\mathbf{1}\{\max_t Z_t > b\}$ and regular variation give $\mathbb{E}Y_1^4 = O(L^5\overline{F}(b))$, so $(m/k)\mathbb{E}[Y_1^2\mathbf{1}\{|Y_1| > \varepsilon\sqrt{k}\}] \to 0$ for fixed L, and the triangular-array CLT gives the deterministic-threshold limit.

Put $q_m = Z_{(k+1)}/b$ and $I_m(q) = \int_q^\infty N_m(y)dy/y$. Since $N_m(1) - 1 = O_p(k^{-1/2})$, local equicontinuity and the deterministic derivative $-\alpha$ at one give $q_m - 1 = \gamma\{N_m(1)-1\} + o_p(k^{-1/2})$, after quantile bracketing establishes tightness at this scale. Ties involve at most L observations a.s. and

contribute $O(L/k) = o(k^{-1/2})$. Hence $\hat{\gamma}_k^H = I_m(1) - \gamma\{N_m(1) - 1\} + o_p(k^{-1/2})$, the same expansion as $\tilde{\gamma}_b = I_m(1)/N_m(1)$. For independent coordinates the variance is a sum of marginal variances and $\sigma_{\mathrm{cl}}^2 = 1$; otherwise covariance terms remain. □

Local equicontinuity is a separate assumption, not a consequence of the scalar variance limit.

B Read-out, segmentation and frozen revisions

DeepSeek has 30 decoder blocks and Goedel and Kimina 36; all read-outs have dimension 4,096. The archived quarter/middle/three-quarter layers are 6/14/21 (DeepSeek) and 8/17/26 (Goedel, Kimina); the middle layer is the principal evaluation. The formal-entry anchor is the last original token ending at the start of the formal proof body, allowing only trailing whitespace. Moving the anchor changes the first raw increment, and refitting the transform can change all whitened scores. The formal-entry analysis fixes calibration fold 0 and shrinkage 0.1; no evaluation labels enter the transform. Semantic replay and token alignment must agree on the first reproducible rejection; unresolved localization, parse failures and resource failures are kept distinct from localizable rejection. Kimina's processing removes terminal text after proof completion.

Table 7: Frozen model revisions and benchmark snapshots.

Artifact	Revision
DeepSeek-Prover-V2-7B	`a8d9e14432b2e8dd9df2a4d4e70f1ba9bc8d9b7b`
Goedel-Prover-V2-8B	`dfd02e6271a58375dfbf3ece0175277cf6b6a89a`
Kimina-Prover-Distill-8B	`74d328a7b1f001ab4871812582fc66d9bf70c68b`
miniF2F (Lean 4)	`70a1249ce240667f6bcdd1ccd62f847f0e065d57`
ProofNet-Verified	`160414332dc196583f6c37c310b420d2a3b07c58`

C Incremental-analysis protocol and supplementary results

Population, labels and features. The comparison retains attempts with $L \geq 2$, complete token boundaries and reliable semantic replay; ambiguous localization, parse errors and missing replay exclude the whole attempt, and no candidate step is dropped. The AST fixes candidate boundaries before execution, the diagnostic is mapped to a node, and prefix and rejection are reproduced in saved local contexts. A training label zero means "not the first rejection", not "verified correct", which differs from the absorbing labels of Section 3. The numeric features t, t/L, $\log(1 + n_{\text{tokens}})$ and mean surprisal use cubic B-splines with four training-quantile knots (linear fallback when knots coincide; constant extrapolation). Categories in fewer than ten training groups, and unseen ones, map to OTHER. No intercept or raw L is added, since attempt-constant scores cancel in the softmax; relative position still lets scores depend on length.

Objective and cross-validation. With G training groups, n_g attempts in group g and $\ell_{im} = \log \sum_{t \leq L_i} e^{r_{it}^{(m)}} - r_{it_i^*}^{(m)}$, each model minimizes

$$\mathcal{L}_m = \frac{1}{G} \sum_g \frac{1}{n_g} \sum_{i \in g} \ell_{im} + \frac{\lambda}{2} \|\eta_m\|_2^2, \tag{17}$$

with no division by L_i and no class reweighting; preprocessing row weights are $1/(G n_{g(i)} L_i)$. Five outer group folds; within each, three inner folds choose $\lambda \in \{10^{-4}, \ldots, 10\}$ (decades) by pooled inner-validation equal-group NLL, ties within 10^{-10} favoring the larger penalty. Scalers, knots and category maps are refitted on each training split. The chosen λ is 0.1 throughout except M1 in one Goedel fold ($\lambda = 1$). All 190 optimizations per prover converged with $\max|\nabla| < 2 \times 10^{-5}$. Metrics are $H_m = G^{-1} \sum_g n_g^{-1} \sum_{i \in g} \mathbf{1}\{\hat{t}_i^{(m)} = t_i^*\}$, $\Delta H = H_1 - H_0$ and $\Delta\mathrm{NLL} = G^{-1} \sum_g n_g^{-1} \sum_{i \in g} (\ell_{i0} - \ell_{i1})$; the ± 1 metric uses $\mathbf{1}\{|\hat{t} - t^*| \leq 1\}$.

Uncertainty and sample support. Twenty thousand bootstrap replicates sample whole groups, identical for both models, without refitting; 95% percentile intervals plus 97.5% intervals for the two-cell family. The 97.5% family intervals for ΔNLL are $[-0.0064, 0.0237]$ (DeepSeek) and $[-0.0967, -0.0180]$ (Goedel). Standardized jump coefficients by outer fold are $0.208, 0.171, 0.231, 0.200, 0.248$ (DeepSeek) and $0.211, 0.246, 0.418, 0.251, 0.111$ (Goedel). A fixed-fit calculation gives a group standard error of about 1.20 pp and roughly 209/167 groups for 80% power at 5% against the $+3$ pp target, compared with 166/133 available.

Table 8: Paired exact outcomes (attempt counts / equal-group shares) and mean absolute error.

Model	Both correct	M1 only	M0 only	Both miss	MAE M0→M1
DeepSeek	191 / 39.8%	14 / 3.7%	6 / 1.0%	241 / 55.5%	3.29 → 3.23
Goedel	148 / 55.7%	5 / 1.4%	6 / 2.1%	100 / 40.7%	3.01 → 3.10

Table 9: Descriptive length strata (ΔH, pp). Groups can appear in several strata. No eligible proof in this population first fails at its last step.

Model	Stratum	Attempts	Groups	ΔH
DeepSeek	$L = 2$–10	188	83	$+3.01$
DeepSeek	$L = 11$–24	148	92	0.00
DeepSeek	$L \geq 25$	116	68	$+0.86$
Goedel	$L = 2$–10	120	64	$+0.91$
Goedel	$L = 11$–24	76	53	-0.31
Goedel	$L \geq 25$	63	48	-1.39

Table 10: Examples selected by smallest SHA256 trace ID within each prover/outcome class.

Model	Outcome	Exercise	t^*/L	M0	M1
DeepSeek	helped	Dummit–Foote 1.1.34	7/25	1	7
DeepSeek	harmed	Artin 10.2.4	6/42	6	1
DeepSeek	unchanged	Rudin 5.5	4/50	9	9
Goedel	helped	Rudin 2.28	1/4	3	1
Goedel	harmed	Dummit–Foote 1.1.5	9/16	9	2
Goedel	unchanged	Artin 2.4.19	2/12	2	2

D TEMPERATURE-1.0 REPLICATION AND THRESHOLD DIAGNOSTICS

Temperature 1.0. Eligibility, the formal-entry transform (calibration fold 0, shrinkage 0.1), the argmax localizers and M0/M1 (refitted once on temperature-0.6 miniF2F failures) were frozen before evaluation. Seven one-sided hypotheses were tested with Holm correction: Ha, exact jump argmax above uniform (three provers); Hb, ± 1 jump argmax above ± 1 surprisal argmax (three provers); Hc, M1 above M0 for DeepSeek. Tests use one-sided cluster-t approximations over tasks; intervals use 20,000 task-bootstrap replicates. M0/M1 for Kimina were not fitted because only ten training groups were available.

For DeepSeek, Hc has Holm-adjusted $p = 0.479$.

Threshold diagnostics. Thresholds are quantiles of whitened increments on external verified calibration proofs. Verified controls are restricted to the tasks or exercise groups that contribute eligible failures; an earlier version with unpaired verified controls is retained in the archive as a sensitivity analysis. Cells with fewer than 30 exceedances or few verified groups are shown without intervals. The prefix analysis of Table 3 excludes attempts with $t^* = 1$.

Positional control for the prefix analysis. For each failed attempt with $t^* = k \geq 2$, the indicator $H(q, k) = \mathbf{1}\{\arg\max_{t \leq k} w_{q,t} = k\}$ is compared with the same indicator at a cutoff that is not an error. The positional control uses cutoffs $s < t^*$ of failed attempts, estimates the rate $r(k)$ for

Table 11: Temperature-1.0 miniF2F evaluation with frozen rules (%). p: Holm-adjusted over seven hypotheses.

Model	N/tasks	Jump	Uniform	Surprisal	Jump ±1	Surpr. ±1	p (Ha)	p (Hb)
DeepSeek	544/140	20.40	10.31	15.99	40.07	39.34	4.2×10^{-4}	0.547
Goedel	456/124	20.18	9.11	14.04	36.40	34.21	1.9×10^{-4}	0.547
Kimina	498/122	19.08	10.72	12.05	35.14	24.90	6.1×10^{-4}	0.0021

each prefix length k from all failed attempts with $t^* > k$, and standardizes it to the distribution of t^*; lengths without support are dropped (at most 0.7% of prefixes). Write $\mathcal{I}_{\text{supp}}$ for the retained attempts. The positional contrast is

$$\Delta_{\text{pos}} = 100 \, \frac{1}{|\mathcal{I}_{\text{supp}}|} \sum_{i \in \mathcal{I}_{\text{supp}}} \big[H(i, t_i^*) - r(t_i^*) \big] \quad \text{(pp)}.$$

The hit rate entering this contrast is evaluated on the same supported lengths as the control. Sensitivity analyses drop the first increment (candidates $t = 2, \ldots, k$, attempts with $t^* \geq 3$) and use equal group weights. Intervals use 20,000 joint bootstrap resamples of tasks or exercise groups, recomputing $r(k)$ in each replicate.

Table 12: Prefix-hit minus positional control, sensitivity analyses (pp; 95% intervals).

Dataset	Model	Main	No first increment	Equal groups
miniF2F	DeepSeek	+12.7 [6.8, 18.5]	+12.0 [5.7, 18.2]	+7.3 [2.1, 12.5]
miniF2F	Goedel	+14.4 [8.1, 20.8]	+13.4 [7.0, 19.9]	+11.8 [5.2, 18.6]
miniF2F	Kimina	+14.1 [7.1, 21.6]	+13.4 [5.9, 21.2]	+15.8 [8.7, 23.1]
ProofNet	DeepSeek	+8.2 [3.1, 13.4]	+8.9 [2.9, 15.2]	+8.9 [3.7, 14.2]
ProofNet	Goedel	+16.7 [10.0, 23.1]	+19.1 [11.5, 26.5]	+11.9 [4.5, 19.7]

By prefix length, the positional rate exceeds $1/k$ at short prefixes (68–92% against 50% at $k = 2$), so the positional trend is concentrated in the first few steps.

Table 13: Threshold diagnostics at all quantiles (%). Columns as in Table 4; intervals in the archive. †descriptive only.

Dataset	Model	q	$\mathbb{P}(M > u)$ failed	$\mathbb{P}(M > u)$ verified	$\mathbb{P}(T = F)$	$\mathbb{P}(F = t^*)$
miniF2F	DeepSeek	0.90	61.7	36.1	31.6	9.8
		0.95	44.1	25.0	48.9	15.0
		0.975	23.1	13.4	54.9	18.9
		0.99	9.7	8.1†	58.8	2.0
miniF2F	Goedel	0.90	63.4	42.0	32.4	14.4
		0.95	50.5	29.0	37.2	17.7
		0.975	38.4	20.8	50.6	20.4
		0.99	19.0	10.8	50.6	11.8
miniF2F	Kimina	0.90	69.1	47.0	28.1	11.6
		0.95	57.7	35.5	38.0	19.8
		0.975	48.8	27.2	41.1	23.4
		0.99	39.8	21.2	48.8	28.7
ProofNet	DeepSeek	0.90	90.8	88.6†	14.7	21.0
		0.95	84.7	62.9†	25.9	22.5
		0.975	73.7	42.9†	40.8	22.6
		0.99	52.6	22.9†	50.8	20.8
ProofNet	Goedel	0.90	93.1	85.0†	18.4	27.5
		0.95	87.8	75.0†	23.5	26.1
		0.975	79.8	67.5†	24.9	28.7
		0.99	65.7	35.0†	40.7	32.0

Online detection. Detection thresholds are set on inner out-of-fold verified attempts at target false-positive rates of 10% (primary) and 20%. Few independent verified groups are available, so realized rates on evaluation attempts depart from the targets.

E AGREEMENT CONTROLS AND SENSITIVITY

Variants and comparisons. This appendix reports the exact and the near rule; the region itself and a one-step margin around it; comparison with gap controls at matched coverage; comparison with length-standardized argmax references; comparison with length- and candidate-matched top-K references; and use of agreement as a detector of failed proofs.

Rules and matched coverage. The exact rule outputs J_i when $J_i = S_i$; the near rule fires when $|J_i - S_i| \leq 1$ and outputs the region $R = [\min(J_i, S_i), \max(J_i, S_i)]$. Precision and coverage are $\mathbb{P}(t^* \in R \mid \text{fire}, E)$ and $\mathbb{P}(\text{fire} \mid E)$. Table 14 compares the exact rule with gap controls that rank proofs by one signal's relative top-two gap $g(u) = (u_{(1)} - u_{(2)})/|u_{(1)}|$, select as many proofs and predict that signal's argmax (Geifman & El-Yaniv, 2019). For Goedel on ProofNet, agreement improves precision by 28.1 pp over the jump-gap control (99.5% interval $[2.4, 52.8]$) and by 26.3 pp over the surprisal-gap control ($[3.8, 45.2]$); for Goedel on miniF2F the jump-gap control is higher. These intervals use 50,000 task/group resamples with Bonferroni correction over ten contrasts.

Table 14: Exact localization when jump and surprisal maxima agree, and precision of gap controls at the same coverage (%). Intervals: nominal 95% task/group bootstrap. Kimina has 16 contributing tasks and is descriptive only.

		Agreement				Matched coverage	
Dataset	Model	Hits/agr.	Coverage	Precision	Interval	Jump gap	Surpr. gap
miniF2F	DeepSeek	36/68	12.9	52.9	[37.3, 66.2]	39.7	29.4
miniF2F	Goedel	22/45	10.0	48.9	[26.2, 66.7]	71.1	11.1
miniF2F	Kimina	5/19	5.9	26.3	–	26.3	21.1
ProofNet	DeepSeek	38/78	17.1	48.7	[36.2, 61.1]	33.3	46.2
ProofNet	Goedel	35/57	21.8	61.4	[46.3, 75.0]	33.3	35.1

Length- and candidate-matched references. Agreement selects short proofs (Table 15), where any localizer is more accurate. Each reference is therefore computed per selected proof and averaged. The random reference places a contiguous window of $K = |R|$ (or $|R \pm 1|$) steps uniformly in a proof of length L, with boundaries handled exactly. The single-signal reference gives the jump or the surprisal its K highest steps on all eligible failures of the same length L, including selected proofs, and reweights these accuracies to the joint (L, K) distribution of the selected proofs:

$$B_m = \frac{1}{A} \sum_{i:\ \text{selected}} \text{mean}_{q:\, L_q = L_i}\, \mathbf{1}\{t_q^* \in \text{top-}K_i \text{ of method } m \text{ on } q\}.$$

Each bootstrap replicate recomputes the per-length accuracies and weights. With $K = 1$ this is the ordinary argmax; with larger K the single signal receives the same candidate budget as the rule, without adjacency. Table 16 gives the results. With one candidate, three contrasts exclude zero: exact agreement over surprisal for DeepSeek on miniF2F and Goedel on ProofNet, and over jump for DeepSeek on ProofNet. For the near rule, DeepSeek on miniF2F exceeds both references and Goedel on miniF2F exceeds surprisal. With a one-step margin, no contrast excludes zero. Table 17 gives the descriptive comparison of exact agreement within one step with each single argmax within one step at the same length distribution.

Length strata and weighting. For the near rule (Table 18), the region contains the first rejection in 56.8–76.4% of selected proofs with $L \leq 10$, against 34.1–46.5% for a random window, and in 31.1–40.0% of those with $11 \leq L \leq 24$, about three times the 10.3–12.2% of a random window, on 19–45 selected proofs per cell; long proofs are rarely selected. For the exact rule, the corresponding short-proof rates are 55.6–72.1% against 22.5–31.0% (Kimina: 25.0% against 25.2%), and for $L \geq 25$ only 2–6 proofs per cell are selected. With equal group weights, exact-agreement precision is 35.6/36.6/23.8% on miniF2F and 48.2/57.8% on ProofNet.

Table 15: Selection and localization for the exact rule ($J = S$) and the near rule ($|J - S| \leq 1$) (%). Region $R = [\min(J, S), \max(J, S)]$ has one step for the exact rule and at most two for the near rule; $R \pm 1$ is clipped at the proof boundaries (at most four steps). $\bar{L}$: mean length of selected proofs.

		Exact rule					Near rule				
Dataset	Model	Sel.	Cov.	$\bar{L}$	$t^* \in R$	$t^* \in R \pm 1$	Sel.	Cov.	$\bar{L}$	$t^* \in R$	$t^* \in R \pm 1$
miniF2F	DeepSeek	68/528	12.9	11.5	52.9	67.6	169/528	32.0	11.5	59.8	72.2
miniF2F	Goedel	45/448	10.0	10.7	48.9	64.4	135/448	30.1	14.1	58.5	74.1
miniF2F	Kimina	19/324	5.9	14.7	26.3	47.4	74/324	22.8	11.3	44.6	51.4
ProofNet	DeepSeek	78/456	17.1	8.1	48.7	80.8	174/456	38.2	8.0	55.7	81.0
ProofNet	Goedel	57/262	21.8	6.8	61.4	87.7	118/262	45.0	8.8	60.2	78.0

Table 16: Agreement against references matched to the selected proofs' length L and candidate count K (%). Window: a uniformly placed contiguous window of K steps. Jump/Surprisal: accuracy of the signal's top-K steps on all eligible failures of the same length, reweighted to the selected (L, K) distribution. Δ: agreement minus reference, pp, with 99.875% intervals (Bonferroni over 40 contrasts; 50,000 paired task/group resamples); bold: interval excludes zero; (–): fewer than 20 contributing groups.

Dataset	Model	Rule	Region	Agree	Window	Jump	Surpr.	Δ vs jump	Δ vs surprisal
miniF2F	DeepSeek	exact	R	52.9	17.6	39.3	29.4	+13.7 [−0.8, 28.5]	**+23.5 [5.0, 40.1]**
		exact	$R \pm 1$	67.6	54.5	62.9	70.4	+4.7 [−10.8, 20.0]	−2.8 [−20.3, 11.4]
		near	R	59.8	30.3	47.5	44.3	**+12.3 [0.3, 26.4]**	**+15.5 [4.7, 24.9]**
		near	$R \pm 1$	72.2	57.4	66.7	73.8	+5.5 [−4.7, 19.0]	−1.6 [−14.3, 9.0]
miniF2F	Goedel	exact	R	48.9	16.9	45.0	25.6	+3.9 [−19.1, 22.1]	+23.3 [−3.9, 41.6]
		exact	$R \pm 1$	64.4	55.1	66.3	69.9	−1.9 [−27.6, 18.4]	−5.5 [−28.9, 13.3]
		near	R	58.5	30.4	49.1	41.1	+9.4 [−3.2, 20.8]	**+17.4 [4.9, 31.0]**
		near	$R \pm 1$	74.1	55.3	68.1	70.3	+6.0 [−5.7, 16.0]	+3.8 [−12.1, 21.1]
miniF2F	Kimina	exact	R	26.3	13.8	20.4	14.0	+5.9 (–)	+12.3 (–)
		exact	$R \pm 1$	47.4	30.9	45.6	36.3	+1.8 (–)	+11.1 (–)
		near	R	44.6	24.3	40.3	40.4	+4.3 [−30.4, 30.7]	+4.2 [−28.8, 29.6]
		near	$R \pm 1$	51.4	42.6	57.8	56.7	−6.4 [−37.4, 16.2]	−5.3 [−37.3, 22.3]
ProofNet	DeepSeek	exact	R	48.7	26.0	37.4	45.5	**+11.3 [1.1, 22.2]**	+3.2 [−12.1, 18.3]
		exact	$R \pm 1$	80.8	67.8	77.7	81.0	+3.0 [−7.4, 13.8]	−0.3 [−14.3, 14.7]
		near	R	55.7	38.5	51.4	53.5	+4.3 [−3.5, 12.2]	+2.2 [−7.3, 12.4]
		near	$R \pm 1$	81.0	68.2	77.3	79.8	+3.7 [−3.2, 11.6]	+1.2 [−9.6, 12.0]
ProofNet	Goedel	exact	R	61.4	25.2	47.9	36.0	+13.5 [−3.4, 28.8]	**+25.4 [7.7, 39.9]**
		exact	$R \pm 1$	87.7	64.2	78.2	76.1	+9.5 [−7.5, 22.3]	+11.6 [−3.6, 27.2]
		near	R	60.2	34.9	54.0	48.3	+6.2 [−3.8, 14.9]	+11.9 [−0.6, 21.2]
		near	$R \pm 1$	78.0	63.0	76.1	76.5	+1.9 [−8.7, 9.7]	+1.5 [−10.6, 11.9]

Detection. Treating a firing as a prediction that the proof failed, over eligible failed and verified proofs, gives the positive predictive values in Table 19. In all ten cells the excess over a length-only reference has an adjusted interval containing zero; none shows a statistically significant gain in detection precision beyond this reference. Counting a firing as correct only when its region contains the first rejection gives 1.6–46.7% of all firings.

Table 17: Exact agreement within one step versus each single argmax within one step, reweighted to the selected length distribution (%; descriptive).

Dataset	Model	Agreement ± 1	Jump argmax ± 1	Surprisal argmax ± 1
miniF2F	DeepSeek	67.6	61.0	57.6
miniF2F	Goedel	64.4	63.4	62.4
miniF2F	Kimina	47.4	44.5	28.7
ProofNet	DeepSeek	80.8	77.7	70.5
ProofNet	Goedel	87.7	76.0	69.2

Table 18: Near rule by length stratum (%, descriptive). Sel.: selected over eligible failures in the stratum; $t^* \in R$: localization; Win.: random window of the same size.

		$L \leq 10$			$11 \leq L \leq 24$			$L \geq 25$		
Dataset	Model	Sel.	$t^* \in R$	Win.	Sel.	$t^* \in R$	Win.	Sel.	$t^* \in R$	Win.
miniF2F	DeepSeek	110/174	76.4	41.3	45/202	31.1	11.6	14/152	21.4	3.8
miniF2F	Goedel	81/132	74.1	44.8	35/133	40.0	11.5	19/183	26.3	4.0
miniF2F	Kimina	44/92	56.8	34.1	22/113	31.8	11.3	8/119	12.5	5.6
ProofNet	DeepSeek	136/188	62.5	46.5	28/150	39.3	12.2	10/118	10.0	4.4
ProofNet	Goedel	90/120	67.8	43.1	19/77	36.8	10.3	9/65	33.3	5.3

Table 19: Agreement as a detector of failed proofs, over eligible failed and verified proofs (%). PPV: share of firings on failed proofs. Length-only: $\Pr(\text{failed} \mid L)$ averaged over the firings' lengths. Excess: PPV minus length-only, pp, with Bonferroni intervals over ten contrasts. Sens.: share of failed proofs selected; FA: share of verified proofs selected; Loc.: firings whose region R contains t^*, over all firings.

Dataset	Model	Rule	Firings	PPV	Length-only	Excess	Sens.	FA	Loc.
miniF2F	DeepSeek	exact	341	19.9	21.8	−1.9 [−7.0, 3.4]	12.9	35.2	10.6
		near	640	26.4	27.4	−1.0 [−4.2, 2.2]	32.0	60.8	15.8
miniF2F	Goedel	exact	219	20.5	18.5	+2.0 [−4.9, 9.3]	10.0	21.5	10.0
		near	536	25.2	23.1	+2.1 [−1.8, 5.7]	30.1	49.6	14.7
miniF2F	Kimina	exact	318	6.0	5.1	+0.9 [−1.7, 4.0]	5.9	34.4	1.6
		near	574	12.9	11.2	+1.7 [−1.5, 5.0]	22.8	57.6	5.7
ProofNet	DeepSeek	exact	114	68.4	71.0	−2.6 [−11.9, 6.1]	17.1	32.4	33.3
		near	242	71.9	70.9	+1.0 [−3.5, 5.5]	38.2	61.3	40.1
ProofNet	Goedel	exact	75	76.0	67.1	+8.9 [−2.4, 19.1]	21.8	22.8	46.7
		near	167	70.7	67.6	+3.1 [−1.6, 8.0]	45.0	62.0	42.5

F Details of unconfirmed diagnostic hypotheses

Archived tail comparisons. The older middle-layer, temperature-0.6 analysis uses the prompt-end anchor. With $M_k^{(r)} = k^{-1}\sum_{j\le k}[\log(Z_{(j)}/Z_{(k+1)})]^r$, the moment estimate is $\hat{\gamma}_k^M = M_k^{(1)} + 1 - \frac{1}{2}\big(1 - (M_k^{(1)})^2/M_k^{(2)}\big)^{-1}$ (Dekkers et al., 1989), and the GPD estimate fits shape and scale to excesses over $Z_{(k+1)}$ (Pickands, 1975). Hill is positive by construction; moment and GPD estimates keep their signs. Threshold selection is repeated under task resampling with the transform fixed.

Table 20: Archived tail estimates (prompt-end anchor). k: selected extremes; Tasks: distinct contributors; –: failed GPD fit.

Model	Subset	Hill	Moment	GPD	k	Tasks
DeepSeek	Verified	0.135	−0.062	−0.063	347	47
DeepSeek	Refuted	0.058	−0.317	−0.359	41	9
DeepSeek	$t < t^*$	0.116	−0.421	−0.326	237	36
DeepSeek	$t = t^*$	0.077	0.002	−0.013	20	13
DeepSeek	$t > t^*$	0.132	0.057	0.020	593	56
Goedel	Verified	0.171	0.168	0.172	1311	86
Goedel	Refuted	0.183	0.214	0.263	1253	79
Goedel	$t < t^*$	0.190	0.113	0.090	332	52
Goedel	$t = t^*$	0.204	0.030	–	22	16
Goedel	$t > t^*$	0.186	0.224	0.253	940	72
Kimina	Verified	0.149	0.135	0.146	773	64
Kimina	Refuted	0.089	−0.315	−0.482	36	16
Kimina	$t < t^*$	0.178	0.133	0.128	600	60
Kimina	$t = t^*$	0.054	0.092	0.147	20	10
Kimina	$t > t^*$	0.066	−0.036	–	20	9

Local peak prominence. For an interior earliest maximum at j, $R = w_j/((w_{j-1} + w_{j+1})/2)$. On tasks with both outcomes we regress $\log R$ on a refuted indicator, $\log L$ and $(\log L)^2$ with task fixed effects and equal task weights, resampling tasks 10,000 times with Bonferroni intervals across models.

Table 21: Length-adjusted outcome effect on local peak prominence, refuted versus verified.

Model	Paired tasks	Factor	98.33% interval
DeepSeek	44	0.998	[0.964, 1.040]
Goedel	54	0.976	[0.945, 1.013]
Kimina	42	1.017	[0.982, 1.054]

G Controlled deduction and training-dynamics protocols

Deduction. Calibration/evaluation successes are 171/1,000 and 84/500 at temperature 1.0 ($c = 0.26838$; Brier 0.126903 vs. 0.139785) and 188/1,000 and 91/500 at 0.6 ($c = 0.25541$). Syntax guidance restricts format only. The 2,000-replicate bootstrap resamples tasks within assigned length in both splits and refits the predictor. The experiment evaluates how success probability depends on assigned reasoning length.

Modular addition. One block, width 128, four heads, a 512-unit ReLU MLP, no layer normalization; modulus 113 with 3,830 training and 8,939 held-out pairs; full-batch AdamW (learning rate 10^{-3}, weight decay 1, $\beta = (0.9, 0.98)$) for 40,000 steps, checkpoints every 100 steps; five real-label seeds paired with shuffled-label controls. The transition is the first of five consecutive checkpoints with at least 90% test accuracy, and pre/post summaries use five checkpoints on either side. The primary 5% tail contains 638 of 12,769 normalized block-increment norms, a bounded finite input distribution rather than a sequence of proof steps.