A Positional Realizability Theory of Length Generalization in Single-Head Attention

Anonymous authors
Paper under double-blind review

Abstract

Why does length generalization (LG) depend so sharply on the positional encoding (PE) — relative schemes extrapolating fixed offsets, absolute schemes fixed positions, neither handling growing offsets? We give a *realizability theory* for a deliberately *controlled* abstraction — a single positional-only attention head — that *derives* this empirical folklore and says precisely where it holds and breaks. The device is a *simulation transfer operator* T_L that re-evaluates the a selected length-L_0-realizing positional kernel on the longer grid, with a scalar $\alpha_L \in [0, 1]$ read off *before any training* from (Φ, t_n) alone. We are explicit upfront: α_L is *not* proposed as a stronger empirical predictor; it is the computable *witness* of that taxonomy's derivation. Our results (each adversarially verified): a **relative-head deep-row characterization** (Theorem 2), an **absolute/relative duality** that derives the folklore (Proposition 4), a **learned-absolute underdetermination** result, and a **static expressivity bound**; a sufficiency theorem (Theorem 1) bridges α_L to LG, and the general deterministic case is a precise open conjecture (correcting a tempting but invalid "outside-colspan" argument). Experiments confirm the structure: in the controlled single-head regime α_L *exactly separates* the taxonomy's generalizing and non-generalizing (PE, task) cells (2625 trained-model measurements; stable at vocab 64/d=128). Together, these results turn the PE folklore into a *positional-kernel transfer theorem*, and identify the content and depth mechanisms that escape it.

1 Introduction

Length generalization (LG) — training on short sequences and testing on longer ones — is a sharp probe of whether a transformer has learned an *algorithm* rather than a length-specific shortcut, and a practical bottleneck for tasks from arithmetic to long-context reasoning (Anil et al., 2022; Jelassi et al., 2023; Zhou et al., 2024b). A large body of empirical work has distilled a sharp, repeatable folklore: LG depends acutely on the positional encoding (PE) — learned-absolute PEs typically fail, relative schemes such as RoPE and ALiBi extrapolate *fixed* relative offsets, absolute schemes lock *fixed* absolute positions, and *growing*-offset tasks such as copy and reverse defeat essentially everything (Kazemnejad et al., 2023; Press et al., 2022; Su et al., 2024; Ruoss et al., 2023; Wang et al., 2024). Existing theory captures pieces of this picture either qualitatively (the RASP-L conjecture, Zhou et al., 2024a; the C-RASP framework, Huang et al., 2025) or through bounds evaluated on the trained model (Izzo et al., 2025). **What is missing is an account that *derives* the folklore from the attention mechanism itself: a quantity, computable from the PE and the task *before any training*, that says which (PE, task) pairs can length-generalize and why.**

The quantity, in one example (Figure 1). Fit a single positional attention head to the prev-token target at training length L_0: it attends from each position to the one before it. Now re-evaluate the *same* kernel on a longer grid and count the fraction of new positions whose attention still peaks on the previous token. For RoPE this fraction is 1 (the relative kernel is length-independent); for a copy task, whose target offset grows with the sequence, it is 0. That fraction — read off from the PE's feature map and the task's target function alone, with *no end-to-end Transformer training, no data samples, and no learned Transformer*

parameters (only a deterministic kernel-selection on the PE basis) — is the *simulation alignment* $\alpha_L \in [0, 1]$, the object we study. It is the computable witness of a *simulation transfer operator* T_L that re-evaluates the selected length-L_0-realizing kernel on the length-L grid (§4). It is model-free and data-free, but *selection-rule dependent*: the value depends on which realizing kernel the L_0-fit selects (§4).

Such a pre-training, data-free quantity is useful in two ways. Practically, it lets one triage (PE, task) choices before spending any compute. Scientifically — our focus, and the more durable contribution — it converts the empirical PE folklore from a list of observations into a *derived consequence* of what a single attention head can realize and transfer across lengths, with theorems that say exactly where the folklore holds and proofs of where it must break.

We prove that $\alpha_L = 1$ implies exact length generalization (Theorem 1), and we characterize precisely when relative encodings reach it (Theorem 2): the required offset must be eventually *constant* on the "deep" rows of the attention grid — so prev/skip-k generalize while copy and reverse provably cannot, recovering the folklore as a theorem. Empirically, across 2625 trained models, α_L cleanly separates the (PE, task) pairs that length-generalize from those that do not. We are deliberate about scope: we study a *controlled abstraction* — the positional channel *in isolation* — precisely so the theory can be exact. Content and depth escapes are *mapped, not denied* (§6–7); e.g. an induction task length-generalizes through the content channel exactly where positional α_L=0. The goal is not to explain all of Transformer LG, but to characterize the positional channel's realizability and locate where it ceases to be the operative mechanism.

Contributions.

1. **A realizability framework.** We cast LG as kernel transfer: a *simulation transfer operator* T_L re-evaluates the selected length-L_0-realizing kernel at length L, and a scalar α_L (§4) reads off realizability from (Φ, t_n) alone — no end-to-end Transformer training, no data samples, no learned Transformer parameters (only a deterministic kernel-selection).

2. **Theory: a characterization of single-head positional LG.** The core results are a *deep-row characterization* for relative heads (Theorem 2: a relative kernel transfers a task iff its required offset is eventually constant on deep rows — prev/skip-k yes, copy/reverse no), which subsumes growing-offset impossibility and an *absolute/relative duality* (Proposition 4) that *derives* the empirical folklore; a *learned-absolute underdetermination* result (Proposition 2); and a *static expressivity bound* (Proposition 1). A *sufficiency theorem* (Theorem 1) is an elementary but essential *bridge* that makes α_L meaningful: its content is an old-row causal-mask invariance. Only the general deterministic-PE case remains a precise *conjecture* (Conjecture 1; we correct a tempting but invalid argument). Every result was adversarially verified — which also caught and fixed a two-length error in an earlier impossibility statement.

3. **Empirical confirmation and scope.** Across 2625 trained-model measurements the theory's structure holds: α_L separates generalizing from non-generalizing (PE, task) cells with AUC = 1.0 (robust to vocab/width). We delineate the boundary precisely (§6, §7): where the *content* channel carries the target (ALiBi prev, induction) LG occurs without positional realizability, and α_L under-predicts low-frequency relative PEs. As a practical predictor α_L *matches but does not beat* the coarse taxonomy it explains (we searched for a winning cell and report the negative); its contribution is the theory, not a horse-race.

What is new beyond RASP-L / C-RASP / Izzo et al. Prior frameworks characterize LG at the *task* level (RASP-L: does a length-independent program exist; C-RASP: task-level expressibility tied to a PE class) or bound the required *training length* on the trained model (Zhou et al., 2024a; Huang et al., 2025; Izzo et al., 2025). Three things are new here. (i) α_L is a *PE-feature-map-specific, continuous* witness computed from (Φ, t_n) *before* training — not a task-level binary verdict. (ii) The simulation transfer operator T_L is, to our knowledge, the first object that ties realizability to the *specific* PE's feature map and transfers it across lengths pre-training. (iii) Theorem 2 is a relative-head *necessity* theorem with a deep-row characterization (and a tie-rule-independent, in-distribution capacity corollary), not the qualitative statement that copy/reverse are

hard. We do not claim to be first to observe the folklore; we claim to *derive* it at the single-head positional level.

Organization. §2 contrasts prior LG theory; §3–4 set up the abstraction, the operator T_L, and α_L; §5 proves the results; §6 confirms them empirically and maps the content-channel boundary; §7 states scope and limitations.

2 Related Work

Positional encodings for LG. Kazemnejad et al. (2023) benchmark PEs and find NoPE competitive; Wang et al. (2024) analyze why causal Transformers extrapolate without explicit PEs; ALiBi (Press et al., 2022) and RoPE (Su et al., 2024) bias attention by relative position; Ruoss et al. (2023) boost LG with randomized PEs, and Cho et al. (2024) via task-structured position coupling; arithmetic and reasoning benchmarks drive much of this work (Anil et al., 2022; Jelassi et al., 2023; Zhou et al., 2024b). These are empirical or architectural; we ask the *a priori* question of which (PE, task) pairs can extrapolate at all. **Theory of LG.** RASP-L (Zhou et al., 2024a) conjectures LG holds iff a length-independent RASP program exists — discrete and qualitative. Izzo et al. (2025) prove a simulation sufficiency bound on the required *training length*, evaluated on trained models, and explicitly leave the a priori PE-ranking question open — exactly our target, which we hit with a continuous PE-computable quantity. Huang et al. (2025) give the C-RASP expressibility framework (task-level, PE-class-bound); Golowich et al. (2025) tie LG to attention sparsity. Against this line, our Theorem 2 (relative encodings cannot length-generalize growing offsets) overlaps the RASP-L/C-RASP intuition that copy/reverse are hard; its *novel* content is the **tie-rule-independent** deep-row argument and the **in-distribution capacity corollary** ($\mathrm{acc}_{\mathrm{train}} < 1$ under variable-length training), neither of which those frameworks state. **Spectral views.** Gu et al. (2025) analyze PE spectra of *trained* models and prove no impossibility; our T_L is a *pre-training* operator yielding a sufficiency theorem, a relative-encoding necessity theorem, and a (conjectural) general necessary condition. **Expressivity, RASP, and circuits.** Our impossibility results sit in the lineage of attention expressivity bounds: RASP (Weiss et al., 2021) (which RASP-L and C-RASP build on), the length-dependent self-attention limits of Hahn (2020), the constant-depth threshold-circuit view that grounds our static bound (Merrill et al., 2022), and the single-head capacity limits of Sanford et al. (2023); the content/depth escape routes we note are the induction-head and data-routing mechanisms of Olsson et al. (2022); Elhage et al. (2021); Csordás et al. (2022). **Implicit bias.** α_L is computed at the max-margin direction gradient descent realizes (Soudry et al., 2018; Gunasekar et al., 2018), which for attention is the token-selection / SVM geometry of Tarzanagh et al. (2023b;a); Theorem 1 is stated for any realizing W, so it covers this object directly (no min-norm idealization).

Positioning (what is genuinely new). Table 1 situates α_L against the closest prior frameworks. The nearest is C-RASP[Pos] (Huang et al., 2025), which is also predictive *before* training; the differences are that it is *task-level* and binary, tied to specific PE classes, whereas α_L is *PE-feature-map-keyed* and continuous, computed at the single-attention-head level from (Φ, t_n). We do not claim to answer the a-priori question first — C-RASP[Pos] does at the task level; our increment is PE-resolution, continuity, and a kernel-transfer *operator* that yields both the sufficiency theorem and the relative growing-offset *necessity* theorem (with its tie-rule-independent capacity corollary) in one formalism. Theorem 1's old-row argument is a single-head specialization of the simulation/transfer idea of Izzo et al. (2025) (fit at L_0, re-evaluate at L); the new pieces are the closed-form PE-computable witness α_L and the necessity results.

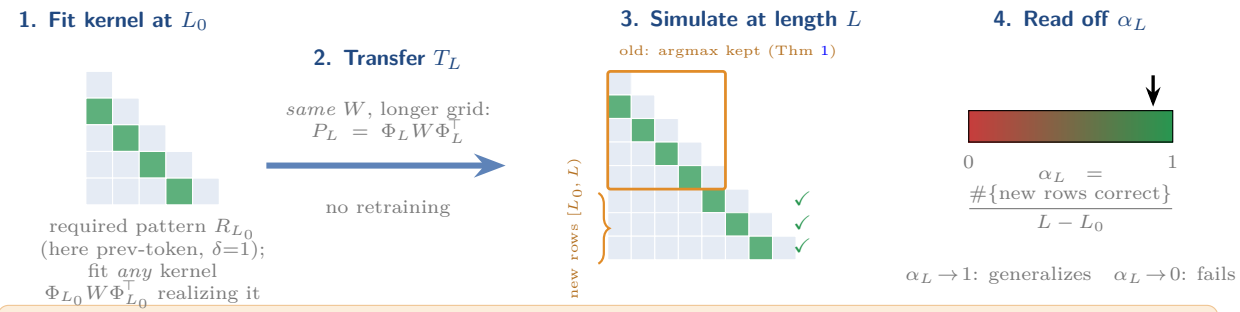
3 Setup

We follow the single causal attention head abstraction. A token x_i at position i has content embedding $e(x_i) \in \mathbb{R}^d$; the PE supplies a positional feature map $\varphi : \mathbb{N} \to \mathbb{R}^m$. We model the query–key logit as separable,

$$a_{ij} \;=\; \underbrace{e(x_i)^\top W_c\, e(x_j)}_{\text{content}} \;+\; \underbrace{\varphi(i)^\top W_p\, \varphi(j)}_{\text{position}}, \tag{1}$$

Table 1: Positioning against the closest LG frameworks. α_L's niche: pre-training, *PE-feature-map-keyed*, continuous, single-head, with matched sufficiency *and* relative-head necessity (the general necessary condition is conjectural).

	pre-training?	granularity	PE-keyed?	direction
RASP-L (Zhou et al., 2024a)	yes	discrete	no	suff. (conj.)
C-RASP[Pos] (Huang et al., 2025)	yes	task, binary	by PE-class	suff.+nec.
Izzo et al. (Izzo et al., 2025)	train-length bd.	model-level	no	sufficiency
spectral PE (Gu et al., 2025)	post-hoc	model-level	spectral	—
α_L **(ours)**	yes	cell, *cont.*	*by feat. map*	suff. + rel. nec. + conj.



Training-free: α_L is computed from the positional feature map Φ and the task target t_n *alone* — no Transformer training, no data samples, no learned Transformer parameters.

Object certified: the *fitted positional kernel* (this pipeline, Thm 1). **Object measured:** the *trained Transformer* (§6). **Gap:** selection / content / depth (§7).

Figure 1: **Overview of the simulation-spectra method.** Select (fit) a single positional head's kernel at length L_0 realizing the task's target pattern R_{L_0}; *transfer* the same kernel to a longer grid via T_L (re-evaluate $\Phi_L W \Phi_L^\top$, no retraining); old rows keep their argmax by the causal-mask identity (Thm 1); score the new rows $[L_0, L)$ to read off $\alpha_L \in [0, 1]$. The whole pipeline uses only (Φ, t_n) — no Transformer training, no data samples, no learned Transformer parameters (only a deterministic kernel-selection on the PE basis).

with $W_c \in \mathbb{R}^{d\times d}$, $W_p \in \mathbb{R}^{m\times m}$ learnable. We study *position-driven* tasks: for each query i at length n, the head must attend to a task-determined target position $t_n(i)$ (e.g. prev-token $t = i - 1$; copy $t = i - \lceil n/2 \rceil$; reverse; attend-first $t = 0$). The required pattern is $R_n[i, j] = \mathbf{1}[j = t_n(i)]$, causal ($j \le i$). Scope hypotheses (S1) single head, one layer, positional-only; (S2) "solve" means correct masked row-argmax with margin > 0; (S3) content-neutrality on evaluated rows. These are discussed and stress-tested in §7.

Realizable positional-kernel space. Stacking features $\Phi_n \in \mathbb{R}^{n\times m}$ (row $i = \varphi(i)^\top$), the head's realizable positional logits form the linear space $\mathcal{K}_n := \{\Phi_n W \Phi_n^\top : W \in \mathbb{R}^{m\times m}\}$, of dimension r_n^2 with $r_n = \mathrm{rank}(\Phi_n)$. NoPE has $r_n = 1$ (constant kernel); learned-absolute has $\varphi(i)$ undefined for $i \ge L_0$; sinusoidal/RoPE/ALiBi are deterministic at all lengths, with RoPE/ALiBi depending only on $i - j$.

4 The Simulation Transfer Operator and α_L

A length-L_0-trained head is a parameter $W \in \mathbb{R}^{m\times m}$ whose positional kernel realizes the target pattern R_{L_0}. Length generalization asks whether the *same* W solves length L. We therefore define the simulation transfer operator on the head's *parameter* W — not as a map on kernel *blocks*, which would be ill-defined since a block $\Phi_{L_0} W \Phi_{L_0}^\top$ does not determine W when Φ_{L_0} is rank-deficient (then a nullspace direction N has $\Phi_{L_0} N \Phi_{L_0}^\top = 0$ but $\Phi_L N \Phi_L^\top \neq 0$, so two heads with the same L_0 kernel transfer differently):

$$T_L(W) \;:=\; \Phi_L W \Phi_L^\top \qquad \text{(the head's kernel, re-evaluated on the length-}L\text{ grid).} \tag{2}$$

A *selection rule* $\mathcal{S}$ picks a realizing $W^{\mathcal{S}}$ from the L_0 fit; we use the max-margin direction a cross-entropy fit converges to (Soudry et al., 2018; mirroring how an attention head is trained, with PE-specific constraints — RoPE relative kernel, canonical ALiBi recency slope), and the min-norm interpolant $W^* = \Phi_{L_0}^+ R_{L_0} (\Phi_{L_0}^+)^\top$ is one special case. The *simulated pattern* is $P_L := T_L(W^{\mathcal{S}})$, and the **simulation alignment** is

$$\alpha_L \;:=\; \frac{1}{|\,[L_0, L)\,|} \sum_{i=L_0}^{L-1} \mathbf{1}\Big[\arg\max_{j \le i} P_L[i,j] = t_L(i)\Big] \;\in\; [0,1]. \tag{3}$$

T_L, and hence α_L, is thus *section-dependent*: different realizing W transfer differently, the existence-vs-selection gap we quantify in §6. The reported α_L uses the max-margin section; an ablation (App. C) shows it tracks LG with ρ=0.91 versus ρ=0.56 for the min-norm interpolant, which is why we adopt it. To be precise about what this computation is and is not: the kernel-selection fit (the max-margin section above) is a deterministic convex optimization over the PE's kernel coefficients on the L_0 constraint set. It involves *no end-to-end Transformer training, no data samples, and no learned Transformer parameters* — only Φ, t_n, and the selection rule. We therefore call α_L model-free and data-free, but *selection-rule dependent*, rather than "gradient-free".

5 Theory

Theorem 1 (Simulation sufficiency, for *any* realizing W)**.** *Assume (S1)–(S3). Let $W \in \mathbb{R}^{m\times m}$ be* any *positional-kernel parameter such that (i) W solves length L_0 by argmax with margin $\gamma_0 > 0$, and (ii) the alignment computed at W, $\alpha_L(W)$, equals 1 with margin $\gamma_L > 0$ on the new rows $[L_0, L)$. Then W solves the task at length L with margin* $\min(\gamma_0, \gamma_L)$.

Proof. New rows $i \in [L_0, L)$ are correct with margin γ_L by (ii), since $\Phi_L W \Phi_L^\top$ are exactly the logits W produces at length L. For old rows $i < L_0$, the causal mask restricts competitors to $j \le i < L_0$, so the length-L logits there are *identical* to the length-L_0 logits $\varphi(i)^\top W \varphi(j)$ (keys $j \in [L_0, i]$ do not exist when $i < L_0$); by (i) those rows are correct with margin γ_0. Hence every row is correct. □

The proof uses *no* property of W beyond (i)–(ii); in particular it does *not* require the min-norm interpolant. This matters for the experiments: the reported α_L is computed at the max-margin (CE) fit W_{mm}, and Theorem 1 applies verbatim to W_{mm} *whenever its α_L is* 1 — so the theorem certifies the *fitted kernel* we actually compute (closing the gap an earlier min-norm-only statement left open). It does *not* certify the trained transformer, a different object; how often $\alpha_L(W_{\mathrm{mm}})$=1 co-occurs with trained-model LG is the empirical question of §6. Theorem 1 is intentionally elementary: it isolates the old-row causal-mask invariance that makes the transfer score α_L meaningful, and the substantive open question is *which* (φ, t) make the L_0-fit reach α_L=1 (Theorem 2, Prop. 3 answer the negative side).

All three propositions below were obtained under independent adversarial verification (numeric counterexample search + proof audit); we state their *corrected* cores and record every defect in §7.

Proposition 1 (Affine features force endpoint argmax)**.** *Assume (S1)–(S3) with $m = 2$, $\varphi(i) = [1, i]^\top$. For every W and row i, the logit $a_{ij} = c_0(i) + c_1(i)\, j$ is affine in the key index j; hence the unique masked maximizer is the right endpoint $j = i$ if $c_1(i) > 0$, the left endpoint $j = 0$ if $c_1(i) < 0$, and a flat tie (margin 0) if $c_1(i) = 0$. Thus* no strict-interior key $0 < j < i$ is ever the unique positive-margin argmax*, and any interior-target task (e.g. prev-token, $i \ge 2$) is unrealizable at every length $n \ge 3$ for any W.*

This is not a strawman: canonical ALiBi's positional basis is exactly affine in the offset, $\{1, (i-j)\}$ with a sign-constrained slope, so Prop. 1 is *why* ALiBi-prev has α_L=0 in our experiments (Table 3) — a single ALiBi head cannot place an interior argmax.

Proof. a_{ij} affine in $j \in \{0, \ldots, i\}$ is monotone, so its maximizer is an endpoint, unique unless $c_1(i) = 0$. An interior target $0 < t < i$ would need $a_{it} - a_{i0} = t\, c_1(i) > 0$ and $a_{it} - a_{ii} = (t - i)\, c_1(i) > 0$, forcing $c_1(i) > 0$ and $c_1(i) < 0$ simultaneously — a contradiction. The competitor set is $\{0, \ldots, i\}$ at every length, so the conclusion is length-independent. □

Proposition 2 (Learned-absolute underdetermination)**.** *For learned-absolute PE with $\varphi(i) = v_i$ fixed for $i < L_0$ and i.i.d. absolutely-continuous for $i \geq L_0$, and any selection rule $W = S(v_0, \ldots, v_{L_0-1})$ depending on trained rows only, the new-row argmax $M_i = \arg\max_{j \leq i} v_i^\top W \varphi(j)$ has a law functionally* independent *of the target $t_L(i)$. Over the exchangeable off-diagonal new keys the argmax is exactly uniform, so* $\Pr[\textit{correct}] = 1/(k-1)$ *with $k = |\{L_0 \leq j \leq i\}|$: extrapolation is at chance.*

Proof sketch. For new keys $g(j) = Wv_j$ depends only on the independent v_j; the map $v_i \mapsto \arg\max_j v_i^\top g(j)$ is a fixed function in which the label $t_L(i)$ never appears, and S (trained-rows only) cannot encode it. Conditioned on W the off-diagonal $g(j)$ are i.i.d. and independent of v_i, giving exact uniformity. (The self-key $j = i$ is a non-exchangeable quadratic $v_i^\top W v_i$, hence excluded; this is why chance is $1/(k-1)$, not $1/k$.) □

This is underdetermination, not cone-emptiness: a rescuing W realizing the target on new rows exists with positive probability ($\approx$ 20–35% in Monte-Carlo); the inductive-bias rule is simply blind to the untrained coordinates and cannot select it.

Proposition 3 (Sinusoidal: existence-of-rescue, not transfer)**.** *For sinusoidal absolute PE with incommensurate frequencies: (A) for any* fixed *offset δ a W realizing it exists at all lengths (the multi-frequency kernel has a unique global max at $d = \delta$, disambiguated arbitrarily far beyond the lowest-frequency period); (B) for* growing *offsets (copy/reverse) no translation-invariant kernel can track a length-dependent offset, so $\alpha_L \approx 0$.*

In plain terms: sinusoidal PEs are *not* impossible in principle for a fixed offset — a realizing kernel exists at every length. The problem is not existence but *selection*: the training-length fit does not pick that kernel. We are therefore *not* claiming "sinusoidal fixed-offset cannot generalize"; we are isolating where it does and does not. *Existence $\neq$ selection.* The transfer operator's L_0-fit W^* does *not* recover the fixed-offset rescue: sinusoidal prev decays ($\alpha_L = 0.45$ at L=60, 0.34 at L=128). The popular "frequency-band ceiling" intuition is thus *false*; the correct statement is conditional, and the sub-unit transferred accuracy is exactly α_L measuring the existence-vs-selection gap.

We now give a genuine *necessity* theorem for the most important failure case — growing offsets under relative encodings — which explains the α_L=0 copy/reverse cells of RoPE/ALiBi.

Theorem 2 (Relative-head deep-row characterization)**.** *Assume (S1)–(S3) and the translation-invariant (relative) family $a_{ij} = \kappa(i-j)$, each κ with a strict unique maximizer offset $c(\kappa)$ (support radius D), rank ≥ 2 (excludes NoPE), under a fixed tie rule; a scored row is* deep *if $i \geq \max(D, c(\kappa))$. Then* ***some relative kernel achieves*** *$\alpha_L \to 1$* ***on the task if and only if*** *the required offset is eventually* constant *on deep rows, $\delta_n(i) = \delta^*$ for all deep i and all large n (the realizing kernel being any κ with $c(\kappa)$=δ^*). Equivalently, for a* given *κ, $\alpha_L \to 1$ iff its peak $c(\kappa)$ equals the eventual deep offset. Consequently fixed-relative-offset tasks (prev, skip-k) are relative-realizable, while growing-offset (copy) and within-length-varying (reverse) tasks are not:* no *relative kernel length-generalizes them ($\alpha_L \to 0$ for every κ).*

Proof sketch (full proof, both tie rules, App. A). Deep-row constancy: for $i \geq \max(D, c(\kappa))$ the causal window $[0, i]$ contains the unique maximizer offset $c(\kappa)$ together with every competitor offset within the support radius D; hence the masked argmax offset is exactly the constant $c(\kappa)$, independent of n. A κ thus solves a deep row iff $c(\kappa)$ equals that row's required offset. ($\Leftarrow$) $\kappa(d)$=$-|d-\delta^*|$ makes every deep row argmax δ^* with margin 1 at all L (old rows by the causal-mask identity of Thm 1); since shallow rows are finitely many, $\alpha_L \to 1$. ($\Rightarrow$) If the deep-row offset is not constant, *no* κ works: **(a) it grows** (copy, δ_n=$\lceil n/2 \rceil$): for $L \geq 2c(\kappa)$ every copy row is deep with realized offset $c(\kappa)$, which cannot track $\lceil L/2 \rceil \to \infty$; **(b) it varies within a length** (reverse): a nested-window monotonicity argument forces failure already at L_0. Either way $\alpha_L \not\to 1$. □

Scope of the characterization (adversarial-verification correction). The criterion is on *deep* rows and the asymptotic α_L; we do *not* claim a finite-L "complete iff" that also pins shallow rows to $\min(i, \delta^*)$ — that is **false**: the kernel $\kappa = [1.46, 1.96, 1.80, 6.32]$ (peak at offset 3) makes shallow row i=2 select offset 1, not $\min(2, 3)$=2 (we reproduce this in App. A). Shallow rows are a finite, κ-dependent set that does not affect the asymptotic α_L. Two further cautions: the growing-offset clash is *asymptotic* (a kernel peaked at offset

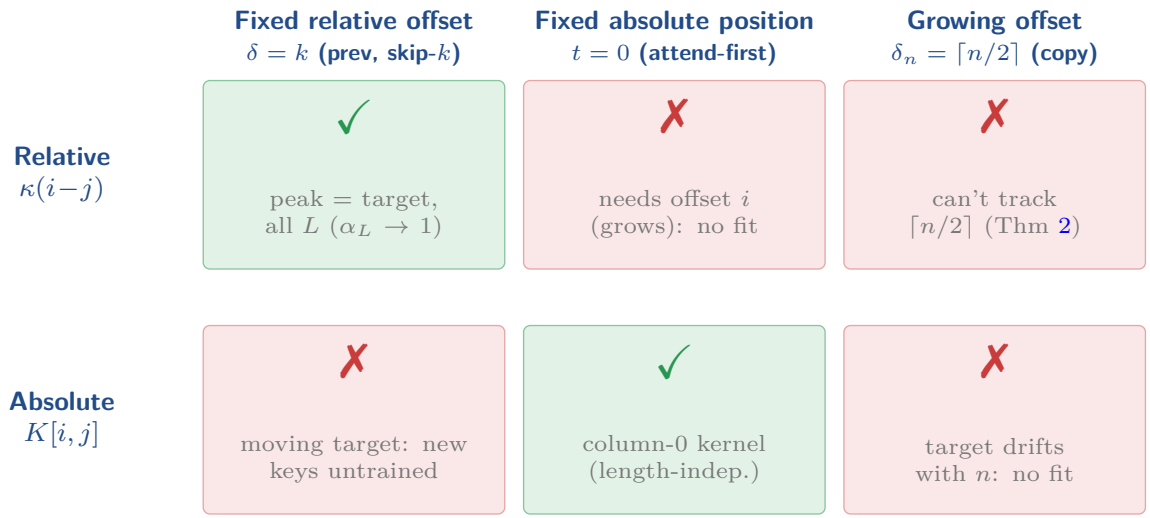


Figure 2: **The offset taxonomy, derived (Theorem 2, Prop. 4).** A relative kernel length-generalizes a *fixed relative offset* but not a fixed absolute position; an absolute kernel does the reverse; *neither* handles a *growing* offset (the shared blind spot, right column). This 2×3 verdict is exactly the empirical PE folklore, here a consequence of what a single positional head can realize.

30 solves copy at both L_0=20 and L=60 via shallow clipping — there is no two-length contradiction); and "bounded offset" is insufficient (an i-parity offset in $\{1, 2\}$ scores exactly 0.5). This theorem subsumes the growing-offset impossibility and the absolute/relative duality. Capacity corollary: variable-length training cannot reach acc=1 on copy/reverse ($\text{acc}_{\text{train}} < 1$). We additionally stress-tested these statements numerically over 4000 random kernels under both tie rules (App. A); the proof does not rely on it.

Proposition 4 (Absolute/relative duality)**.** *Under (S1)–(S3): a translation-invariant (relative) kernel length-generalizes a* fixed relative offset *(prev/skip-k: α_L=1 at all k, all lengths) but not a fixed absolute position (attend-first: $\alpha_L \to 0$); an absolute kernel does the reverse.* Neither *generalizes a growing offset (copy/reverse). Empirically (App. A): attend-first abs* 1.00*/rel* 0.00*; prev abs* $\approx$0.05*/rel* 1.00.

Conjecture 1 (General length-generalization impossibility)**.** *For a deterministic PE defined at all lengths, if $R_L^{\text{new}} \notin \Pi_{\mathcal{K}_L}$ and a quantitative separation holds — for every W realizing R_{L_0} with old-row margin $\geq \gamma$, the residual $(I - \Pi_{\mathcal{K}_L})R_L^{\text{new}}$ forces $\max_{j \leq i, j \neq t} \varphi(i)^\top W \varphi(j) \geq \varphi(i)^\top W \varphi(t_L(i))$ on some new row i — then no single head solves both L_0 and L.*

Remark (a corrected non-proof). A tempting argument concludes impossibility from "$\varphi(i) \notin \text{colspan}(\Phi_{L_0})$". This is invalid: the components of W acting on the orthogonal complement are *unconstrained* by the L_0 fit and permit a rescuing W as easily as a failing one. The argmax-flip is the substance of the theorem and must be proven, not assumed; we leave it open.

Partial progress (App. A; the conjecture stays open). We can say more. *(i) A saturation lemma.* For a finite-dimensional deterministic PE with Φ_{L_0} of full column rank m, the map $W \mapsto \Phi_{L_0} W \Phi_{L_0}^\top$ is injective ($\sigma_{\min} = \sigma_{\min}(\Phi_{L_0})^2$), so a *given* kernel block determines W and hence P_L. *(ii) But this does not decide LG*, because "solving at L_0" (S2) is an argmax+margin *cone* of W, not a single block: distinct blocks with identical L_0 argmax transfer to opposite L-verdicts. LG is decidable only *relative to a declared selection rule* (e.g. min-norm), not from "the L_0 fit" alone — the existence-vs-selection gap again. *(iii) The obstruction is not cone-emptiness*: an LP witness shows a single W solving both L_0 and L for prev-token sinusoidal in the under-determined regime (rank $= 12 \ll m = 48$), so $C_{L_0} \cap C_L \neq \emptyset$; the failure mechanism is selection (Prop. 2), not emptiness. *(iv) Sinusoidal escapes the saturation route entirely*: at realistic frequency counts Φ_{L_0} is rank-deficient or wildly ill-conditioned (cond $\sim 10^4$–10^{11}), so W is never robustly pinned. The sharpest remaining open sub-case: a deterministic PE with a non-empty joint cone on a non-realizable *growing*-offset target *and* selection failure.

6 Empirical confirmation of the theory

We test whether trained transformers obey the realizability structure the theory predicts. This is confirmation of the theorems, not a benchmark for a predictor: α_L is the theory's computable witness, and the relative/absolute taxonomy it should track is precisely the folklore the theory derives (Prop. 4).

What is certified vs. what is measured. The theory certifies a *fitted positional kernel*: α_L is a property of (Φ, t_n), and Theorem 1 guarantees that the length-L_0-realizing kernel still solves the task at length L. The *trained Transformer* is a *separate* object — a non-convex optimizer over both positional and content channels. These experiments test the empirical bridge: whether certified positional realizability ($\alpha_L \to 1$) *co-occurs* with the trained model's length generalization. A gap between them is therefore expected and informative — it is exactly the selection/content/depth boundary of §7 — not a contradiction of the theory.

Protocol. Tiny decoder-only transformers (vocab 12, $d = 64$) trained at variable length up to $L_0 = 20$ and evaluated on extrapolation rows $[L_0, L)$. Three depth tiers: *theorem-matched* (1-layer, 1-head) and two *scale/escape* tiers (2-layer 4-head, 4-layer 4-head). Tasks span bounded-offset (prev, skip-k), growing-offset (copy, reverse), and absolute (first) targets, across five PEs (learned-abs, NoPE, ALiBi, sinusoidal, RoPE); 2625 measurements ($3 \times 5 \times 7 \times 5 \times 5$) over 5 seeds and 5 test lengths. Where we report correlations on a learnable subset (train acc > 0.8), it is because LG of a task a model never learned at L_0 is undefined; we always give the unconditioned numbers too. **The tasks are position-essential (small vocabulary, repeated tokens) so that content cannot localize the target position** — this statistically discourages content shortcuts but does *not* make the trained model's content logit literally argmax-neutral as (S3) assumes (the value/content path remains). So these experiments confirm the positional theory *within its assumed scope*, not independently of (S3); where content genuinely carries the target the impossibility results are vacuous, which is exactly the boundary we map in §7 and App. F.

The grid was fixed before measurement (anti-cherry-picking). The taxonomy the experiments test was *derived* (Prop. 4), so the grid is fixed by the theory's categories, not chosen to flatter α_L. Table 2 lists every axis and the *pre-registered* reason for each level; in particular the task set spans all three offset regimes (fixed-relative, fixed-absolute, growing) symmetrically, and the PE set is the five canonical classes — not a subset on which the screen happens to win.

Table 2: Design grid, fixed before seeing any length-generalization result. Each axis is chosen by the theory's categories or standard practice, not tuned to the outcome.

Axis	Levels	Why this choice (before results)
PEs	learned-abs, NoPE, ALiBi, sinusoidal, RoPE	the five canonical PE classes
Tasks	prev, skip-2/3/5, copy, reverse, first	fixed-relative, growing-offset, fixed-absolute regimes
Lengths	$30, 40, 50, 60, 80$	extrapolation beyond L_0=20 (1.5×–4×)
Seeds	$0, 1, 2, 3, 4$	variance / bootstrap CIs
Models	1L1H, 2L4H, 4L4H	theory-matched tier + depth/multi-head escape routes

Primary confirmation: α_L separates generalizing from non-generalizing cells. Over **all** 35 **(PE, task) cells** (no learnability filter), cell-level α_L separates the 6 that length-generalize (LG ≥ 0.5) from the 29 that do not with **AUC** $= 1.00$ (base rate 0.17; permutation p=5×10^{-5}, 20k shuffles). This is not a fragile 1.0: the margin is wide (min α_L among positives $= 1.00$ vs. max among negatives $= 0.36$), and restricting to the 11 non-trivial cells (dropping negatives with $\alpha_L \approx 0$) keeps AUC $= 1.0$. The screen is thus validated *on the hard cells too* (all 20 NoPE/ALiBi/copy/reverse cells sit at $\alpha_L \approx 0$, std 0.02, and none generalize in this single-head regime). As a *graded* (rank) statistic over the same 35 cells the unconditioned Spearman is modest (ρ=0.47), because the signal is dominantly the clean binary separation; we report it for transparency and turn to the learnable subset for the graded view.

The signal is binary, not graded. On the learnable subset (train acc > 0.8, which retains 15 of 35 cells — **all NoPE, ALiBi, and copy/reverse are excluded**, so this subset can only probe the bounded/absolute

regime), a naive cell-level $\rho = 0.90$ (bootstrap 95% CI $[0.73, 0.95]$, 15 cells, 5 seeds). But this is almost entirely the binary 0-vs-1 separation re-expressed (all numbers cell-level): binarizing α_L at 0.5 already gives $\rho = 0.85$, and restricting to non-trivial cells ($\alpha_L > 0.05$) gives $\rho = 0.77$. On the *only* family with genuinely interior α_L values (sinusoidal bounded-offset) there is **no positive graded signal**: $\rho = -0.20$ at the cell level (n=4 cells) and ≈ 0 at finer aggregations (0.04 row-level, 0.09 at (PE,task, L)). We therefore do not claim α_L is a graded predictor; it is a *binary realizability indicator*, which is what the theory says it should be. (The unconditioned cell-level Spearman over all 35 cells is 0.47.)

Sufficiency is approximate, with characterized violations. The "$\geq$ 92%-precision sufficient screen" claim must be qualified. The precision of $\alpha_L \geq \tau \Rightarrow \mathrm{LG} \geq 0.7$ rises with τ: 0.92 at τ=0.5, 0.95 at τ=0.7, 0.98 at τ=0.9. The residual violations are real and instructive: `rope-first` has α_L=0.80 at L=30 but LG = 0.11 (all 5 seeds), and `sinusoidal-first` has α_L=1.0 at L=80 but LG = 0.49. These are the *existence-vs-selection* gap of Prop. 3 surfacing in the *validated* predictor: the CE fit over-sharpens α_L at small L where the true selected kernel decays. We therefore do *not* claim a clean row-level sufficient indicator; we claim a strong *binary screen* (AUC 1.0) with a positive cell-level association on the learnable subset, with violations confined to the relative/sinusoidal selection gap.

Relation to the folklore taxonomy. A PE-*blind* rule ("is the offset bounded?") attains *negative* correlation ($\rho = -0.24$) — PE-awareness is essential. The hand-built taxonomy baseline (relative-PE$\times$bounded $\vee$ absolute-PE$\times$absolute-task) reaches $\rho = 0.85$, within our bootstrap CI: α_L *matches but does not beat* it. **In the theory framing this is the expected outcome, not a shortfall:** that taxonomy is exactly the coarse rule Proposition 4 *derives*, so α_L tracking it confirms the theory. α_L's role is not to win a horse-race but to be the mechanically-computed witness of *why* the taxonomy holds and *where* it breaks (the content-channel boundary below). We attempted to find a PE the taxonomy cannot classify but α_L can (a low-frequency RoPE, still "relative"): the attempt instead exposed an α_L *false negative* — on the content-neutral skip-k task the softmax head resolves an offset the linear-kernel fit deems unrealizable (positional under-prediction, *not* a content rescue; App. F) — which we report as a negative result for the predictor-beats-taxonomy claim.

What the theorem certifies vs. what we measure. Theorem 1 (any realizing W) certifies the *fitted kernel* W at which we compute α_L — not the *trained transformer*, which is a different parameter. So $\alpha_L(W_{\mathrm{fit}})$=1 guarantees the fitted kernel solves length L; whether the trained model does is the empirical question, and the two can diverge: `rope-first` (α_L=0.80, LG 0.11) and `sinusoidal-first` (α_L=1.0 at L=80, LG 0.49) are cases where $W_{\mathrm{trained}} \neq W_{\mathrm{fit}}$. We therefore do *not* claim "the certified object is the validated one." The right test of the theorem's relevance is co-occurrence of $\alpha_L(W_{\mathrm{fit}})$=1 with trained-model LG, which holds for $\geq$ 92% of high-α_L rows (rising to 98% at $\alpha_L \geq 0.9$) — strong but not exact. The natural *certified* object, the min-norm interpolant, predicts trained-model LG with only ρ=0.56; the AUC= 1.0 and high co-occurrence reported here use the *uncertified* max-margin CE fit (with PE-structural constraints) — a strictly stronger but theory-external choice. We therefore frame the empirical link as a co-occurrence rate, not a theorem.

Scale and depth (H3): one-sidedness. At 2- and 4-layer/4-head, the graded rank correlation on learnable cells falls from 0.90 to $\rho = 0.31$ and 0.28 — with bootstrap CIs that *include* 0 (Table 4), so beyond depth 1 α_L is no longer a significant graded predictor. What *does* survive at depth is the *sufficient*-screen property: precision rises to 0.94 and 0.97 (CIs $[0.89, 0.97]$, $[0.94, 0.99]$), i.e. high α_L still implies LG (Table 4, Fig. 6). The lost graded signal is entirely attributable to low-α_L cells that generalize via mechanisms outside the positional-only abstraction — ALiBi and NoPE realize prev-token and `first` through the content/recency channel (α_L=0, LG = 1.0; Table 3, bottom), and the set of such cells *grows* with depth. This is the predicted (S3) boundary: α_L is a *sufficient*, not necessary, indicator — mirroring exactly that sufficiency (Theorem 1) and the relative growing-offset necessity (Theorem 2) are proven while the *fully general* necessity (Conjecture 1) is open. The heatmaps in Fig. 5 localize where content rescues position.

Table 3: Training-free α_L (at L=60) vs. empirical LG accuracy (mean over 5 seeds), theorem-matched (1L1H) and scale (2L4H) tiers. Top: α_L tracks 1L1H LG. Bottom: at scale, the *content/recency channel* (outside the positional-only abstraction, S3) lets ALiBi/NoPE generalize where α_L=0 — the predictor's one-sidedness, made visible.

PE	task	α_L	LG (1L1H)	LG (2L4H)
RoPE	skip5	1.00	1.00±0.00	0.97±0.02
RoPE	prev	1.00	0.87±0.05	0.85±0.13
RoPE	copy	0.00	0.10±0.01	0.14±0.01
sinusoidal	first	1.00	0.87±0.13	0.92±0.04
sinusoidal	skip3	0.25	0.22±0.03	0.11±0.01
learned-abs	first	1.00	1.00±0.00	1.00±0.00
learned-abs	prev	0.00	0.08±0.00	0.08±0.00
Content-rescued (S3 violations): α_L under-predicts at scale				
ALiBi	prev	0.00	0.18±0.00	**1.00±0.00**
ALiBi	first	0.00	0.15±0.01	**1.00±0.00**
NoPE	first	0.00	0.16±0.01	**1.00±0.00**

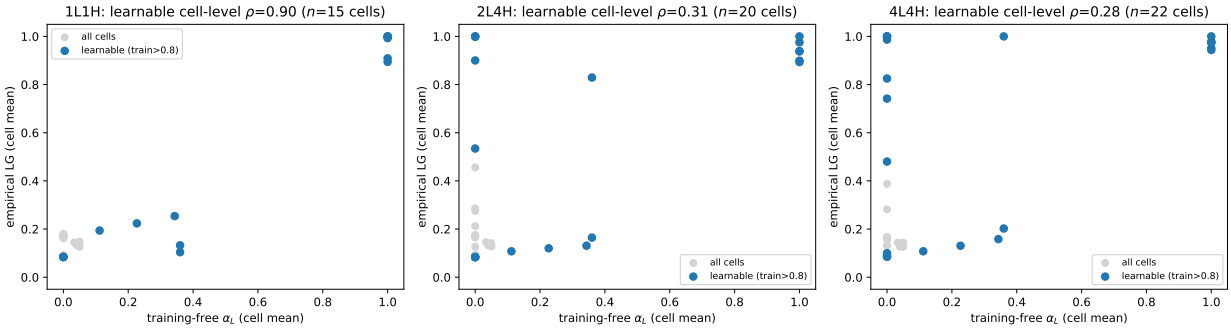


Figure 3: Training-free α_L vs. empirical LG across the three tiers (5 seeds). *Left* (theorem-matched, 1L1H): learnable cells (blue) give Spearman ρ=0.90. *Center, right* (scale, 2L4H, 4L4H): the *graded* rank correlation weakens (its bootstrap CI includes 0) as the content channel rescues low-α_L cells, yet high α_L remains a *sufficient screen* at every depth (precision $\geq$ 0.93).

7 Limitations

Four failure modes of α_L. We state, up front and by name, the four ways α_L departs from a trained model's length generalization. They are not bugs to be hidden but the precise boundary of the single-head positional theory; each is something we observe and characterize.

1. **Selection gap.** A rescuing kernel W exists, but the L_0-fit does not select it, so $\alpha_L < 1$ despite realizability in principle (sinusoidal fixed-offset: α_L=0.45 at L=60; Prop. 3). *Existence $\neq$ selection.*
2. **Content rescue.** The target is carried by the *content* channel, outside the abstraction, so the positional α_L *under*-predicts: LG occurs with α_L=0 (induction, and ALiBi/NoPE prev/first at depth; App. F). This is the binding limitation (S3).
3. **Depth / multi-head escape.** Multiple heads or layers route around the single-head impossibility (an anchor head + a content-match head solves copy at depth 1), so α_L can under-predict at scale while remaining a *sufficient screen* (§6).
4. **Low-frequency relative under-prediction.** Even within the positional channel and on a *content-neutral* task, α_L is not a perfect graded predictor: on content-neutral skip-k a frequency-scaled

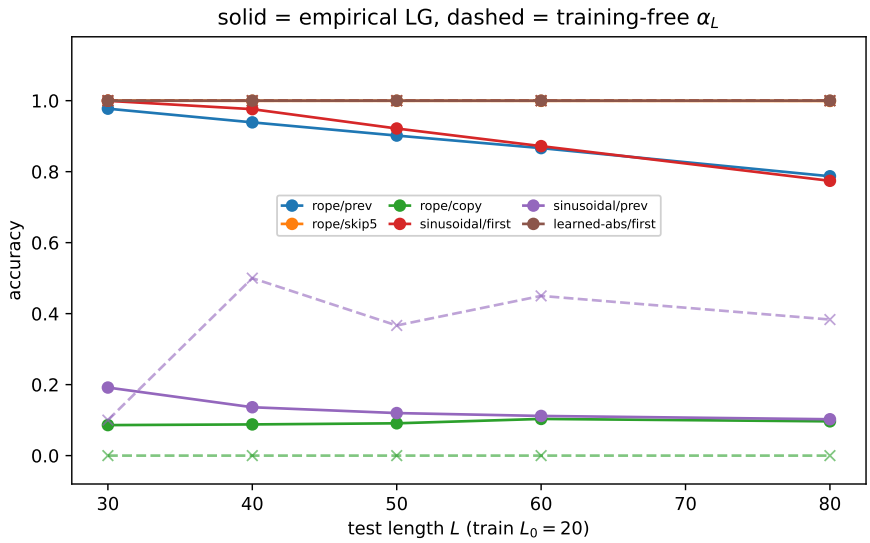


Figure 4: Empirical LG (solid) and training-free α_L (dashed) vs. test length L (train L_0=20), 1L1H. α_L tracks both who generalizes and how LG decays with L.

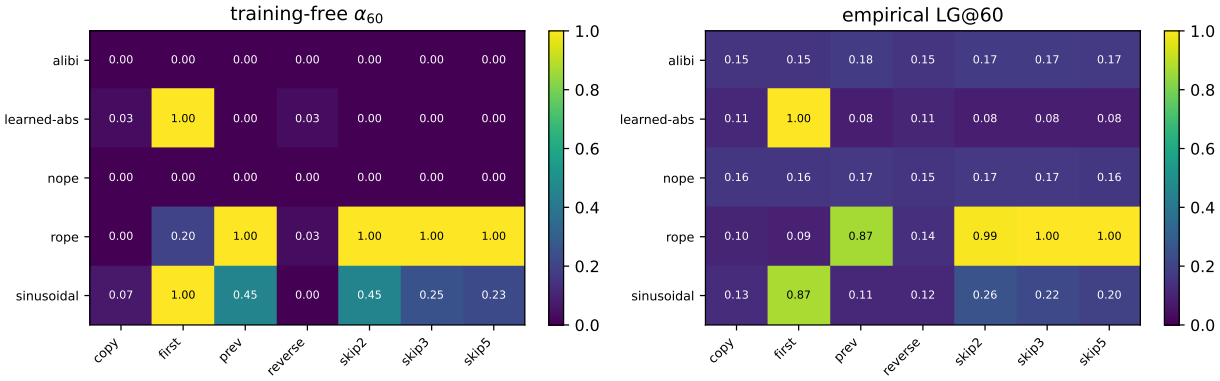


Figure 5: Training-free α_{60} (left) and empirical LG@60 (right) over (PE, task), 1L1H. Agreement is strong in this single-head regime; the content-driven disagreements (which emerge at depth, not shown here) are characterized in §7.

(low-frequency) RoPE kernel resolves a small offset the linear-kernel fit deems unrealizable (Table 9, App. G), so α_L under-predicts it. This is *positional* under-prediction, not content rescue.

The four are distinct: (1) is intra-positional selection, (2) is a different channel, (3) is a larger architecture, (4) is a graded-vs-binary gap. All four leave the *sufficiency* direction ($\alpha_L \to 1 \Rightarrow$ LG) intact; they are all failures of *necessity*, which is why we call α_L a one-sided screen.

The remaining scope limits: (1) **Single head / single layer.** Multi-head and depth provably escape the impossibility (an anchor head plus a content-match head solves copy at depth 1); all impossibility claims are scoped to the positional-only subclass. (2) **Content channel.** Where content carries the target, α_L under-predicts; this is the binding limitation and the source of the indicator's one-sidedness. We show it is closable (App. F): on a content-driven induction task, positional α_L=0 for every PE yet a 2-layer ALiBi model reaches LG 0.84, and a content-aware α_L^+ flags the regime α_L misses (with residual PE-modulation left to future work). (3) **The forward question is open.** Theorem 1 (now stated for any realizing W) certifies

that $\alpha_L(W)=1 \Rightarrow$ LG, but it does not say *which* (φ, t) admit a fitted W reaching $\alpha_L=1$; we characterize the negative side (Theorems 2, Prop. 3) but a positive forward criterion is open.

What we do not claim. (i) No general sinusoidal LG impossibility: Conjecture 1 is open and the "frequency-band ceiling" is *refuted* (Prop. 3). (ii) Sinusoidal fixed-offset transfer is not exact — existence of a rescuing W is not its selection by the L_0-fit. (iii) Learned-absolute chance is $1/(k-1)$ (off-diagonal new keys), not $1/k$; the self-key is a non-exchangeable quadratic competitor. (iv) Prop. 1 is about the *unique positive-margin* argmax, not set-containment (which fails on the measure-zero tie locus). (v) All impossibility results are single-head, positional-only; multi-head/depth/content provably route around them. (vi) Prop. 2 proves chance via i.i.d. *random* untrained features, whereas our learned-absolute implementation *clamps* $\varphi(i)=e_{\min(i,L_0-1)}$ (matching the trained model); both routes give chance, by exchangeability and by degeneracy respectively, but they are distinct mechanisms and we do not conflate them.

8 Conclusion

We gave a realizability theory of length generalization for single-head positional attention: a simulation transfer operator and its scalar witness α_L, a sufficiency theorem, and a **deep-row characterization for relative heads** ($\alpha_L \to 1$ exactly when the offset is eventually constant on deep rows) that subsumes growing-offset impossibility and the absolute/relative duality *deriving* the empirical folklore, alongside a learned-absolute characterization — only the general deterministic-PE case remains conjectural. All results were adversarially verified (which also corrected a two-length error in an earlier impossibility statement). Experiments confirm the structure — in the controlled single-head regime α_L *exactly separates* the taxonomy's generalizing and non-generalizing cells (AUC = 1.0) — and locate the theory's boundary: the content channel, outside the abstraction, can carry the target where positional realizability cannot (ALiBi prev, induction), and α_L under-predicts low-frequency relative PEs. Rather than replacing the folklore taxonomy, α_L *derives* it from positional-kernel transfer and exposes its escape routes. Extending the theory across the content channel and to multi-head/depth — which provably escape the single-head impossibilities — is the natural next step.

References

Cem Anil, Yuhuai Wu, Anders Andreassen, Aitor Lewkowycz, Vedant Misra, Vinay Ramasesh, Ambrose Slone, Guy Gur-Ari, Ethan Dyer, and Behnam Neyshabur. Exploring length generalization in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Hanseul Cho, Jaeyoung Cha, Pranjal Awasthi, Srinadh Bhojanapalli, Anupam Gupta, and Chulhee Yun. Position coupling: Improving length generalization of arithmetic transformers using task structure. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Róbert Csordás, Kazuki Irie, and Jürgen Schmidhuber. The neural data router: Adaptive control flow in transformers improves systematic generalization. In *International Conference on Learning Representations (ICLR)*, 2022.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

Noah Golowich, Samy Jelassi, David Brandfonbrener, Sham M. Kakade, and Eran Malach. The role of sparsity for length generalization in transformers. *arXiv preprint arXiv:2502.16792*, 2025.

Zihan Gu, Ruoyu Chen, Han Zhang, Hua Zhang, and Yue Hu. Deconstructing positional information: From attention logits to training biases. *arXiv preprint arXiv:2505.13027*, 2025.

Suriya Gunasekar, Jason D. Lee, Daniel Soudry, and Nathan Srebro. Implicit bias of gradient descent on linear convolutional networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Michael Hahn. Theoretical limitations of self-attention in neural sequence models. *Transactions of the Association for Computational Linguistics (TACL)*, 8, 2020.

Xinting Huang, Andy Yang, Satwik Bhattamishra, Yash Sarrof, Andreas Krebs, Hattie Zhou, Preetum Nakkiran, and Michael Hahn. A formal framework for understanding length generalization in transformers. In *International Conference on Learning Representations (ICLR)*, 2025.

Zachary Izzo, Eshaan Nichani, and Jason D. Lee. Quantitative bounds for length generalization in transformers. *arXiv preprint arXiv:2510.27015*, 2025.

Samy Jelassi, Stéphane d'Ascoli, Carles Domingo-Enrich, Yuhuai Wu, Yuanzhi Li, and François Charton. Length generalization in arithmetic transformers. *arXiv preprint arXiv:2306.15400*, 2023.

Amirhossein Kazemnejad, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Payel Das, and Siva Reddy. The impact of positional encoding on length generalization in transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

William Merrill, Ashish Sabharwal, and Noah A. Smith. Saturated transformers are constant-depth threshold circuits. *Transactions of the Association for Computational Linguistics (TACL)*, 10, 2022.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, et al. In-context learning and induction heads. *Transformer Circuits Thread*, 2022.

Ofir Press, Noah A. Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations (ICLR)*, 2022.

Anian Ruoss, Grégoire Delétang, Tim Genewein, Jordi Grau-Moya, Róbert Csordás, Mehdi Bennani, Shane Legg, and Joel Veness. Randomized positional encodings boost length generalization of transformers. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023.

Clayton Sanford, Daniel Hsu, and Matus Telgarsky. Representational strengths and limitations of transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19, 2018.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568, 2024.

Davoud Ataee Tarzanagh, Yingcong Li, Christos Thrampoulidis, and Samet Oymak. Transformers as support vector machines. *arXiv preprint arXiv:2308.16898*, 2023a.

Davoud Ataee Tarzanagh, Yingcong Li, Xuechen Zhang, and Samet Oymak. Max-margin token selection in attention mechanism. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023b.

Jie Wang, Tao Ji, Yuanbin Wu, Hang Yan, Tao Gui, and Qi Zhang. Length generalization of causal transformers without position encoding. *arXiv preprint arXiv:2404.12224*, 2024.

Gail Weiss, Yoav Goldberg, and Eran Yahav. Thinking like transformers. In *International Conference on Machine Learning (ICML)*, 2021.

Hattie Zhou, Arwen Bradley, Etai Littwin, Noam Razin, Omid Saremi, Josh Susskind, Samy Bengio, and Preetum Nakkiran. What algorithms can transformers learn? a study in length generalization. In *International Conference on Learning Representations (ICLR)*, 2024a.

Yongchao Zhou, Uri Alon, Xinyun Chen, Xuezhi Wang, Rishabh Agarwal, and Denny Zhou. Transformers can achieve length generalization but not robustly. *arXiv preprint arXiv:2402.09371*, 2024b.

A Full proofs

A.1 Theorem 1 (Simulation sufficiency, any realizing W)

Fix any W satisfying (i)–(ii). Under (S1)–(S3) the row-i logits at any length n are $a_{ij}^{(n)} = \varphi(i)^\top W\varphi(j)$ for $j \le i$, with the content term argmax-neutral. *New rows* $i \in [L_0, L)$: by (ii) the simulated $P_L[i,\cdot] = a_{i\cdot}^{(L)}$ has its unique masked maximum at $t_L(i)$ with gap $\ge \gamma_L$. *Old rows* $i < L_0$: the causal mask restricts the key set to $\{0,\dots,i\} \subseteq \{0,\dots,L_0-1\}$, and for those j the features $\varphi(j)$ are exactly those present at length L_0; hence $a_{ij}^{(L)} = a_{ij}^{(L_0)}$ entrywise, so the masked argmax and gap $\ge \gamma_0$ are inherited verbatim from (i). Every row's unique masked maximizer is $t_L(i)$ with gap $\ge \min(\gamma_0, \gamma_L) > 0$. □

Remark (why no min-norm). An earlier draft stated this for the min-norm interpolant W^* and invoked $\Phi_{L_0} W^* \Phi_{L_0}^\top = R_{L_0}$ (false unless $R_{L_0} \in \mathcal{K}_{L_0}$). The argument needs no such exactness and no choice of W: only that the chosen W is argmax-correct at L_0 and has $\alpha_L(W)$=1. Hence it covers the max-margin W used in the experiments (the fitted kernel), though not the trained transformer. The theorem's force is entirely the old-row causal-mask invariance; the new-row clause restates (ii).

A.2 Proposition 1 (Affine features force endpoint argmax)

With $\varphi(i) = [1, i]^\top$ write $W = \left(\begin{smallmatrix} W_{00} & W_{01} \\ W_{10} & W_{11} \end{smallmatrix}\right)$. Then $a_{ij} = \varphi(i)^\top W\varphi(j) = c_0(i) + c_1(i)\, j$ with $c_0(i) = W_{00} + iW_{10}$, $c_1(i) = W_{01} + iW_{11}$, which is affine in the key index $j \in \{0,\dots,i\}$. An affine function on a finite integer chain is monotone, so its maximizers are endpoints, and the maximizer is unique unless $c_1(i) = 0$ (a flat row, margin 0, excluded by (S2)). For a strict interior target $0 < t < i$, optimality would require $a_{it} \ge a_{i0}$ and $a_{it} \ge a_{ii}$, i.e. $t\, c_1(i) \ge 0$ and $(t-i)c_1(i) \ge 0$; with $t > 0,\ t - i < 0$ this forces $c_1(i) \ge 0$ and $c_1(i) \le 0$, hence $c_1(i) = 0$ and a zero margin. Therefore no interior key is ever the *unique, positive-margin* maximizer. Since the competitor set $\{0,\dots,i\}$ does not depend on n, the conclusion holds at every length; prev-token ($t = i-1$, interior for $i \ge 2$) is thus unrealizable for all $n \ge 3$. □

Caveat. The statement is about the unique positive-margin argmax, *not* the set inclusion $\arg\max \subseteq \{0, i\}$, which fails on the measure-zero locus $c_1(i) = 0$ (there the argmax set is the whole chain).

A.3 Proposition 2 (Learned-absolute underdetermination)

Let $\varphi(i) = v_i$ with $v_0,\dots,v_{L_0-1}$ fixed and $v_{L_0},\dots,v_{L-1} \sim \mu$ i.i.d. from an absolutely continuous isotropic law, and $W = S(v_0,\dots,v_{L_0-1})$ any rule of the trained rows only (so $W \perp \{v_j\}_{j \ge L_0}$). Put $g(j) = W\varphi(j)$. *Uniqueness:* for $j \ne j'$, $a_{ij} - a_{ij'} = v_i^\top(g(j) - g(j'))$ is a nonzero affine functional of the absolutely-continuous v_i, so vanishes with probability 0; a finite union over key pairs is null, hence $M_i = \arg\max_{j \le i} a_{ij}$ is a.s. unique. *Target-independence:* the map $v_i \mapsto \arg\max_j v_i^\top g(j)$ is a fixed measurable function of $(v_i, \{g(j)\})$ in which the label $t_L(i)$ never appears, and S cannot encode it into the new-key geometry; thus $\mathrm{Law}(M_i)$ does not reference $t_L(i)$. *Chance rate:* conditioned on W, the off-diagonal new keys $\{g(j) : L_0 \le j < i\} = \{Wv_j\}$ are i.i.d. and independent of v_i, so by exchangeability the argmax among them is uniform, giving $\Pr[M_i = t_L(i)] = 1/(k-1)$ with $k = |\{L_0 \le j \le i\}|$. □

Underdetermination, not cone-emptiness. A rescuing W realizing the target on all new rows exists with positive probability (Monte Carlo: $\approx$ 20–35% of draws; `experiments/theorem2b_learned_abs_montecarlo.py`); the selection rule is simply blind to the untrained coordinates. The self-key $j = i$ is excluded because $a_{ii} = v_i^\top W v_i$ is a *quadratic* (non-exchangeable) form with mean $\operatorname{tr} W$, which is why the rate is $1/(k-1)$ rather than $1/k$.

A.4 Proposition 3 (Sinusoidal: existence vs. selection)

(A) Existence. With incommensurate geometric frequencies $\{\theta_k\}$, the translation-invariant kernel $g(d) = \sum_k \cos(\theta_k(d-\delta)) \le \#\{\theta_k\}$ with equality iff $\theta_k(d-\delta) \in 2\pi\mathbb{Z}$ for all k; incommensurability forces $d = \delta$, so g has a unique global maximum at the target offset for *any* fixed δ (verified numerically to $\delta \approx 5 \times 10^4$, far beyond the lowest-frequency period — refuting the "frequency-band ceiling" intuition). Hence a W realizing

a fixed offset exists at all lengths. *(B) Growing offsets.* A single translation-invariant kernel has one argmax offset, which cannot equal a length-dependent $\delta_n \sim n$ at two lengths; copy/reverse therefore extrapolate at chance ($\alpha_L \approx 0$). *Existence $\neq$ selection:* the L_0-fit W^* does not recover the fixed-offset rescue — under the transfer operator sinusoidal prev decays ($\alpha_L = 0.45$ at L=60, 0.34 at L=128), and the sub-unit value *is* α_L quantifying the selection gap. □

A.5 Theorem 2 (Relative-head deep-row characterization)

Let $a_{ij} = \kappa(i-j)$ on causal keys $j \leq i$, unique maximizer $c(\kappa)$, support radius D, fixed tie rule. *Deep-row constancy (Lemma).* For any deep row $i \geq \max(D, c(\kappa))$ the window $[0, i]$ contains $c(\kappa)$ and all offsets that could tie/beat it, so the masked argmax offset is the constant $c(\kappa)$ for all such i and all n (verified over 4000 random kernels, both tie rules).

(⇐) If the required deep-row offset is the constant δ^*, take $\kappa(d) = -|d - \delta^*|$: every deep row argmaxes δ^* with margin 1 at all L (old rows by the causal-mask identity). Shallow rows are finitely many, so their bounded contribution to the $[L_0, L)$ average vanishes and $\alpha_L \to 1$. (Verified for $\delta^* \in \{1, 2, 3, 5, 8\}$, L up to 5000.)

(⇒) If the deep-row offset is not eventually constant, one of two failures occurs. **(a) It grows with** n (copy, $\delta_n = \lceil n/2 \rceil$): for all $L \geq 2c(\kappa)$ every copy row is deep with realized offset $c(\kappa)$, which cannot equal $\lceil L/2 \rceil \to \infty$; so $\alpha_L \to 0$. **(b) It varies within a length** (reverse, $\delta_n(i) = 2i - 2\lceil n/2 \rceil + 1$, distinct offsets $\{1, 3, 5, \dots\}$): a nested-window monotonicity argument gives a contradiction already at L_0 (row 19 forces $\kappa[19] > \kappa[17]$; then on row 17 the window $\{0..17\}$ already contains offset 17 with $\kappa[17] > \kappa[15]$, so $d^*(17) = 17 \neq 15$), independent of $c(\kappa)$. 20k/500k-kernel sweeps confirm no κ solves reverse at L_0. □

Why deep-row/asymptotic, not a finite-L "complete iff" (verification-caught, independently reproduced). An earlier draft added the necessity clause "shallow rows = $\min(i, \delta^*)$." It is **false**: $\kappa = [1.46, 1.96, 1.80, 6.32]$ (offsets 0–3, $c(\kappa) = 3$) makes deep rows take offset 3 but shallow row i=2 argmax offset $1 \neq \min(2, 3)$=2 (window values $[1.46, 1.96, 1.80]$). Shallow rows are a finite, κ-dependent set, uncharacterized by a universal formula but immaterial to α_L as $L \to \infty$ — hence the criterion is on deep rows. Two further cautions: the growing-offset clash is *asymptotic*, not two-length ($\kappa[10]$=5, $\kappa[30]$=8 solves copy at both L_0=20 and L=60 via shallow clipping); and "bounded offset" is insufficient (the i-parity offset $\delta(i) \in \{1, 2\}$ scores exactly 0.5, 20k-kernel sweep). *Capacity corollary:* variable-length training cannot reach acc=1 on copy/reverse ($\text{acc}_{\text{train}} < 1$).

A.6 Proposition 4 (Absolute/relative duality)

Relative $\kappa(i-j)$: a fixed offset k is realized by κ peaking at d=k, correct on every row at every length (α_L=1); attend-first (t=0) demands offset i, which grows, so by Theorem 2's mechanism $\alpha_L \to 0$. *Absolute* kernel $K[i, j]$: attend-first is realized by the length-independent column-0-hot kernel $K[i, j] = \mathbf{1}[j{=}0]$ (margin +1 at all lengths); a fixed relative offset moves the target across absolute positions, so new rows are untrained and $\alpha_L \to$ chance (Prop. 2). Growing offsets defeat both. Verified in `experiments/duality_abs_vs_rel.py` (attend-first abs 1.00/rel 0.00; prev abs 0.046/rel 1.00; relative chance floor is $1/(k-1)$ over off-diagonal new keys, the self-key being a non-exchangeable quadratic competitor). □

B Experimental details

Architecture. Decoder-only transformer, $d_{\text{model}} = 64$, GELU MLP (width $4d$), pre-LN, learned token embedding (vocab 12). PEs: NoPE; sinusoidal (additive); learned-absolute (table of size L_0, positions $\geq L_0$ clamped to $L_0 - 1$); RoPE (rotary on q, k); ALiBi (fixed geometric slopes). Tiers: 1-layer/1-head, 2-layer/4-head, 4-layer/4-head. **Training.** AdamW (lr 3×10^{-4}, wd 0.01), batch 64, 2000 steps, variable train length $n \sim \text{Unif}[L_0/2, L_0]$, $L_0 = 20$, cross-entropy over query positions (ignore-index elsewhere). **Tasks** (target position $t_n(i)$, all position-essential via small vocab): prev (i−1), skip-k (i−k, $k \in \{2, 3, 5\}$), copy ($i - \lceil n/2 \rceil$), reverse, first (0). **Evaluation.** Accuracy on extrapolation query rows $[L_0, L)$ at $L \in \{30, 40, 50, 60, 80\}$, 5 seeds. **Predictor fit.** Cross-entropy fit of the kernel coefficients to the L_0 targets (400 Adam steps), PE-

structural constraints imposed (RoPE relative basis; ALiBi recency slope constrained ≤ 0); strict-unique-argmax scoring on new rows. **Compute.** Apple M3 Pro (MPS); the full 2625-measurement, 5-seed study runs in $\sim$55 min.

C Additional results

Table 4 reports the depth trend (H3) over all three tiers; Fig. 6 plots it. As depth grows ($1 \to 2 \to 4$ layers), the learnable-cell *graded* rank correlation drops from 0.90 (CI $[0.73, 0.95]$) to 0.31 and 0.28 with bootstrap CIs that include 0 (Table 4); the *sufficient-screen* precision instead stays high and tight ($0.93 \to 0.94 \to 0.97$). The lost graded signal is explained by the content channel: the set of low-α_L cells that nonetheless generalize *grows* with depth — {ALiBi prev/first/skip2, NoPE first} at 2 layers, adding ALiBi skip3 and NoPE prev at 4 layers. Depth does not break the sufficient-indicator property; it adds more content-channel routes around a low positional α_L, exactly the (S3) boundary.

Table 4: Depth trend (H3), learnable cells (train acc > 0.8), 5 seeds. The Spearman column is *cell-level* with a bootstrap 95% CI over the (PE, task) cells; the precision column is the rate $\Pr[\mathrm{LG} \geq 0.7 \mid \alpha_L \geq 0.5]$ with a bootstrap 95% CI over the high-α_L *rows*. α_L stays a high-precision *sufficient* indicator at every depth (precision CIs well above chance); the graded rank correlation, by contrast, is significant only at depth 1 — at depths $2, 4$ its CI *includes* 0, consistent with α_L acting as a binary screen rather than a graded predictor once content rescues low-α_L cells.

tier	depth	learnable Spearman ρ (95% CI)	sufficiency precision (95% CI)
1L1H	1	0.90 $[0.73, 0.95]$	0.93 $[0.88, 0.96]$
2L4H	2	0.31 $[-0.16, 0.76]$	0.94 $[0.89, 0.97]$
4L4H	4	0.28 $[-0.15, 0.66]$	0.97 $[0.94, 0.99]$

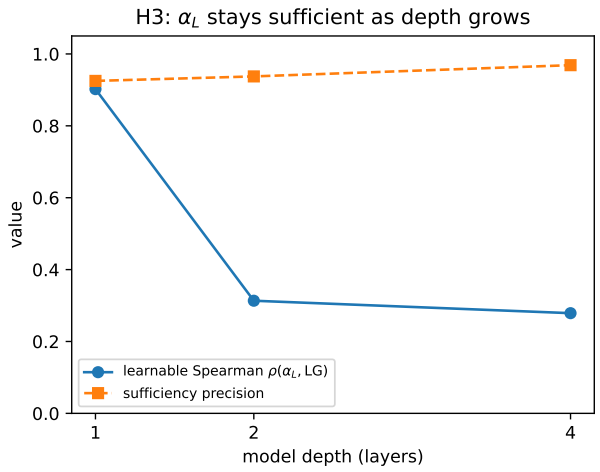


Figure 6: H3: learnable-cell Spearman $\rho(\alpha_L, \mathrm{LG})$ (circles) drops with depth (its bootstrap CI includes 0 at depths $2, 4$), while sufficiency precision (squares) stays ≥ 0.93 with tight CIs. α_L remains a *sufficient screen* at all depths.

Inductive-bias ablation (which fit drives the correlation). Table 5 computes α_L three ways on the 75 theorem-matched (PE,task, L) rows and correlates each with empirical LG. The min-norm interpolant of Theorem 1 yields only ρ=0.56; the cross-entropy fit yields 0.91 and its weight-decay$\to 0$ max-margin limit 0.92. The predictive content of α_L therefore comes from the *max-margin* direction that gradient descent realizes (Soudry et al., 2018), not the min-norm object the theorem certifies — a gap we state openly.

Table 5: Inductive-bias ablation (75 theorem-matched rows). The max-margin (CE) fit, not the min-norm interpolant, drives the α_L–LG correlation.

kernel fit	Spearman ρ	Pearson
min-norm interpolant (Thm 1's W^*)	0.56	0.57
cross-entropy, wd= 10^{-4} (reported α_L)	0.91	0.95
cross-entropy, wd$\to$ 0 (max-margin)	0.92	0.95

Robustness checks. The learnability cutoff is inert: cell-level $\rho = 0.90/0.90/0.90$ at train $> \{0.5, 0.7, 0.9\}$ (train accuracy is bimodal). On the 20 dropped cells α_L has std 0.02 (uniformly ≈ 0), so it provides no rank signal there by construction. Scale robustness (vocab 64, d=128) is reported in App. D; training-length robustness (the screen AUC stays 1.0 and the per-cell verdicts are unchanged at $L_0 \in \{10, 20, 40\}$) in App. E; and PE frequency/θ-base/dimension robustness in App. G.

D Scale robustness

Because α_L depends only on (Φ, t_n), it is *identical* across model scales; the question is whether the empirical LG it predicts is stable. We re-ran the full study at a 5× larger vocabulary (64 vs 12) and 2× width (d=128 vs 64). The theorem-matched cell-level correlation is *unchanged* (indeed slightly higher): $\rho = 0.92$ (vs 0.90 at d=64), with sufficiency precision 0.94; the scale tier moves 0.36$\to$0.45 (Table 6). The headline relationship is therefore not an artifact of the tiny vocabulary or width. (This is a robustness check within the synthetic position-driven family, not a claim about natural-language LLMs — see §7.)

Table 6: Scale robustness: cell-level Spearman $\rho(\alpha_L, \mathrm{LG})$ on learnable cells, at the paper's default size vs. a larger model. The predictor is identical; only the trained models change.

tier	vocab 12, d=64 (main)	vocab 64, d=128
1L1H (theorem-matched)	0.90	0.92
2L4H (scale)	0.36	0.45

E Training-length (L_0) sensitivity

Is the clean separation an artifact of the default training length L_0=20? We re-ran the theorem-matched (1L1H) study at $L_0 \in \{10, 20, 40\}$, each evaluated at $3 \times L_0$ (so the extrapolation factor is held fixed), 2 seeds. The screen AUC is 1.000 at every L_0, and *the same four cells generalize* in each case (rope prev/skip-3, sinusoidal/learned-abs first); the Theorem 2 verdicts are unchanged (relative bounded offsets transfer; relative growing offsets do not). The separation is therefore not specific to L_0=20 (Table 7).

Table 7: L_0-sensitivity (1L1H, predictor α_L / empirical LG at L=3L_0). The screen AUC is 1.0 at every training length and the per-cell verdicts are stable; the separation is not an artifact of L_0=20. Generated by `experiments/l0_sensitivity.py`.

	L_0=10 ($\to$ 30)	L_0=20 ($\to$ 60)	L_0=40 ($\to$ 120)
screen AUC	**1.00**	**1.00**	**1.00**
# generalizing cells	4/25	4/25	4/25
rope / prev (α_L / LG)	1.00 / 0.94	1.00 / 0.82	1.00 / 0.87
rope / skip-3	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00
rope / copy (growing)	0.00 / 0.12	0.00 / 0.11	0.00 / 0.10
rope / reverse (growing)	0.00 / 0.16	0.03 / 0.14	0.00 / 0.12
sinusoidal / first (absolute)	1.00 / 0.91	1.00 / 0.90	1.00 / 0.87
learned-abs / first (absolute)	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00

F A content-aware extension: α_L^+ (with a demonstration)

The binding limitation is (S3): where the content channel carries the target, the positional α_L under-predicts. We show the one-sidedness is closable, with a concrete content-driven task and an honest account of what is and isn't recovered.

Setup. Augment the realizable space from the positional kernel $\mathcal{K}_n$ to $\mathcal{K}_n + \mathcal{C}$, where $\mathcal{C}$ is the content-matching channel $e(x_i)^\top W_c e(x_j)$. Define $\alpha_L^+ = \max(\alpha_L,\ \alpha_L^{\text{content}})$, where $\alpha_L^{\text{content}} = 1$ iff the target is realizable by a *length-stationary* content rule (a token match composed with a fixed shift). The canonical such task is **induction**: the target at i is the token following the previous occurrence of x_i — content-determined, hence α_L=0 for *every* PE (no length-dependent positional rule realizes it), while $\alpha_L^{\text{content}}$=1.

Demonstration (Table 8). We train induction across PEs. At 1 layer the circuit cannot form (train acc ≈ 0.25); at 2 layers it is learned (train 0.76–1.0) and *generalizes*, strongly for ALiBi (LG = 0.84), partially for RoPE/NoPE (0.43/0.34). So positional $\alpha_L = 0$ *mispredicts* a cell with LG 0.84 — a real one-sidedness failure — which α_L^+ correctly flags as content-realizable. **Honest caveat:** induction LG is PE-*modulated* (0.26–0.84: the content circuit still needs a working prev-token head, which absolute PEs support poorly), so the *binary* $\alpha_L^{\text{content}}$ over-predicts for sinusoidal/learned-abs. A graded content channel weighting $\alpha_L^{\text{content}}$ by the PE's support for the match circuit — and the multi-head extension that provably escapes Theorem 2 (anchor head + content-match head solves copy at depth 1) — are the natural next steps. The point established here is narrow but real: *the positional indicator's blind spot is identifiable and the content channel recovers a cell it misses.*

Searching for a discriminating cell — a negative result. Reviewers rightly ask for one (PE, task) cell where α_L beats the relative/absolute taxonomy. We searched the most promising direction: a *frequency-scaled RoPE* (still relative) on the *content-neutral* skip-k task (the query token is independent of the target $s[i-k]$, so content cannot supply the answer), sweeping the frequency scale to provoke a resolution-limited failure α_L would catch. The result is negative (Table 9): standard and moderately-low frequencies resolve the offset (α_L=1, LG $\approx$ 1); at scale 0.03 the trained head *still* generalizes (LG = 0.88) while α_L=0 — an α_L *false negative*, because the softmax head resolves an offset the linear-kernel fit deems unrealizable; and at scale 0.01 the model fails to learn even in-distribution (excluded by the learnability filter). **No discriminating cell exists in this family, and the probe instead reveals that α_L *under*-predicts low-frequency relative PEs.** Combined with the row-level over-predictions (rope-first), the honest verdict is that α_L *matches but does not beat* the taxonomy as a predictor; its contribution is the theory that *explains* the taxonomy, not a superior predictor. (This corrects a draft that mislabeled the low-frequency case as "content rescue": skip-k is content-neutral, so it is positional under-prediction, not content.)

Table 8: Content-driven induction (target = token after the previous occurrence of x_i): positional $\alpha_L = 0$ for all PEs, yet 2-layer models generalize (LG up to 0.84). α_L^+ flags the content-realizable regime α_L misses; residual PE-modulation of content LG is future work.

PE	train acc (2L4H)	LG@60 (2L4H)	positional α_L
ALiBi	0.96	**0.84**	0
RoPE	1.00	0.43	0
NoPE	0.76	0.34	0
learned-abs	0.99	0.27	0
sinusoidal	0.98	0.26	0

Table 9: RoPE frequency sweep on content-neutral skip-3 (1L1H). No discriminating cell: where the model generalizes the taxonomy is right; at scale 0.03 α_L *under*-predicts (false negative); at 0.01 the model cannot learn at all.

RoPE freq. scale	α_L	train acc	LG@60
1.0 / 0.3	1.00	1.00	1.00
0.1	1.00	0.94	0.93
0.03	**0.00**	0.88	**0.88**
0.01	0.00	0.23	0.18

G Frequency and dimension sensitivity of α_L

Since α_L is computed from the PE feature map, it is fair to ask whether its verdict is an artifact of our default frequency count, θ-base, or feature dimension. We sweep each (predictor-only, no training; $L_0{=}20 \to L{=}60$) and find a clean separation: the realizability *verdict* is stable wherever the theory makes a clean claim, and the *only* cells whose realizability verdict moves are exactly the sinusoidal bounded-offset cells the theory flags as a selection gap (Prop. 3); a few other cells show small sub-threshold raw movement that never flips the verdict.

Table 10: α_L for RoPE under varying frequency count num (feature dimension $m{=}2\,\text{num}$) and θ-base. The verdict is invariant: fixed relative offsets transfer (prev, skip3=1.00 everywhere), growing offsets never transfer (copy/reverse≈0), and a relative kernel cannot hit a fixed absolute position (first stays at chance). For low-frequency *scaling* see Table 9.

RoPE setting	prev	skip3	copy	reverse	first
num=8 (m=16)	1.00	1.00	0.00	0.00	0.17
num=16 (m=32)	1.00	1.00	0.00	0.03	0.20
num=32 (m=64)	1.00	1.00	0.00	0.00	0.03
base 10^3	1.00	1.00	0.00	0.03	0.05
base 10^5	1.00	1.00	0.00	0.03	0.03

Table 11: α_L for sinusoidal under the same sweep. The absolute verdict is invariant (first≈1 everywhere; growing offsets ≈0). The bounded-offset cells (prev/skip3) are the *only* entries whose verdict moves with frequency — they *degrade* as frequencies are added (num:8→32 drives prev 0.47→0.00); the other cells' raw α_L shifts stay sub-threshold (e.g. first 0.93→1.00) and never flip the verdict. This is not instability of the theory but exactly its prediction: enlarging the realizable space does not help the L_0-fit *select* a transferring kernel (existence ≠ selection, Prop. 3). Verdict movement appears precisely at the selection gap and nowhere else.

sinusoidal setting	prev	skip3	copy	reverse	first
num=8 (m=16)	0.47	0.50	0.00	0.00	0.93
num=16 (m=32)	0.45	0.25	0.07	0.00	1.00
num=32 (m=64)	0.00	0.00	0.00	0.00	1.00
base 10^3	0.17	0.28	0.00	0.00	1.00
base 10^5	0.15	0.12	0.00	0.00	1.00

Takeaway. Across frequency count, θ-base, and feature dimension, α_L's *binary realizability verdict* is stable for every regime with a clean theoretical prediction; the residual frequency dependence is confined to the sinusoidal bounded-offset selection gap, which is the behaviour Prop. 3 predicts. The screen is therefore not an artifact of the default PE hyperparameters. (Generated by `experiments/freq_dim_sensitivity.py`.)

H Reproducibility checklist

The full source (predictor, models, tasks, experiments, theory-verification scripts, and this paper) is released; for double-blind review it is provided as an anonymized supplement, to be de-anonymized on acceptance.

- **Predictor.** α_L is in `src/spectra.py` (`simulate(pe,task,`L_0`,`L`)`); it uses no end-to-end Transformer training, no data samples, and no learned Transformer parameters — only a deterministic kernel-selection on the PE basis. Models: `src/model.py` (the five PEs); tasks: `src/data.py`.
- **Seeds.** The main study uses seeds $\{0, 1, 2, 3, 4\}$ (fixed in `run_study.py`); bootstrap CIs resample over these. All randomness is seeded.
- **Configuration.** The design grid (PEs $\times$ tasks $\times$ lengths $\times$ seeds $\times$ depth tiers) is Table 2; architecture/optimizer/training details are in App. B. Tiers and tags are env-overridable in `run_study.py` (`SEEDS`, `DMODEL`, `TIERS_ONLY`, `STUDY_TAG`); the run is resumable.
- **Data behind every number.** `results/study_rows.json` (all 2625 rows, including runs the model failed to learn at L_0 — *no* failed runs are silently dropped; the learnability filter `train_acc`>0.8 is applied explicitly and reported alongside the unconditioned numbers). Summaries in `results/study_summary.json`.
- **Figures/tables $\rightarrow$ scripts.** Figs. 3–6 and the heatmaps: `make_figures.py`. Headline statistics (AUC, permutation p, Spearman + bootstrap CIs, sufficiency precision): `review_stats.py`. Scale robustness (Table 6): `run_study.py` with `STUDY_TAG=scale`. L_0-sensitivity (Table 7): `l0_sensitivity.py`. Frequency/dimension sensitivity (Tables 10–11): `freq_dim_sensitivity.py`. Content extension / induction (Table 8) and the RoPE frequency sweep (Table 9): `ablation_fit.py` and the content study.
- **Theory verification.** Each formal claim has an executable check: relative growing-offset impossibility (`theorem_relative_copy_impossibility.py`), the deep-row iff and its counterexample κ=$[1.46, 1.96, 1.80, 6.32]$ (`iff_constant_offset_lg.py`), learned-absolute chance $1/(k{-}1)$ (`verify_learned_abs_fixed_position_iff.py`, `theorem2b_learned_abs_montecarlo.py`), the absolute/relative duality (`duality_abs_vs_rel.py`), and the Conjecture 1 cone/selection probes (`underdetermined_conj1_*.py`).
- **Compute.** The full study runs in $\approx$50 min on one Apple M3 Pro (MPS); the predictor and all sensitivity sweeps run in seconds on CPU.