You Need Better Attention Priors

GOAT: learning structural defaults inside standard attention

Paper authors: Elon Litman and Gabe Guo

Research paper discussion

arXiv:2601.15380v2 | ICML 2026
Reading date: October 10, 2026

Content evidence + learned positional log-prior $\longrightarrow$ one SDPA call

Weak content still needs a routing decision

The design question

If no key is clearly relevant, should attention be uniform, local, or directed to a default token?

Method	How position enters	Structural choice
RoPE	Rotate content Q/K	Position interacts with content
ALiBi	Add a fixed distance bias	Monotone recency
GOAT	Add a learned log-prior	Relative patterns + key defaults

GOAT reserves independent coordinates for structure in the attention logits.

Paper, Secs. 1 and 4; Appendix A and F, PDF pp. 1–5, 12, 15–16.

Softmax solves a one-sided transport problem

For one query, distribute unit mass over L keys:

$$p^* = \arg\min_{p \in \Delta^{L-1}} \left[-\langle p, s \rangle - \tau H(p) \right], \qquad H(p) = -\sum_j p_j \log p_j.$$

- s_j: unscaled query–key affinity; $-s_j$: transport cost.
- $\tau > 0$: temperature; only the query-side simplex is constrained.
- There is no key marginal constraint and no Sinkhorn iteration.

The implicit prior

$\mathrm{KL}(p \| \mathcal{U}) = -H(p) + \log L$: entropy regularization is KL regularization toward a uniform distribution.

Paper, Eqs. (2)–(3), Sec. 2 and Appendix B, PDF pp. 2, 12.

A different prior gives an additive logit term

$$\min_{p \in \Delta^{L-1}} -\langle p, s \rangle + \tau \, \mathrm{KL}(p \| \pi) \quad \Longrightarrow \quad p_j^* = \frac{\pi_j e^{s_j/\tau}}{\sum_k \pi_k e^{s_k/\tau}}.$$

Closed-form attention

$$p^* = \mathrm{softmax}(s/\tau + \log \pi).$$

- Use positive π on allowed keys; apply the causal mask separately.
- A log-prior $\mathcal{K}$ needs no explicit normalization: row constants cancel.
- A 9:1 prior is balanced by a $\log 9$ content-logit advantage for key 2.

Paper, Proposition 3.1, Eqs. (4)–(10), PDF p. 3. Last bullet: teaching example.

GOAT allocates a fixed head budget to structure

$$z_{ij} = \underbrace{\frac{q_{c,i}^{\mathsf{T}} k_{c,j}}{\sqrt{d_c}}}_{\text{content}} + \underbrace{\mathcal{K}_{ij}^{\mathrm{rel}}}_{\text{relative position}} + \underbrace{u_h(j)}_{\text{key-only default}}, \qquad o_i = \sum_j \mathrm{softmax}_j(z_i) v_j.$$

Dimensions

$d_p = 2R + 2$ positional dimensions
$d_c = d_h - d_p$ content dimensions
Value vectors retain all d_h dimensions.

Example: $d_h = 64$

$R = 4$: $d_p = 10$, $d_c = 54$
$R = 8$: $d_p = 18$, $d_c = 46$
More positional capacity uses content capacity.

Paper, Sec. 4 and Algorithm 1, PDF pp. 4–5. Dimension examples follow the main-text formula.

Two coordinates realize one directional Fourier mode

$$\mathcal{K}_{ij}^{\mathrm{rel}} = \sum_{r=1}^{R} \left[\alpha_r \cos(\omega_r(i-j)) + \beta_r \sin(\omega_r(i-j))\right].$$

$$k_j^{(r)} = \begin{bmatrix} \cos \omega_r j \\ \sin \omega_r j \end{bmatrix}, \qquad q_i^{(r)} = \begin{bmatrix} \alpha_r \cos \omega_r i + \beta_r \sin \omega_r i \\ \alpha_r \sin \omega_r i - \beta_r \cos \omega_r i \end{bmatrix}.$$

- Their dot product is exactly the rth relative log-prior term.
- α_r controls symmetric patterns; β_r controls directionality.
- Signed coefficients are valid: probabilities come from exponentiation.

Paper, Eqs. (12)–(19); Appendices C–D, PDF pp. 3–4, 12–14.

One positional coordinate can express a key default

$$u_h(j) = \lambda_h j + g_h(\phi_{\mathrm{abs}}(j)) + b_h \mathbf{1}\{j = 0\}.$$

- g_h: a small SiLU MLP over sinusoidal and normalized position features.
- $\lambda_h j$: learned key-linear trend; b_h: optional first-token bump.
- Constant query coordinate 1 times key coordinate $u_h(j)$ adds the bias.

ALiBi fits into the same coordinate

For causal keys $j \le i$, $-m(i-j) = mj - mi$. The term $-mi$ is a row constant, so a key-linear bias gives the same softmax weights.

Paper, Eqs. (21)–(22), Appendix G.3–G.4, PDF pp. 4, 20–22.

Pre-scaling makes one standard SDPA call exact

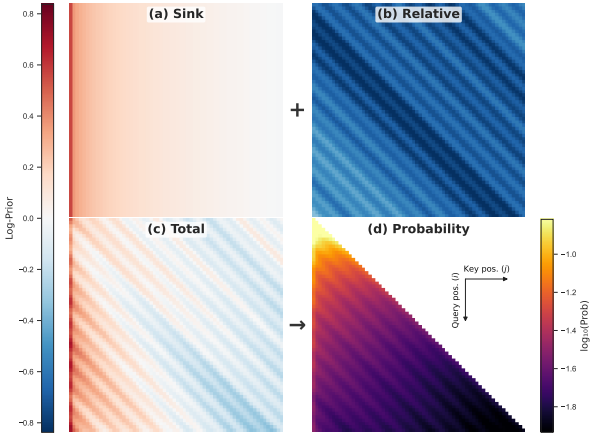
$$q'_i = \begin{bmatrix} \sqrt{d_h/d_c}\, q_{c,i} \\ \sqrt{d_h}\, q_{\mathrm{rel},i} \\ \sqrt{d_h} \\ 0 \end{bmatrix}, \qquad k'_j = \begin{bmatrix} k_{c,j} \\ k_{\mathrm{rel},j} \\ u_h(j) \\ 0 \end{bmatrix}.$$

$$\frac{(q'_i)^{\mathsf{T}} k'_j}{\sqrt{d_h}} = \frac{q_{c,i}^{\mathsf{T}} k_{c,j}}{\sqrt{d_c}} + \mathcal{K}_{ij}^{\mathrm{rel}} + u_h(j).$$

- No explicit dense positional bias matrix; standard attention kernels apply.
- Dense attention still has quadratic computation; positional features add work.

Paper, Eqs. (23)–(27) and Algorithm 1, PDF pp. 4–5. Zero coordinate is alignment padding.

A copy-mixture task separates global and relative structure



Copy the first token, copy the previous token, or sample noise.

- Key-only term: first-position preference.
- Relative term: diagonal, repeating patterns.
- Mask + normalization produce the causal prior.

Paper, Fig. 1, PDF p. 1; Sec. 6, p. 6. Columns: key position; rows: query position.

Low content contrast makes the posterior approach the prior

Now $s_{ij} = q_{c,i}^{\mathsf{T}} k_{c,j}/\sqrt{d_c}$ is a scaled logit. Define

$$\omega_i = \max_k s_{ik} - \min_k s_{ik}.$$

Theorem 5.1: collapse to prior

$$\pi_{ij} e^{-\omega_i} \le p_{ij} \le \pi_{ij} e^{\omega_i} \quad \Longrightarrow \quad p_i \to \pi_i \text{ as } \omega_i \to 0.$$

- Subtract the minimum score: the normalizer lies in $[1, e^{\omega_i}]$.
- A peaked prior yields a sink; a uniform prior yields uniform weights.
- The result describes a given prior; it does not force training to learn a sink.

Paper, Eq. (28), Sec. 5 and Appendix E.1, PDF pp. 5, 14.

The sink bound retains an explicit dependence on length

Fix Q/K. Perturb only non-sink values by $\|\Delta v_k\| \le \varepsilon$:

$$\|\Delta o_i\| \le \varepsilon \Psi, \qquad \Psi = 1 - p_{ij^*}.$$

If the prior margin is $\delta = \min_{k \neq j^*}(\mathcal{K}_{ij^*} - \mathcal{K}_{ik}) > 0$,

$$\lim_{\omega_i \to 0} \Psi \le \frac{L-1}{e^{\delta} + L - 1}.$$

What length robustness requires

To guarantee $\Psi \le \eta$ using this bound: $\delta \ge \log \frac{(1-\eta)(L-1)}{\eta}$. Fixed margin does not give a uniform guarantee as L grows.

Paper, Eqs. (31)–(34), PDF p. 6; Appendix E.2, p. 15. Rearranged bound: reading analysis.

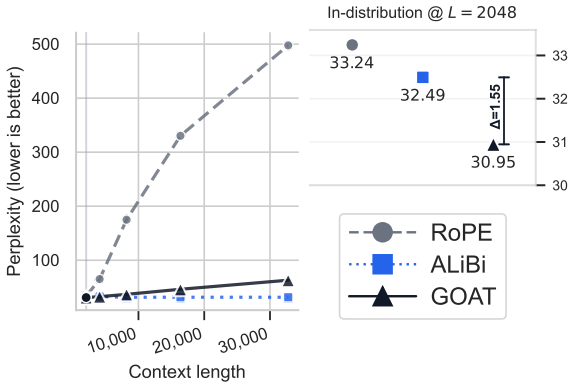
Evaluation spans language, retrieval, and spatial structure

Experiment	Setup	Main measurement
C4	125M; 4B tokens; train at 2048	PPL under length extrapolation
FineWeb	1.23B; about 3.13B tokens	PPL from 2K to 16K
Passkey / NIAH	Synthetic retrieval; 8-digit key	Span-token accuracy
Human genome	125M decoder LM	NLL, throughput, memory
ImageNet-1k	ViT-Small; train at 224^2	Top-1 at larger resolutions
Spatial VQA	Four-image CIFAR-10 scenes	Held-out scan-order accuracy

Language experiments share a training window, not a dataset or training budget.

Paper, Sec. 6; Tables 2, 8, 10, 11, PDF pp. 6–9, 23–25.

C4 improves the quality–extrapolation trade-off



At the 2048 training length:

Model	PPL ↓
RoPE	33.24
ALiBi	32.49
GOAT	**30.95**

RoPE degrades rapidly. ALiBi is flatter at very long lengths; GOAT is not best at every length.

Paper, Fig. 2(a), PDF p. 7. Caption reports 1.55 PPL gain; printed labels differ by 1.54.

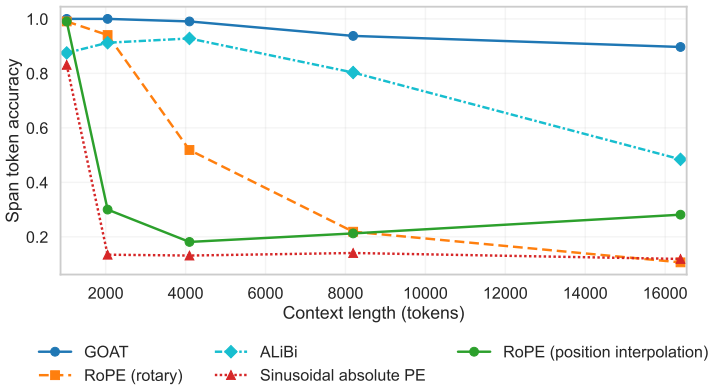
At 1.23B, the main gain appears beyond the training window

Model	2K	4K	8K	16K
RoPE	20.82	28.53	89.45	417.43
GOAT	**20.75**	**21.04**	**22.69**	**24.77**

- FineWeb; 22 layers, 32 heads, width 2048; training length 2048.
- About 3.13B tokens at the selected GOAT checkpoint.
- In-window PPL is almost unchanged; extrapolation degrades much more slowly.

Paper, Table 1, PDF p. 8; Table 8, p. 24. All entries are perplexity; lower is better.

Passkey retrieval stays strong at long distances



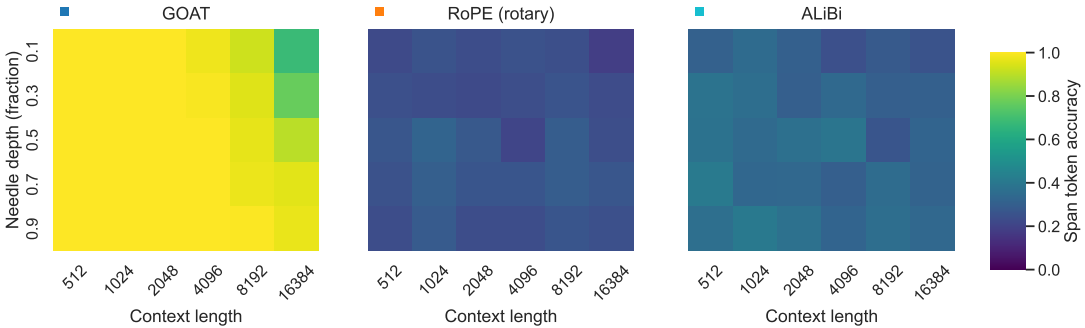
At about 16K:

- GOAT: ∼90%
- ALiBi: ∼48%
- RoPE: ∼11%

Metric: accuracy of tokens within the target span, not full-key exact match.

Paper, Fig. 3(a), PDF p. 7. Values are approximate graph readings; caption uses $L_{\text{train}} = 1024$ as an example.

NIAH gains span depths, with some loss at the longest length



GOAT is strongest across the grid. At 16K, shallow needles are visibly less reliable than those in shorter contexts.

Paper, Fig. 3(c), PDF p. 7. Selected original GOAT, RoPE, ALiBi panels and shared colorbar; data unchanged.

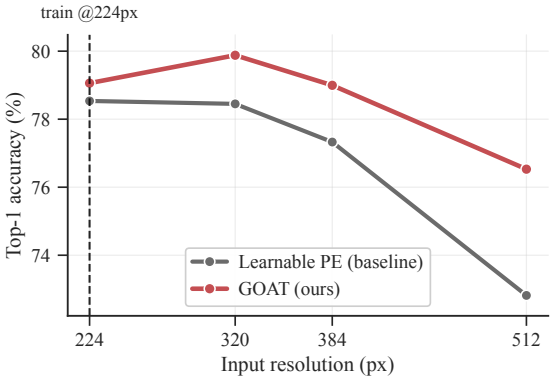
DNA modeling improves NLL and memory, with similar throughput

Metric	RoPE	GOAT	Change
Validation NLL ↓	1.2054	**1.1294**	−6.30%
Bits/base ↓	1.7390	**1.6294**	−0.1096
Training tokens/s ↑	**139,886**	138,171	−1.23%
Peak allocated memory (GB) ↓	2.86	**1.83**	−36.01%

- Human reference genome, 125M decoder models, matched training budgets.
- NLL and bits/base express the same predictive quality in different units.
- The memory result is for this setup; it is not a fixed saving of the mechanism.

Paper, Fig. 4(a)–(b), PDF p. 9. Relative changes calculated from displayed values.

A 2D prior transfers better to higher image resolutions



- ViT-Small, patch size 16, 150 epochs at 224×224.
- At 512: about 76.6% vs. 72.8% Top-1, without fine-tuning.
- Axis-wise features impose relative 2D structure; learning sets its shape.

Paper, Fig. 5(a), PDF p. 8; Table 11, p. 25. Accuracy values are approximate graph readings.

Canonical coordinates help when image scan order changes

Four CIFAR-10 images on a canonical 2×2 grid; test scan orders were unseen during training.

Method	Answer NLL ↓	Exact answer accuracy ↑
No PE	0.359	17.2%
RoPE	0.367	16.5%
GOAT	**0.350**	**20.3%**

- 50k/10k train/test scenes; answer-only spatial questions.
- GOAT gains 3.8 percentage points over RoPE; absolute accuracy remains low.

Paper, Tables 9–10, PDF pp. 24–25. Image tokens use 2D coordinates; text uses 1D positions.

More Fourier modes mainly improve extrapolation

R	d_p	PPL 2K	PPL 4K	PPL 8K
1	4	35.49	34.81	38.28
2	6	35.52	34.95	38.64
4	10	35.49	34.57	35.97
8	18	**35.18**	**34.15**	**35.15**

- $R = 2 \rightarrow 8$: 8K PPL improves by 3.49; 2K PPL improves by 0.34.
- The head budget is fixed: larger R reduces content Q/K dimensions.
- These ablation PPLs differ from Fig. 2; compare within this table.

Paper, Table 3, PDF p. 23. Main-figure and ablation configurations are not fully reconciled in the paper.

Learned structure is directional and stronger in early layers

1.23B spectral statistics

- 287/704 heads have amplitude > 1.
- 662/704 have dominant antisymmetric coefficients.
- Median $|\beta|/|\alpha| = 1.84$.
- Mean amplitude: 2.04 in layer 0, 0.70 in layer 21.

Training loss checkpoints

	RoPE	GOAT
C4, 15k steps	3.535	3.520
FineWeb, 10k	3.003	2.993

Reported runs show no clear early optimization penalty.

Paper, Tables 6–7, PDF p. 24. Spectral statistics describe learned heads; they are not component-removal ablations.

The useful contribution is an explicit, efficient routing prior

1. **Mechanism:** learn structural logits separately from content, then factor them into ordinary Q/K coordinates.
2. **Evidence:** large extrapolation gains over the tested RoPE baseline, including at 1.23B; gains also appear in retrieval and spatial tasks.
3. **Trade-offs:** positional capacity uses content dimensions; sink bounds need sufficient margin; dense attention cost remains.

A productive next comparison

Hold content capacity and training budget fixed, compare adapted long-context baselines, and ablate relative directionality versus key-only defaults.

Synthesis of Secs. 4–6 and Appendices G–K. Last block: reading analysis.

Paper and discussion materials

Elon Litman and Gabe Guo.

You Need Better Attention Priors.

ICML 2026; arXiv:2601.15380v2, revised August 23, 2026.

- Versioned paper: `https://arxiv.org/abs/2601.15380v2`
- Author code link: `https://github.com/elonlit/goat`
- Companion Chinese notes: `reading-notes.zh.md`

All experiment results and original plots are from this paper.
Supplementary derivations and technical qualifications follow.

Backup: finite rank and boundedness imply finite frequencies

Assume a relative log-prior with fixed d_p-dimensional positional features:

$$\mathcal{K}_{ij} = \kappa(i-j) = \langle \varphi(i), \psi(j) \rangle, \qquad \sup_{d \in \mathbb{Z}} |\kappa(d)| < \infty.$$

1. Translates $\kappa(\cdot - j)$ span a finite-dimensional shift-invariant space.
2. The shift operator yields exponential-polynomial sequences.
3. Boundedness in both directions excludes growth and nonconstant polynomials.

$$\kappa(d) = \sum_{m=1}^{M} c_m e^{\mathrm{i}\theta_m d}, \qquad M \le d_p.$$

This class result does not select the geometric frequency grid, nor constrain all of $u_h(j)$.

Paper, Appendix G.1–G.2, Theorem 2, PDF pp. 16–19.

Backup: a special frequency needs separate dimension counting

The real-mode statement in Eq. (73) gives $R \leq \lfloor d_p/2 \rfloor$.

Counterexample to that bound as stated

On integer positions, $\kappa(d) = (-1)^d$ is bounded and nonconstant, but

$$\kappa(i-j) = (-1)^i(-1)^j$$

uses just one dimension. For $d_p = 1$, the printed bound would require $R = 0$.

- Nondegenerate conjugate frequencies use two real dimensions per pair.
- Constant and Nyquist (π) modes each use only one dimension.
- The finite-frequency conclusion and GOAT's ordinary mode construction survive.

Paper, Appendix G.2, Eq. (73), PDF p. 17. Counterexample and correction: reading analysis.

Backup: a mean-lag constraint gives a geometric prior

Let $d = i - j \in \{0, \dots, L-1\}$ and prescribe mean lag μ:

$$\max_q H(q) \quad \text{subject to} \quad \sum_d q_d = 1, \quad \sum_d d q_d = \mu.$$

$$q_d^* = \frac{e^{-\lambda d}}{Z(\lambda)}, \qquad \frac{d}{d\lambda}\mathbb{E}_{q^*}[d] = -\operatorname{Var}_{q^*}(d) < 0.$$

- For $L \geq 2$ and $0 < \mu < L - 1$, the required λ is unique.
- Recency corresponds to $\lambda > 0$; $\lambda = 0$ gives the uniform prior.
- $\log q^*_{i-j} = \lambda j + c_i$: a key-only trend is sufficient.

Paper, Appendix G.3, Theorem 3 and Proposition G.6, PDF pp. 20–21.

Backup: frequency and initialization changes are modest

Frequency grid

Setting	2K PPL	8K PPL
Base 1k	35.48	38.42
Base 10k	35.52	38.64
Base 100k	35.62	38.51
Learned	**35.39**	**37.99**

Initialization

Setting	2K PPL	8K PPL
Zero	**35.52**	38.64
Random	35.54	38.88
ALiBi start	35.56	**38.48**

Changes across broad fixed grids are smaller than the mode-count effect. Learning frequencies or using a recency warm start brings a small gain.

Paper, Tables 4–5, PDF p. 23.

Backup: details that matter when implementing the paper

- **2D formula:** Eq. (20) omits the plus signs between cosine and sine terms. Axis-wise concatenation implements the additive form of Eq. (12).
- **Dimension names:** C4 Table 2 reports $d_{\text{pos}} = 12$, Pos Rank 2; the mapping to $d_p = 2R + 2$ is not explained.
- **C4 protocol:** Fig. 2 and ablation PPLs differ; Table 2's batch and Table 7's tokens/step require an effective-batch explanation.
- **Sink behavior:** a key-only bias changes routing. It neither removes the sink token nor makes its value vector zero.

Paper, Eqs. (12), (20), Algorithm 1, Tables 2–7, PDF pp. 3–5, 23–24. Qualifications: reading analysis.