Copying Under Repetition

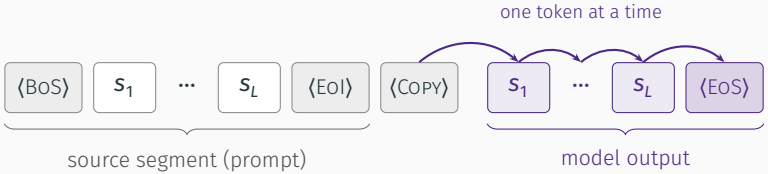
From Prefix Matching to Position Hashing

Anxin (Bob) Guo [1] Tianlong Huang [2] Chenxiao Yang [3] Zhiyuan Li [3]

[1]Northwestern University [2]Duke University [3]TTIC

Autoregressive copying

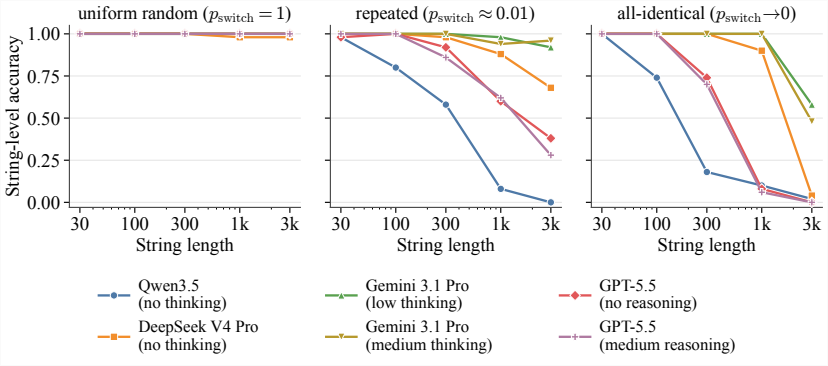
Given a source string between delimiters and a ⟨Copy⟩ cue, the model must reproduce the source one token at a time and stop at ⟨EoS⟩:



Language models do this constantly: quoting a retrieved document, reproducing code, carrying an answer out of a chain of thought.

It looks like the easiest thing one could ask a language model to do.

… except on repeated strings

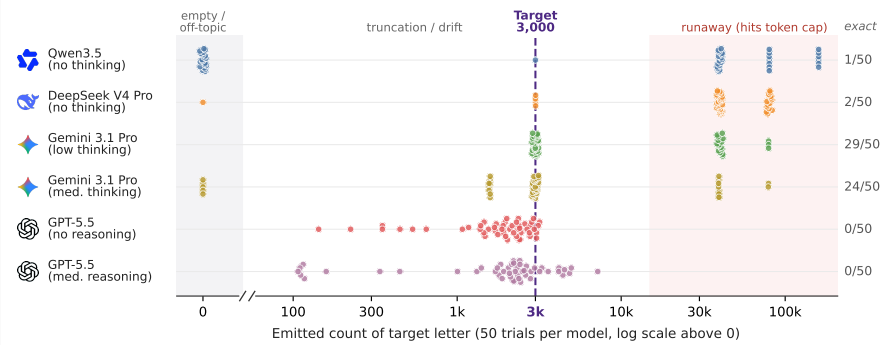


String-level accuracy

random	wtjjqutbagbrjzonghicgopzfccvfhbpsoesyevngcmhawcujftgnojqgtitudjiqhcqzx…
$p_{\text{switch}} \approx 0.01$	…xxxxxxxxxxxxxxxxxxxxxxxeeeeeeeeeeeeeeeeeeeeeeeeggggggggggggggggggggggg…
all-identical	kkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkk…

What the failures look like

All-identical sources of length **3,000** (aaaa…a):



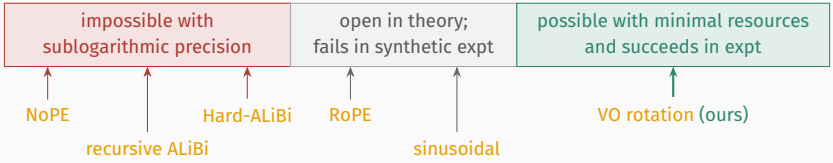
A specific type of error

Repetition makes the next letter *trivial* to decide. But this is exactly when models fail to correctly stop!

This talk

A potential explanation through the lens of *expressivity*.

- Most *position encodings* (PEs) struggle to encode relative distances.
- Some PEs simply cannot terminate correctly using $0.9 \log N$ precision, even with arbitrary width!
- We identify a PE where copying is easily expressible with $O(\log \log N)$ precision. With it, a small Transformer can learn to copy strings of length 5k.
- Along the way: what expressivity framework is suitable for this question (spoiler: not the popular *circuit complexity* and *formal language* ones).



Plan

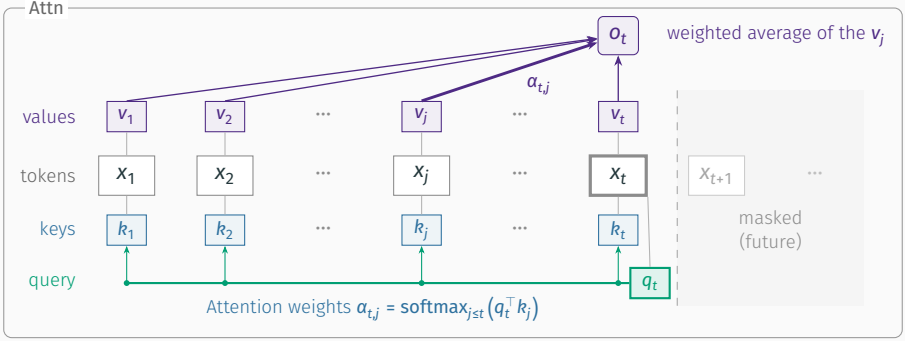
1. Background: attention and positional encodings
2. The right expressivity question
3. Lower bounds
4. Upper bounds
5. Experiments and takeaways

A causal self-attention head

Position t produces *query* vector q_t; every position j produces a *key* k_j and a *value* v_j.

$$\alpha_{t,j} = \mathrm{softmax}_{j \le t}\left(q_t^\top k_j\right) = \frac{e^{q_t^\top k_j}}{\sum_{\ell \le t} e^{q_t^\top k_\ell}}, \qquad o_t = \sum_{j \le t} \alpha_{t,j}\, v_j.$$

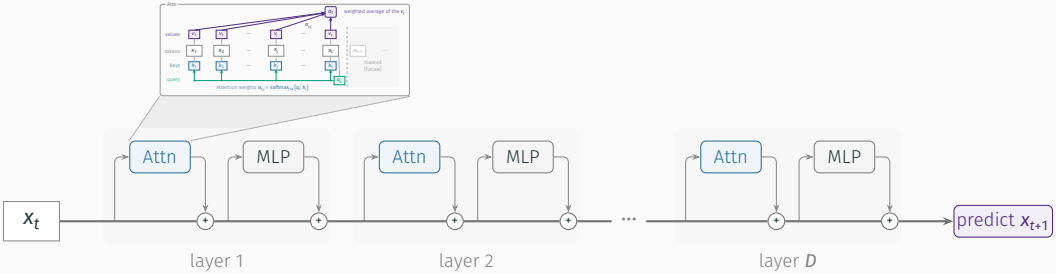
A *soft lookup*: score every visible key, return a convex combination of their values.



Stack it up: this is NoPE

Each token's vector goes through D layers: *causal self-attention*, then an MLP.
Each adds its output to the vector it read (a residual connection).

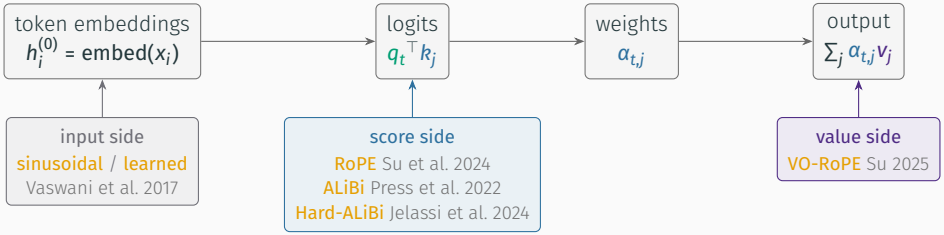
$$u_t = h_t^{(\ell-1)} + \mathrm{Attn}^{(\ell)}\left(h^{(\ell-1)}\right)_t, \qquad h_t^{(\ell)} = u_t + \mathrm{MLP}^{(\ell)}(u_t).$$



No position information is added anywhere: this is NoPE (Kazemnejad et al. 2023).

Adding position explicitly

Most models do not stop at NoPE: they add position explicitly.



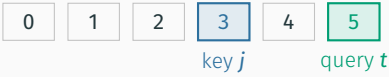
Roughly: we show impossibility for some score-side PEs, and a constructive positive result for VO-RoPE with integer periods.

RoPE: rotate the queries and keys

Rotate q_t by $t\theta$ and k_j by $j\theta$. The relative rotation in the inner product is $(j - t)\theta$.

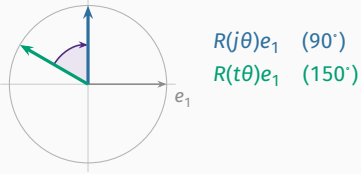
$$\langle R(t\theta)\, q_t,\ R(j\theta)\, k_j \rangle \ = \ \langle q_t,\ R\big((j - t)\theta\big)\, k_j \rangle.$$

Token positions



Move both. The gap stays at two.

Rotation phases ($\theta = 30^\circ$)



Query to key: $(j - t)\theta = -60^\circ$

Multiple frequencies $\theta_1, \theta_2, \ldots$ run in parallel to resolve both near and far distances. RoPE is the default PE in open-weight LLMs.

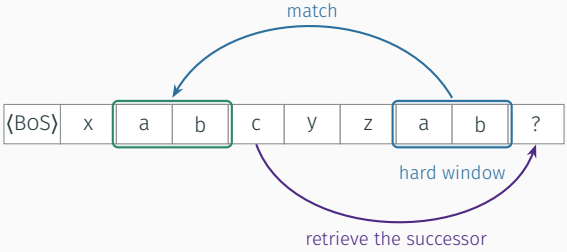
ALiBi, Hard-ALiBi, and how copying was solved before

$$\text{NoPE}: \quad \alpha_{t,j} = \mathrm{softmax}_{j \le t}\left(q_t^\top k_j\right)$$

$$\text{ALiBi (Press et al. 2022)}: \quad \alpha_{t,j} = \mathrm{softmax}_{j \le t}\left(q_t^\top k_j - 2^{-h} \cdot (t-j)\right)$$

$$\text{Hard-ALiBi (Jelassi et al. 2024)}: \quad \alpha_{t,j} = \mathrm{softmax}_{j \le t}\left(q_t^\top k_j - \mathbf{1}\{t - j > w\} \cdot \infty\right)$$

A hard window allows the model to directly read off *the previous* w *tokens* exactly. One can construct a small transformer that copies using this content hashing mechanism.



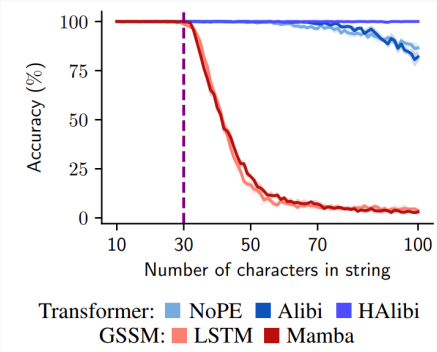
The prior positive result, and its limitations

Main theorem (Jelassi et al. 2024)

A two-layer transformer with Hard-ALiBi copies strings of length N using $O(\log N)$ width and $O(\log \log N)$ precision — and length generalizes strongly in practice.

Limitation: the guarantee assumes the source contains no repeated n-gram.

That assumption breaks down completely on aaaa…a.



Trained on length ≤ 30; evaluated up to 100.

Plan

1. Background: attention and positional encodings
2. The right expressivity question
3. Lower bounds
4. Upper bounds
5. Experiments and takeaways

The regime we study: repetition + low precision

Two ingredients make copying hard for real language models:

1. High repetition

"Content hashing" breaks down. Model must find the next token via position, not content.

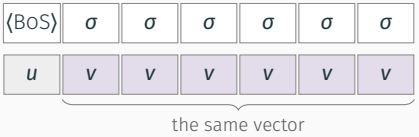
2. Low (sub-log) precision

Real LLMs run at BF16 (7 mantissa bits) or FP8 (3–4 bits) over contexts of $N \sim 10^5$–10^6 ($\log_2 N \approx 17$–20).

Mantissa precision p is **sub-logarithmic** in context length N: $p \ll \log_2 N$.

What repetition breaks

On an all-identical source, every token in layer 1 carries the same value vector $\boldsymbol{v}$:



$$o_t = \alpha_0\, u + (1 - \alpha_0)\, v$$

only α_0 depends on position!

The key observation

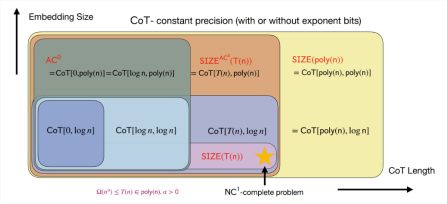
Score-side PEs only reweight identical vectors — position leaks in only through the delimiter weight α_0. In low precision, it is hence very hard to distinguish different relative positions.

To make this formal, we need the correct expressivity framework.

Framework 1: circuit complexity (length-specific)

Circuit complexity (Merrill & Sabharwal 2023, 2024; Li et al. 2024): fix the input length n and compare transformers with circuit classes.

- Natural for upper bounds: constant-precision transformers without CoT $\subseteq \mathsf{AC}^0$, for instance.
- Hard for lower bounds: a circuit family is wired differently for each n. Hard for one transformer to simulate (Merrill & Sabharwal 2023, 2024).
- Li et al. get equalities by letting the transformer *also* depend on n.

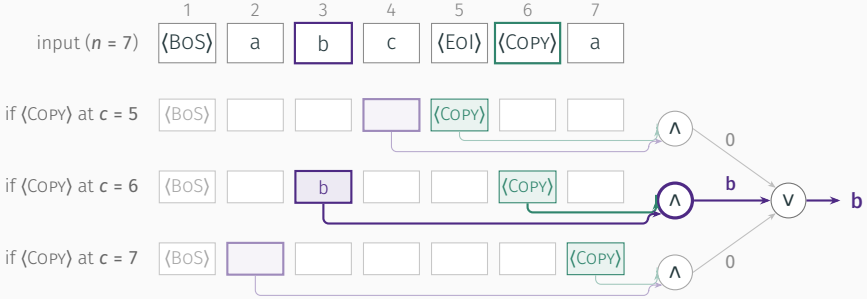


Li et al. (ICLR 2024), *Chain of Thought Empowers Transformers to Solve Inherently Serial Problems.*

Framework 1: the only "circuit upper bound" is uniform AC^0

Fix input length n. Then autoregressive copying is easy for circuits.

For each candidate ⟨COPY⟩ position c, an AND checks if x_c = ⟨COPY⟩, and if so, passes on the value of cell $n - c + 2$. An OR picks out the correct (nonzero) result.



Depth 2, linear size. But wired in separately for each n!

Framework 2: fixed parameters for unbounded length

A classic line of work fixes one transformer and sends the length $n \to \infty$ (*formal languages* and *Turing-machine simulation*).

Positive results typically rely on extra power:

- High precision ($\geq \log n$ bits) for exact counting (Yang et al. 2025);
- Hard attention for exact selection (Yang et al. 2024; Merrill & Sabharwal 2024).

Without them, unbounded copying is impossible:

- Under fixed precision and soft attention, a fixed model reduces to finite-state prediction (Li & Cotterell 2025).
- With periodic/local PEs, only $O(\log n)$ bits can cross from source to output (Huang et al. 2025).

Takeaway for Framework 2

With soft attention and low precision, copying is impossible as $n \to \infty$.

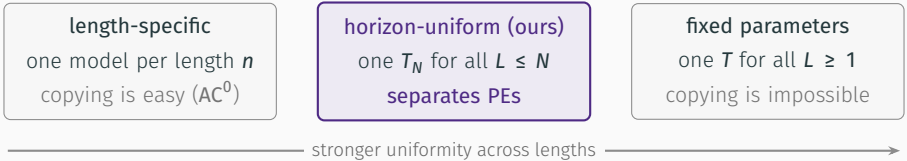
Both frameworks are degenerate here

Real models have a fixed finite context window, and inside it they copy well — except under repetition. We want a framework to capture this.

Definition (Horizon-uniform copying)

Fix a design horizon N. One model T_N with fixed parameters and precision must, for every $L \le N$ and every string, emit $s_1, \dots, s_L, \langle \text{EOS} \rangle$ and stop.

Resources may grow with N, but not with L or the string.

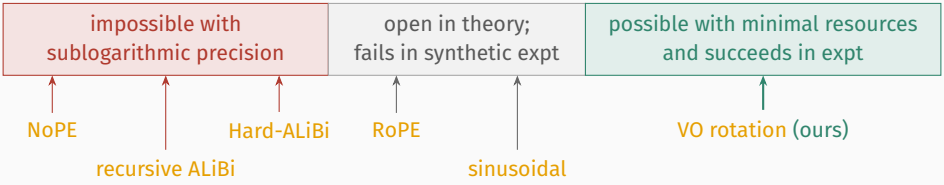


What the question becomes

For models under certain architectures and PEs,

How must the resources scale with N, in order to copy arbitrary strings of length $\leq N$?

(resources here are width, depth, and precision.)



Plan

1. Background: attention and positional encodings
2. The right expressivity question
3. Lower bounds
4. Upper bounds
5. Experiments and takeaways

Eventually-constant bias-only transformers

Defined class T_{ECB} as transformers with the following types of PEs: position enters only as an additive bias inside softmax,

$$\alpha_{t,j} = \mathrm{softmax}_{j \le t}\left(e_{t,j}\right), \qquad e_{t,j} = q_t{}^{\top} k_j - b(t-j),$$

where $b(d)$ stops changing beyond some distance. Width and the number of heads are arbitrary.

- NoPE: no bias.
- Recursive ALiBi: a linear penalty by repeatedly subtracting the same slope.
- Hard-ALiBi: a hard local window.

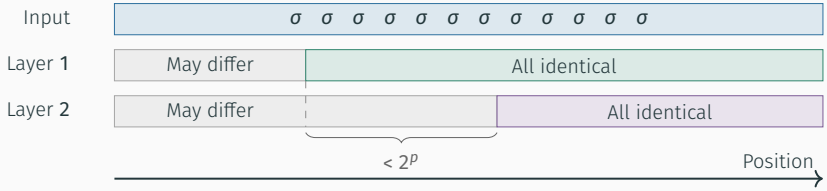
The last bullet [§1]

Hard-ALiBi $\in T_{\mathrm{ECB}}$: the mechanism that wins on random strings must fail under repetition.

Why a long run stops registering

Recall $o_t = \frac{\sum_{j \le t} e^{q_t^\top k_j} v_j}{\sum_{j \le t} e^{q_t^\top k_j}}$, summed left to right in p-bit floats. Inside a run of σ's, each new σ adds the same term to both sums. After $< 2^p$ steps, the next term is absorbed by rounding and the sums stop changing.

Layer by layer, the tail of the run becomes bit-identical:



Past a *saturation length* of $O(D \cdot 2^p)$, one more σ leaves the prediction exactly the same. The model forgets *how long the run is*.

Correct copying forces a loop

Theorem

If the saturation length is $o(N)$, then for every $L \in [N/3, N/2]$ the model fails to copy σ^L: it either makes mistakes early on, or outputs σ forever.

Proof. Steps $L - 1$ and L are both past saturation, so they make the same prediction. □

- Width does not help: rules out NoPE and Hard-ALiBi of arbitrary width and depth $o(N^{0.1})$ at $0.9 \log N$ precision.
- Matches the observed failure patterns in frontier LLMs!

Plan

1. Background: attention and positional encodings
2. The right expressivity question
3. Lower bounds
4. Upper bounds
5. Experiments and takeaways

From RoPE to VO-RoPE

Core idea: let position act on the returned values, not the weights.

VO-RoPE (Su 2025): rotate each value by $-j$ on the way in, and the output by $+t$ on the way out:

$$o_t = R^t \sum_{j \le t} \alpha_{t,j} R^{-j} v_j \xrightarrow{\text{one-hot on } j} R^{t-j} v_j.$$

Taking $v_j = \mathbf{e}_1$, the one-hot attention output $R^{t-j}\mathbf{e}_1$ actually encodes $(t - j) \bmod$ period via rotation!

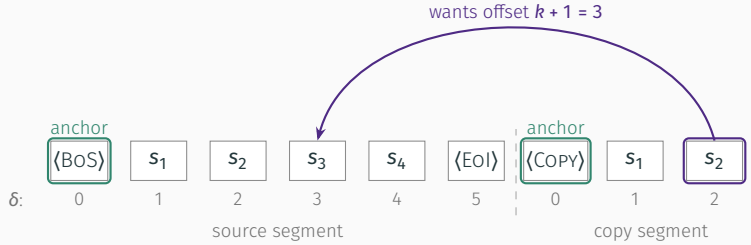
	rotates	position lands in
RoPE	(Q, K)	attention scores
VO-RoPE	(V, O)	residual stream

Additional idea: instead of heuristically chosen rotation periods, we use *small prime numbers* as periods.

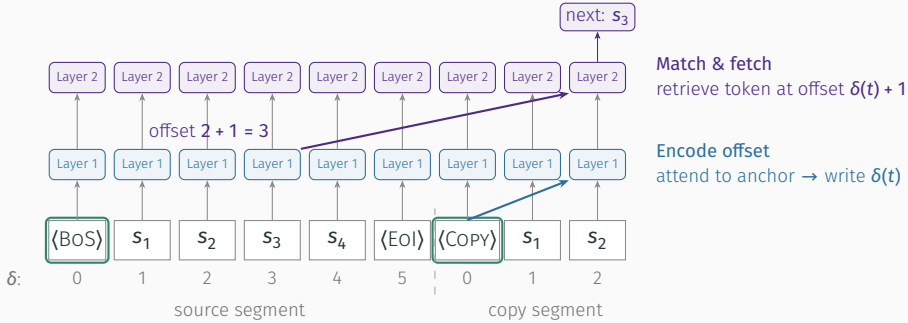
Anchors and offsets

Anchors: ⟨BOS⟩ and ⟨COPY⟩. Offset $\delta(t)$: distance from the nearest anchor.

- Source s_k has offset k; query at offset k wants offset $k + 1$.
- Offset $k + 1$ occurs only on the source side. The goal is hence to retrieve a token with this offset.



Overview: a two-layer copying transformer



Layer 1 (write the offset):

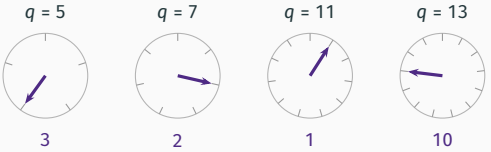
- Attend one-hot to the nearest anchor (pure content match — anchors are unique).
- Head h (prime period q_h, value $\mathbf{e}_1$) writes $\delta(t) \bmod q_h$ via $R_{q_h}^{\delta(t)} \mathbf{e}_1$.

Chinese Remainder Theorem for addressing

Distinct primes $q_1, \dots, q_H$ with $\prod_h q_h > N + 1$ identify $\delta \in \{0, \dots, N + 1\}$ uniquely from $(\delta \bmod q_1, \dots, \delta \bmod q_H)$.

$$H = \Theta\left(\frac{\log N}{\log \log N}\right), \qquad q_{\max} = \Theta(\log N).$$

Each clock has only $O(\log N)$ ticks $\implies$ $O(\log \log N)$ precision suffices.



$\delta = 23 \longmapsto (3, 2, 1, 10)$

Layer 2: retrieval

- **Query**: turn each of its own clocks forward by **+1** step (asking for offset $k + 1$).
- **Attention score** $\alpha_{t,j}$: sum unit-circle inner products across clocks $\Longrightarrow$ uniquely maximized when all residues match.



With low precision, softmax actually *underflows* to zero when residues don't match. We hence have exact hard attention selection.

The construction

Theorem

For every fixed finite vocabulary and all sufficiently large N, there is a two-layer $T_N \in T_{V0}$ that copies every string of every length $L \le N$ under arbitrary repetition and terminates correctly, with

- $H = \Theta(\log N / \log\log N)$ *prime-period heads, each with two rotary value coordinates;*
- *precision* $p = O(\log\log N)$*;*
- *dimension* $d = O(|V| + \log N / \log\log N)$*.*

Content hashing vs. position hashing

	Jelassi et al. 2024	this work
retrieves by	n-gram content	anchor-relative offset
positional channel	Hard-ALiBi (score)	VO-RoPE (value)
repetition	excluded by assumption	arbitrary
termination	not considered	same rule as copying
depth	2 layers	2 layers
width	$O(\log N)$	$O(\log N / \log\log N)$
precision	$O(\log\log N)$	$O(\log\log N)$

Plan

1. Background: attention and positional encodings
2. The right expressivity question
3. Lower bounds
4. Upper bounds
5. Experiments and takeaways

Synthetic setup

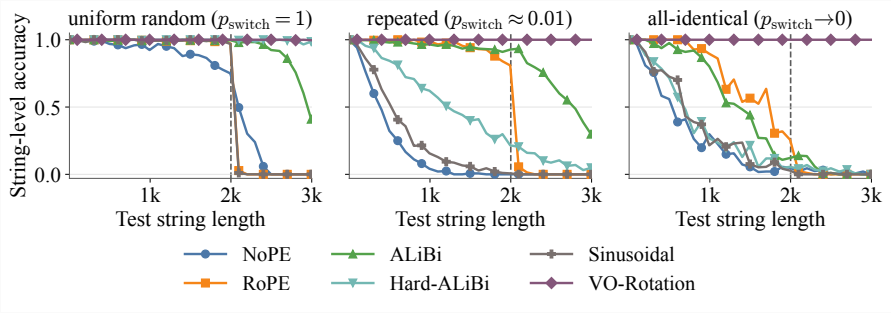
- **Task:** $\langle \text{BOS} \rangle\, s\, \langle \text{EOI} \rangle\, \langle \text{COPY} \rangle \;\mapsto\; s\, \langle \text{EOS} \rangle$; **26** lowercase letters, $L \sim \text{Unif}\{1, \ldots, 2000\}$.
- **Training mix:** **10%** i.i.d., **10%** all-identical, **80%** same-letter blocks (block length $\sim \text{Geom}(p_{\text{switch}})$, mean $1/p_{\text{switch}} \in [1, 301]$).
- **Metric:** string-level exact match (including $\langle \text{EOS} \rangle$); 3 seeds.

Examples

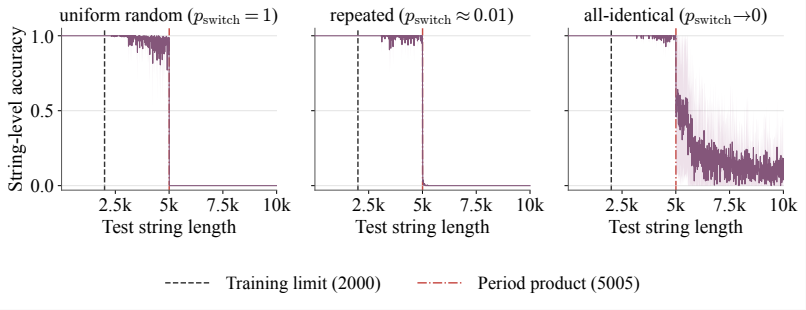
i.i.d., $L = 34$	`oryooqsgfzqputzfdojecrzwubtmouxtuf`
blocks, random $p_{\text{switch}} \approx 1/188$, $L = 362$	$\text{q}^{11}\,\text{c}^{17}\,\text{g}^{28}\,\text{z}^{271}\,\text{m}^{35}$
all-identical, $L = 14$	`eeeeeeeeeeeeee`

Near the training limit — still inside the training range

Model	Layers	Width	$p_{\text{switch}} = 1$	$p_{\text{switch}} = 0.01$	$p_{\text{switch}} \to 0$
NoPE	6	512	0.750 ± 0.244	0.000 ± 0.000	0.049 ± 0.084
RoPE	6	512	0.972 ± 0.030	0.809 ± 0.071	0.257 ± 0.367
ALiBi	6	512	0.997 ± 0.006	0.927 ± 0.038	0.122 ± 0.080
Hard-ALiBi	6	512	1.000 ± 0.000	0.215 ± 0.129	0.056 ± 0.006
Sinusoidal	6	512	0.969 ± 0.021	0.007 ± 0.006	0.028 ± 0.039
VO-RoPE	2	256	**1.000** ± 0.000	**1.000** ± 0.000	**1.000** ± 0.000



The cliff is where the theory puts it



- Trained on $L \leq 2000$; generalizes to **0.98–0.998** up to $L = 5000$.
- Collapses right at $5005 = 5 \cdot 7 \cdot 11 \cdot 13$ — where all CRT residues wrap around.

Three things to take away

1. **Repetition isolates position.** Content matching fails; termination becomes the real bottleneck.
2. **Where position enters matters.** Score-side biases may fail at *any* width; value-side VO-RoPE solves copying in 2 layers and $O(\log \log N)$ bits.
3. **Horizon-uniform expressivity.** One model per horizon N avoids both trivial (AC^0) and impossible (unbounded) extremes.

Open problems

- The RoPE gap: outside T_{ECB} yet fails empirically — lower bound or hidden construction?
- **Accumulation rules:** does saturation survive tree-structured or compensated summation?
- **Value-side optimality:** is $\Theta(\log N / \log\log N)$ heads optimal?
- **Hybrid VO/RoPE:** what do value rotations cost on semantic tasks?

AI disclaimer

- Claude, Gemini, GPT helped with drafting the paper, writing proofs, and generating code.
- But I told them specifically what to do. Ideas are almost completely human-generated.
- DeepSeek and Qwen participated as experimental subjects.
- Claude was cruelly abused during the making of the slides.

Thank you.