PaTH Attention

Position Encoding via Accumulating Householder Transformations

Paper authors: Songlin Yang, Yikang Shen, Kaiyue Wen, Shawn Tan, Mayank Mishra, Liliang Ren, Rameswar Panda, Yoon Kim

NeurIPS 2025 | arXiv:2505.16381v2 (3 Feb 2026)
Independent paper discussion | 13 Oct 2026

Let the events between two tokens transform how they match.

The path between tokens can carry state

A concrete question

After writes, distractors, and swaps, which earlier record is still relevant?

$$\text{RoPE logit} = \mathbf{q}_i^\top \mathbf{R}^{i-j} \mathbf{k}_j$$

- Queries and keys depend on content; the positional transform depends on distance.
- State tracking requires composing ordered events between the endpoints.
- PaTH makes each step of that transformation depend on the input.

Paper, Secs. 1–2, PDF pp. 1–3.

PaTH inserts a learned transformation along every path

$$\mathbf{H}_t = \mathbf{I} - \beta_t \mathbf{w}_t \mathbf{w}_t^\top, \qquad \beta_t = 2\,\mathrm{sigmoid}(\mathbf{u}^\top \mathbf{x}_t + b), \quad \|\mathbf{w}_t\|_2 = 1$$

Causal attention, with product order explicit

$$\widetilde{A}_{ij} = \mathbf{k}_j^\top \underbrace{\mathbf{H}_{j+1}\mathbf{H}_{j+2}\cdots\mathbf{H}_i}_{\text{content-dependent path}} \mathbf{q}_i, \qquad j \le i$$

$$A_{ij} = \mathrm{softmax}_{j\le i}(\widetilde{A}_{ij}), \qquad \mathbf{o}_i = \sum_{j\le i} A_{ij}\mathbf{v}_j$$

$\mathbf{w}_t$: low-rank projection $\rightarrow$ width-3 convolution $\rightarrow$ L_2 normalization.

Paper, Sec. 2.2 and footnote 2, PDF pp. 2–3. Attention scale omitted as in the paper.

One direction can shrink, disappear, or change sign

For $\mathbf{z} = \mathbf{z}_\perp + \alpha \mathbf{w}_t$,

$$\mathbf{H}_t \mathbf{z} = \mathbf{z}_\perp + (1 - \beta_t)\alpha \mathbf{w}_t.$$

β_t	Along $\mathbf{w}_t$	Interpretation
$(0, 1)$	Positive contraction	Attenuate
1	Zero	Erase
$(1, 2)$	Negative contraction	Flip and attenuate
2	Exact reflection	Idealized proof endpoint

Different directions generally do not commute: event order matters.

Parameterization: Sec. 2.2, PDF p. 2. Eigenvalue interpretation derived here. Exact $\beta = 2$: Appendix A, p. 17.

PaTH keeps a recurrent path and a softmax readout

A related linear RNN maintains a matrix state:

$$\mathbf{S}_t = \mathbf{S}_{t-1}\mathbf{H}_t + \mathbf{v}_t\mathbf{k}_t^\top, \qquad \mathbf{o}_t = \mathbf{S}_t\mathbf{q}_t.$$

Define $c_{tj} = \mathbf{k}_j^\top(\mathbf{H}_{j+1}\cdots\mathbf{H}_t)\mathbf{q}_t$.

Linear recurrence, unrolled

$$\mathbf{o}_t = \sum_{j\le t} \mathbf{v}_j c_{tj}$$

History is compressed into $\mathbf{S}_t$.

PaTH attention

$$\mathbf{o}_t = \sum_{j\le t} \mathbf{v}_j \frac{e^{c_{tj}}}{\sum_{r\le t} e^{c_{tr}}}$$

Explicit historical values remain available.

Paper, Sec. 2.2, PDF p. 3. This connection does not give PaTH a fixed-size recurrent cache.

A reflection can implement a symbolic swap

$$\mathbf{w} = \frac{\mathbf{e}_a - \mathbf{e}_b}{\sqrt{2}}, \quad \beta = 2 \quad \Longrightarrow \quad (\mathbf{I} - 2\mathbf{w}\mathbf{w}^\top) \text{ swaps coordinates } a, b.$$

Theorem 2.1

A one-layer, two-head PaTH Transformer with $\log n$ precision can solve an NC^1-complete problem under AC^0 reductions.

- Compose swaps of five elements; test whether the final permutation is identity.
- The separation from TC^0 assumes $\mathsf{TC}^0 \neq \mathsf{NC}^1$.
- The construction uses the ideal reflection endpoint; learned performance is tested separately.

Paper, Theorem 2.1, PDF p. 3; construction in Appendix A, pp. 17–18.

FoX adds a separate mechanism for forgetting

$$f_t = \text{sigmoid}(\mathbf{u}_f^\top \mathbf{x}_t + b_f) \in (0, 1)$$

PaTH-FoX attention

$$A_{ij} \propto \underbrace{\left(\prod_{s=j+1}^{i} f_s \right)}_{\text{forgetting}} \exp\left(\underbrace{\mathbf{k}_j^\top \mathbf{H}_{j+1} \cdots \mathbf{H}_i \mathbf{q}_i}_{\text{path-dependent matching}} \right)$$

- PaTH changes the geometry of matching.
- FoX attenuates older unnormalized attention weights.
- The scalar product of gates is outside the exponential.

Paper, Sec. 2.2, PDF p. 3.

A triangular representation makes paths computable

Stack $\mathbf{w}_t^\top$ into $\mathbf{W}$ and define $\mathbf{D} = \mathrm{diag}(\beta_t)$.

$$\mathbf{U} = (\mathbf{I} + \mathrm{strictLower}(\mathbf{D}\mathbf{W}\mathbf{W}^\top))^{-1}\mathbf{D}$$

$$\mathbf{H}_L \cdots \mathbf{H}_1 = \mathbf{I} - \mathbf{W}^\top \mathbf{U}\mathbf{W}$$

All causal logits from shared matrix operations

$$\widetilde{\mathbf{A}} = \mathrm{lower}(\mathbf{Q}\mathbf{K}^\top) - \mathrm{lower}(\mathbf{Q}\mathbf{W}^\top)\mathbf{U}\,\mathrm{strictLower}(\mathbf{W}\mathbf{K}^\top)$$

Use local triangular solves inside blocks; avoid a global $L \times L$ solve.

Paper, Secs. 3.1–3.3, PDF pp. 4–5. $\mathbf{U}$ denotes the paper's $\mathbf{T}^{-1}$; product orientation made explicit here.

A right-to-left scan accumulates cross-block paths

Precompute boundary-adjusted $\overleftarrow{\mathbf{Q}}_b$, $\overrightarrow{\mathbf{K}}_b$, and forward block product $\mathbf{F}_b$.

1. Compute the causal diagonal block with the local formula.
2. Set $\mathbf{Q}_{\text{work}} = \overleftarrow{\mathbf{Q}}_b$.
3. For earlier blocks $c = b-1, \ldots, 0$:

$$\widetilde{\mathbf{A}}_{b,c} = \mathbf{Q}_{\text{work}} \overrightarrow{\mathbf{K}}_c^{\top} \quad \longrightarrow \quad \text{online softmax and value accumulation}$$

$$\mathbf{Q}_{\text{work}} \leftarrow \mathbf{Q}_{\text{work}} \mathbf{F}_c^{\top}$$

Read this block first, then cross its transition. Query blocks can execute in parallel; the full attention matrix stays out of HBM.

Paper, Sec. 3.3, PDF p. 5. $\mathbf{F}_c = \mathbf{H}_{\text{first}} \cdots \mathbf{H}_{\text{last}}$ fixes orientation and the missing identity term.

Decoding updates historical keys in place

$$\mathbf{k}_j^{(t)} = \mathbf{k}_j^{(t-1)} - \beta_t \mathbf{w}_t (\mathbf{w}_t^\top \mathbf{k}_j^{(t-1)}), \quad j < t$$

$$(\mathbf{k}_j^{(t)})^\top \mathbf{q}_t = \mathbf{k}_j^\top \mathbf{H}_{j+1} \cdots \mathbf{H}_t \mathbf{q}_t$$

- The transformed keys feed a standard softmax decoding readout.
- No separate cache of all past transition vectors is needed.
- Prefill must align keys to the prompt endpoint before decoding begins.
- Historical-key updates add $O(td)$ work per head and rewrite the cache.

Paper, Sec. 3.4, PDF p. 6, Eq. (1). Complexity observation follows from the rank-one update.

The experiments ask three distinct questions

Question	Evidence	Setting
State tracking	FFLM; A_5; MQRAR-N	Small Transformers
Language modeling	Perplexity; commonsense; long context	760M; 50B tokens; 4K context
Model conversion	Math; code; general tasks	SmolLM2-1.7B; Qwen2.5-7B-Instruct

RoPE: static position transform. FoX: learned forgetting.

SBA: ordered stick-breaking weights, evaluated on synthetic tasks.

Paper, Sec. 4, PDF pp. 7–10; training details in Appendix B, p. 20.

PaTH tracks sparse writes through heavy distraction

FFLM: return the bit from the most recent `write` operation.

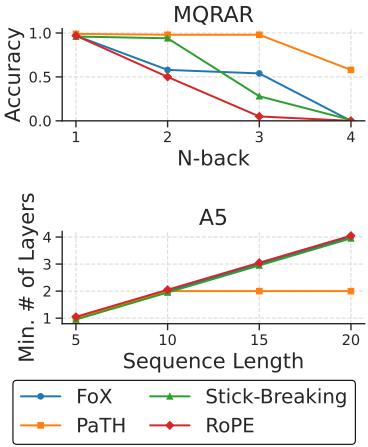
`w 1 r 1 w 0 i 1 i 0 i 1 r` $\longrightarrow$ `0`

Error rate (%) ↓	ID	OOD sparse	OOD dense
RoPE	6.9	40.3	0.01
SBA	9.6	38.9	0
FoX	8.3	36.3	0
PaTH	0	0.0001	0

All models: 1 layer, 2 heads, 64 dimensions.
Sparse: 98% ignore operations. Dense: 10% ignore operations.

Paper, Table 1 and Sec. 4.1, PDF p. 7. 0.0001% is an error fraction of 10^{-6}.

Ordered recall improves, and fewer layers solve swaps



Top: MQRAR-N
Recall the N-th previous assignment.
PaTH stays strong through $N = 3$ and degrades at $N = 4$.

Bottom: A_5 word problem
Minimum depth for $> 90\%$ accuracy.
At length 20, PaTH uses about 2 layers; baselines use about 4.

MQRAR: 2 layers, width 256, 2 heads; 32 KV pairs, length 768.

Paper, Fig. 2 and Sec. 4.1, PDF pp. 7–8. Original figure; approximate layer counts read from plot.

Language modeling improves at a fixed token budget

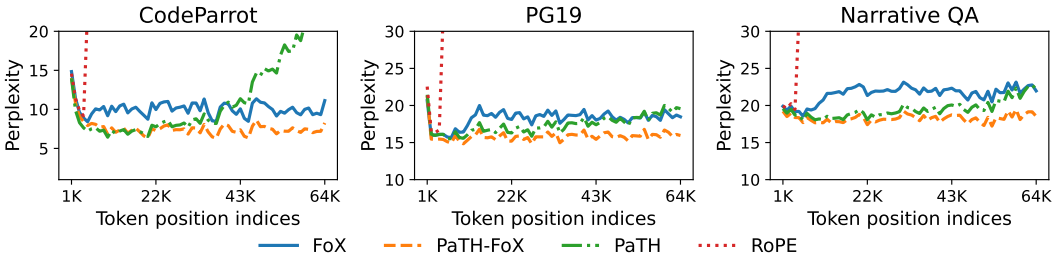
760M parameters; 50B FineWeb-Edu tokens; context length 4096.

Model	Wiki. ppl↓	LMB. ppl↓	LMB. acc↑	Avg.↑
RoPE	19.01	19.77	40.4	52.7
FoX	18.33	18.28	41.7	53.1
PaTH	18.03	16.79	44.0	**54.2**
PaTH-FoX	**17.35**	**16.23**	**44.1**	**54.2**

- PaTH: +1.5 percentage points in the six-task average over RoPE.
- PaTH-FoX lowers perplexity further, with the same rounded average.
- FoX remains better than PaTH on PIQA and WinoGrande.

Paper, Table 2, PDF p. 8. Average: six accuracy metrics, excluding perplexity; HellaSwag and ARC-c use acc_n.

Forgetting stabilizes extrapolation beyond 4K



- Horizontal axis: token position, up to 64K. Lower perplexity is better.
- FoX and PaTH-FoX remain usable to 64K; the combination has lower perplexity.
- PaTH alone deteriorates beyond roughly 32K, most visibly in code.

Paper, Fig. 3 and Sec. 4.2, PDF p. 8. Original figure; models trained at 4096 context length.

Dense retrieval benefits strongly from the combination

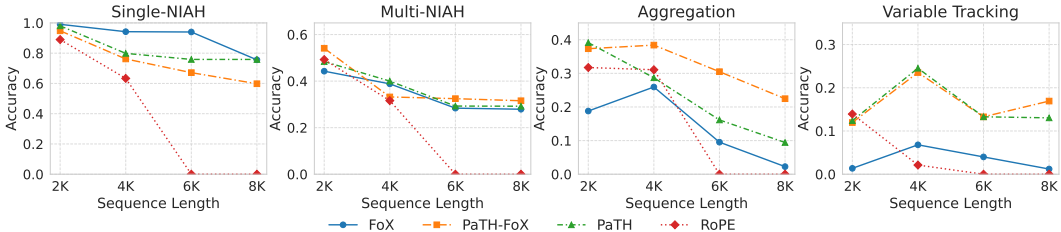
	RULER ↑		PhoneBook ↑	
Model	8K	16K	4K	8K
RoPE	1.3	0.0	15.6	0.0
FoX	29.5	4.9	38.5	17.7
PaTH	**34.8**	18.7	20.8	0.0
PaTH-FoX	34.0	**22.6**	**93.8**	**66.6**

A useful distinction

PaTH's state-tracking gains do not automatically yield dense exact retrieval.
At 8K PhoneBook, PaTH scores 0.0; PaTH-FoX scores 66.6.

Paper, Table 3, PDF p. 9. RULER averages 13 zero-shot tasks; training context is 4K.

The advantage differs across RULER task categories



- FoX leads at some lengths on Single-NIAH.
- PaTH variants clearly improve Variable Tracking over FoX and RoPE.
- Variable Tracking remains difficult: the best curve reaches only about 0.25.

Paper, Fig. 4, PDF p. 9. Original figure. Panel vertical scales differ; all report accuracy.

Narrative reasoning and mixed tasks show a tradeoff

	BABILONG ↑		LongBench-E ↑	
Model	8K	16K	4–8K	8–16K
RoPE	0.0	0.0	3.7	2.0
FoX	8.2	4.4	16.9	11.7
PaTH	16.8	**11.6**	**22.5**	14.4
PaTH-FoX	**19.2**	10.0	21.8	**16.1**

- PaTH leads on BABILONG at 16K; the combination leads at 8K.
- LongBench-E columns are length buckets, with mixed task scores.
- Which distant facts should be forgotten depends on the task.

Paper, Table 3, PDF p. 9. Table 7 gives PaTH-FoX 21.9 rather than 21.8 for 4–8K; this slide uses Table 3.

SmolLM2 conversion improves code after further training

100M tokens: attention-output MSE → 3B tokens: output KL
→ 21B tokens: DCLM / Python-Edu / MegaMathWeb (1:1:1).

SmolLM2-1.7B	GSM8K	HumanEval	MBPP+
Base, before further training	8.6	16.4	38.6
RoPE, after further training	19.9	23.1	47.1
FoX, converted and trained	15.5	21.3	48.2
PaTH, converted and trained	**20.1**	**25.6**	**51.3**

PaTH over continued RoPE: +0.2 / +2.5 / +4.2 points.
The starting checkpoint precedes WSD decay and high-quality math/code training.

Paper, Sec. 4.3 and Table 4, PDF p. 9. Base-to-final gains also include the additional training data.

A mature Qwen model gives mixed conversion results

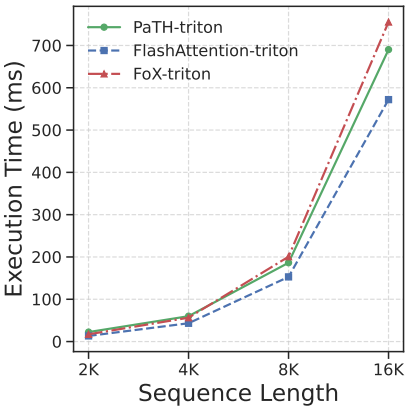
Qwen2.5-7B-Instruct: distillation only, without math/code continued pretraining.

Task	RoPE teacher	PaTH student	Change (points)
MMLU	74.21	73.28	−0.93
WinoGrande	71.51	68.90	−2.61
TheoremQA	18.12	21.88	+3.76
GSM8K	80.29	80.67	+0.38
MATH	69.10	65.38	−3.72
HumanEval	82.32	77.44	−4.88

Earlier conversion is an author hypothesis; model history and training protocols also differ.

Paper, selected rows of Table 5 and Sec. 4.3, PDF pp. 9–10.

The kernel remains quadratic and adds practical cost



Single H100; batch 32;
32 heads; head dimension 64.

- PaTH is slower than FlashAttention-triton.
- It is faster than FoX-triton at longer lengths.
- With $B \approx d$, attention stays $O(L^2 d)$.

Larger heads and full decoding services require separate evaluation.

Paper, Fig. 1, Sec. 3.3 and Sec. 6, PDF pp. 5–6, 10. Original figure; block-count correction in backup.

PaTH connects state transitions with explicit retrieval

What to retain

Content-dependent, noncommuting path products improve state tracking.
FoX adds complementary forgetting for several long-context tasks.

- **Compute:** quadratic attention, extra block transforms, and historical-key writes.
- **Numerics:** BF16 products can become unstable; clip β below 2.
- **Geometry:** a single rank-one reflection-like step does not span arbitrary RoPE rotations.
- **Transfer:** gains depend on the task and conversion protocol.

Paper, Secs. 4, 6–7, PDF pp. 7–10.

References and materials

Songlin Yang et al.
PaTH Attention: Position Encoding via Accumulating Householder Transformations.
NeurIPS 2025; arXiv:2505.16381v2, revised 3 February 2026.

`https://arxiv.org/abs/2505.16381v2`

Original figures and numerical results are from this version.
All page locators refer to the PDF, counted from page 1.

The paper links its implementation in Flash Linear Attention:
`https://github.com/fla-org/flash-linear-attention`

Backup: exact block equations, complexity, proof ideas,
BABILONG details, LongBench-E examples, and training settings.

Backup: fix orientation before implementing UT products

Let $\mathbf{U}_b = (\mathbf{I} + \mathrm{strictLower}(\mathbf{D}_b\mathbf{W}_b\mathbf{W}_b^\top))^{-1}\mathbf{D}_b$.

$$\overleftarrow{\mathbf{Q}}_b = \mathbf{Q}_b - \mathrm{lower}(\mathbf{Q}_b\mathbf{W}_b^\top)\mathbf{U}_b\mathbf{W}_b$$

$$\overrightarrow{\mathbf{K}}_b = \mathbf{K}_b - (\mathbf{U}_b\,\mathrm{strictLower}(\mathbf{W}_b\mathbf{K}_b^\top))^\top\mathbf{W}_b$$

$$\mathbf{F}_b = \mathbf{H}_{\mathrm{first}} \cdots \mathbf{H}_{\mathrm{last}} = \mathbf{I} - \mathbf{W}_b^\top\mathbf{U}_b^\top\mathbf{W}_b$$

- A lower-triangular UT factor without the transpose gives the *reverse* product.
- The paper's block product on p. 5 omits $\mathbf{I}-$; its orientation also needs care.
- Apply a causal $-\infty$ mask before softmax, not a zero mask.

Derived with explicit forward order from Secs. 3.1–3.3, PDF pp. 4–5; notation corrected for consistency.

Backup: count the cross-block transitions explicitly

Cost per query/key block pair:

$$O(B^2 d + B d^2).$$

Across $O((L/B)^2)$ causal pairs:

$$\boxed{O\left(L^2 d + \frac{L^2 d^2}{B}\right)}.$$

- Local solves and boundary transforms: $O(LB^2 + LBd)$.
- Explicit dense block products add $O(Ld^2)$ preprocessing.
- $B \approx d$ retains the $O(L^2 d)$ dominant scaling.

The paper prints Ld^2/B in the second term; multiplication by the number of block pairs gives $L^2 d^2/B$.

Derivation from the per-block cost in Sec. 3.3, PDF pp. 5–6.

Backup: two heads distinguish the identity permutation

Compose five-element swaps using exact Householder reflections.

$$s_0 = n\left(\sum_{r=1}^{5} r\pi(r) - 54.5\right)$$

$s_0 = n/2$ for identity, $\quad s_0 \le -n/2$ otherwise.

Head 1: path-dependent

$$a_0 = \frac{e^{s_0}}{e^{s_0} + n}$$

Only the start token has a nonzero value.

Head 2: uniform reference

$$a_{\mathrm{ref}} = \frac{1}{n+1}$$

All logits are zero.

Output the sign of $a_0 - a_{\mathrm{ref}}$.

Paper, Appendix A, PDF pp. 17–18. Equality/inequality signs above follow directly from the displayed score.

Backup: the flip-flop construction erases older writes

Theorem A.1, as stated in the paper

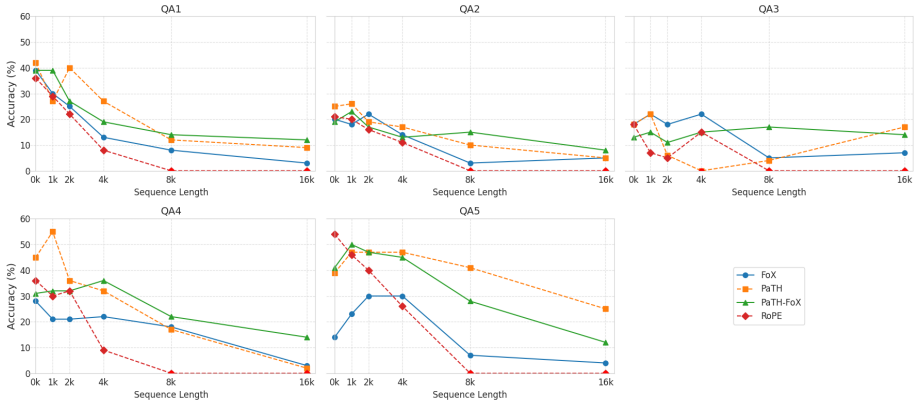
For inputs up to length n, a two-layer PaTH Transformer with $O(\log n)$ precision achieves FFLM accuracy above $1 - 1/n^{100}$.

1. Layer 1 associates each bit with its preceding instruction.
2. Layer 2 inserts a projection at a new write, blocking older write paths.
3. The read position selects the latest surviving write value.

The coordinate formulas on p. 19 contain mismatched indices: write bits are in h_{10}, h_{11}, while later value/gate equations use h_8, h_9. The two final output tests are also identical. This slide explains the intended construction.

Paper, Appendix A, PDF pp. 18–19. The empirical one-layer results are separate (Table 1, p. 7).

Backup: BABILONG gains vary by reasoning skill



QA1–QA5 cover facts, multi-fact chains, and argument relations.

QA3 stays difficult; PaTH is particularly strong on several QA5 settings.

Paper, Fig. 5 and Table 6, PDF p. 20. Original figure. Main text says QA0–QA5; appendix shows QA1–QA5.

Backup: LongBench-E averages hide task-specific effects

Dataset	Length bucket	PaTH	PaTH-FoX
gov_report	0–4K	22.3	10.2
triviaqa	8–16K	18.4	31.6
hotpotqa	8–16K	8.8	7.7
repobench-p	8–16K	19.2	19.3

- PaTH can retain useful details for short-bucket summarization.
- Forgetting helps some longer-bucket tasks substantially.
- The long-bucket average does not imply a win on every task.

Reporting difference: Table 3 gives PaTH-FoX averages 23.4 / 21.8 / 16.1; Table 7 gives 23.5 / 21.9 / 16.1.

Paper, Tables 3 and 7, PDF pp. 9, 21. Selected task-specific scores; higher is better.

Backup: pretraining and extrapolation conditions

760M training	Reported setting
Data / tokenizer	FineWeb-Edu / Mistral
Tokens / context / batch	50B / 4096 / 2M tokens
Optimizer	AdamW; weight decay 0.01
Learning rate	Peak 10^{-3}; start/end 3×10^{-5}
Schedule	Cosine; 1B-token warmup
Gradient clipping / init std.	1.0 / 0.02
Hardware / reported duration	8 H100 GPUs / 2–3 days per model

Extrapolation corpora: CodeParrot, PG-19, NarrativeQA.
Long-context tasks: RULER, BABILONG, PhoneBook, LongBench-E.

Paper, Sec. 4.2, PDF pp. 8–9; Appendix B, PDF p. 20.