MultiHashFormer: Hash-based Generative Language Models

Research group paper discussion

Paper: Huiyin Xue, Atsuki Yamaguchi, Nikolaos Aletras

arXiv:2606.28057v2 EMNLP 2026
Reading presentation: 22 October 2026

A reversible hash interface enables token generation

Research question

Can a causal LM share a fixed set of bucket embeddings across a growing vocabulary while still generating an unambiguous next token?

- Encode each token as a unique multi-hash signature.
- Decode its coordinates, then normalize over registered tokens.
- At 1B/3B, the main variant improves many benchmark scores; a 32K-to-48K vocabulary expansion adds no model parameters.

Paper, Abstract; Secs. 1, 3, 5; PDF pp. 1–8.

A single hash bucket cannot recover a token

Single-ID hashing

$$\texttt{cat} \mapsto 40, \qquad \texttt{map} \mapsto 40$$

Predicting bucket 40 leaves two possible words.

Multi-ID hashing

$$\texttt{cat} \mapsto (12, 40, 56)$$

$$\texttt{map} \mapsto (99, 40, 3)$$

Coordinate collisions are allowed; the full signatures differ.

The output interface must be invertible on the registered vocabulary.

Paper, Sec. 1 and Fig. 1; PDF pp. 1–2. Example from the paper.

Signature capacity grows combinatorially

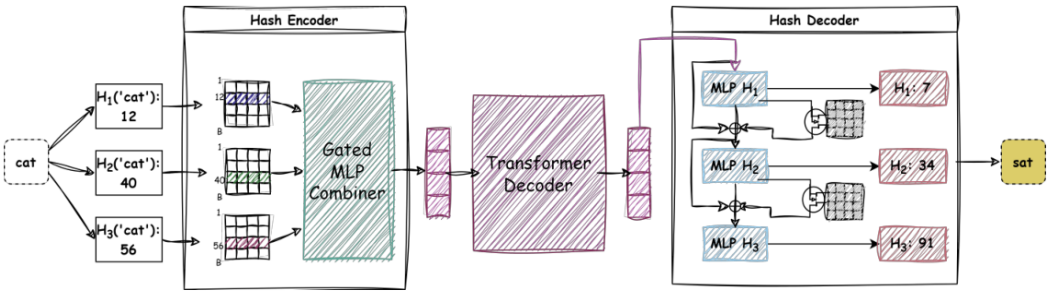
- H independent MMH3 hashes; B buckets per hash, with bucket 0 reserved for padding.
- For non-padding tokens, $\mathcal{H}_i(w) = \mathrm{MMH3}(w, \mathrm{seed}_i) \bmod (B-1) + 1$.
- The authors resolve a full-signature collision by rehashing until an unused signature is assigned.

Different scaling laws

Non-padding signature combinations: $(B-1)^H$; learned bucket embeddings: HBd parameters.

Paper, Sec. 3.1; PDF p. 3. Capacity is an upper bound on possible signatures.

The interface surrounds a standard Transformer decoder



- Hash Encoder combines the input buckets into one d-dimensional vector.
- Hash Decoder reconstructs the next token's signature after the backbone.

Paper, Fig. 1; PDF p. 2. Three hash heads are illustrative; main experiments use H=3 or 4.

The encoder learns how to combine shared buckets

$$u_i = E^{(i)}_{\mathcal{H}_i(w),:}, \quad a_i = \sigma(u_i W_1) W_2, \quad \tilde{\alpha}_i = \mathrm{softmax}(a)_i$$

$$e(w) = \left(\sum_{i=1}^{H} \tilde{\alpha}_i u_i \right) W_s$$

- $E^{(i)} \in \mathbb{R}^{B \times d}$; W_1 reduces to a 64-dimensional bottleneck in the experiments.
- The gate weights the buckets of the current token; the Transformer adds sequence context afterward.

Paper, Secs. 3.1–3.2 and 4.1; PDF pp. 3, 5.

A soft cascade predicts the next token's coordinates

$$c^{(1)} = h_t, \qquad o^{(i)} = E^{(i)} c^{(i)}, \qquad p^{(i)} = \mathrm{softmax}(o^{(i)})$$

$$\bar{e}^{(i)} = E^{(i)\top} p^{(i)}, \qquad c^{(i+1)} = c^{(i)} + W_{\mathrm{up}}^{(i)\top} \sigma\left(W_{\mathrm{dn}}^{(i)\top} \begin{bmatrix} c^{(i)} \\ \bar{e}^{(i)} \end{bmatrix} \right)$$

- Output projections reuse the input bucket embeddings.
- The next head receives a differentiable expectation over buckets, not a sampled ID.

Paper, Sec. 3.3; PDF pp. 3–4.

Inference removes probability mass on invalid signatures

Training objective: target coordinates

$$P_{\text{train}}(w \mid c) = \prod_{i=1}^{H} p^{(i)}_{\mathcal{H}_i(w)}, \qquad \mathcal{L} = -\sum_i \log p^{(i)}_{\mathcal{H}_i(w)}.$$

Inference distribution: registered tokens only

$$P_{\text{inf}}(w \mid c) = \frac{\prod_i p^{(i)}_{\mathcal{H}_i(w)}}{\sum_{w' \in \mathcal{V}_{\text{actl}}} \prod_i p^{(i)}_{\mathcal{H}_i(w')}}.$$

Paper, Sec. 3.4; PDF p. 4. The denominator scans the actual vocabulary.

The comparison includes a parameter-matched baseline

1B-scale model	Embedding	Decoder
Standard	65M	0.9B
Standard+2L	65M	1.0B
MHF H3B10K	65M	0.9B
MHF H4B16K	138M	0.9B

- H3B10K directly matches Standard's reported parameter counts.
- H4B16K uses more embedding parameters; Standard+2L spends its extra budget on backbone depth.

Paper, Sec. 4.1 and Table 1; PDF pp. 5–6. Counts are rounded as reported.

The study tests scaling, semantics, and adaptation

Question	Protocol
Core abilities	10 tasks: LAMBADA, five commonsense tasks, three reading tasks, ReCoRD.
Rare words	Card-660: human similarity versus hidden-state cosine similarity.
Vocabulary growth	6B Arabic/Chinese/Hindi + 2B English continual-training tokens; 32,768 to 48,122 tokens.

- 100M models train on 10B English tokens; 1B/3B models on 100B.
- Core benchmarks mix normalized accuracy, accuracy, and ReCoRD F1; 0-shot and 5-shot settings are specified per task.

Paper, Secs. 4.2–4.4; PDF pp. 5–6.

MHF lifts LAMBADA accuracy at 1B and 3B

Scale	Model	LAMBADA accuracy (%)
1B	Standard	30.41
1B	MHF H4B16K	**35.34**
3B	Standard	28.64
3B	MHF H4B16K	**37.26**

Absolute score difference

+4.93 percentage points at 1B; +8.62 percentage points at 3B.

Paper, Sec. 5.1 and Table 1; PDF p. 6.

The 1B gains are broad, with an OBQA reversal

Model	ARC-E	COPA	OBQA	ReCoRD
Standard	59.85	64.00	**31.60**	63.20
Standard+2L	60.23	65.00	31.20	63.78
MHF H3B10K	61.95	65.00	**34.00**	64.51
MHF H4B16K	**64.35**	**71.00**	29.80	**64.90**

- H4B16K beats Standard on 9/10 main tasks; OBQA is the exception.
- H3B10K beats Standard on 8/10 under equal reported parameters.

Paper, Table 1; PDF p. 6. ReCoRD is F1; ARC-E and OBQA use normalized accuracy.

At 100M, deeper backbones often do better

Model	Main run LAMBADA	3-seed mean (SD)
Standard	15.33	15.18 (0.4)
Standard+4L	**19.64**	**19.19 (0.6)**
MHF H3B10K	16.40	16.67 (0.5)
MHF H4B16K	18.79	18.80 (0.4)

- The authors attribute this to limited contextual disambiguation and representation capacity in small decoders.
- The three-seed appendix preserves the same LAMBADA ordering.

Paper, Sec. 5.1/Table 1; PDF pp. 6–7. Appendix F/Table 10; PDF p. 18.

Rare-word similarity improves in the penultimate layer

1B model	Pearson r	Spearman ρ
Standard	0.26	0.25
Standard+2L	0.29	0.29
MHF H3B10K	**0.32**	0.29
MHF H4B16K	0.30	**0.32**

- Card-660 correlates model cosine similarities with human ratings of rare-word pairs.
- These are representation correlations, not rare-word generation accuracy.

Paper, Sec. 5.2 and Table 2; PDF p. 7. Penultimate hidden state shown.

Vocabulary growth adds zero MHF model parameters

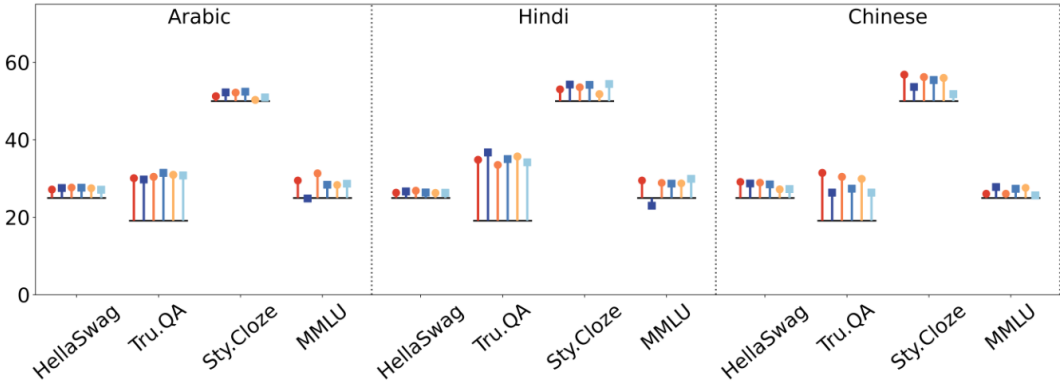
	Standard	MHF
Vocabulary	32,768 $\rightarrow$ 48,122	32,768 $\rightarrow$ 48,122
Added embedding parameters	about 31M	**0M**

Observed post-adaptation outcomes

H4B16K exceeds Standard on 12/22 tasks at 1B and 13/22 at 3B across Arabic, Hindi, Chinese, and English.

Paper, Secs. 4.4, 5.3 and Fig. 2; PDF pp. 5–8. Both systems receive continual pre-training.

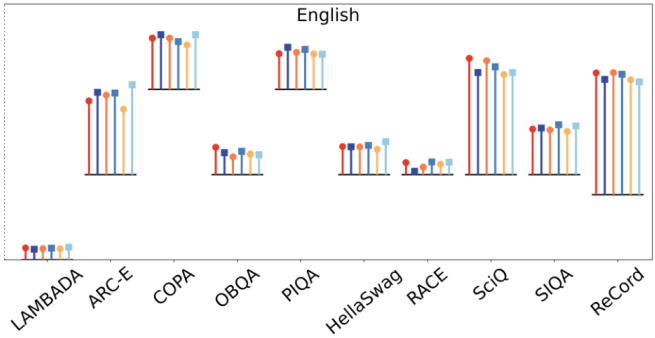
Added-language outcomes vary by task



- Arabic, Hindi, and Chinese each contribute four tasks to Figure 2.
- Red/orange circles: Standard; blue squares: MHF. Darker shades denote 1B models.

Paper, Fig. 2 and Sec. 5.3; PDF pp. 7–8. Cropped original panels; metrics vary by task.

English capabilities remain competitive after adaptation



- Authors count H3B10K and H4B16K as better or comparable on 6/10 and 9/10 English tasks, respectively.
- Red/orange circles: Standard; blue squares: MHF. Darker shades denote 1B models.

Paper, Fig. 2 and Sec. 5.3; PDF pp. 7–8. Cropped original English panel.

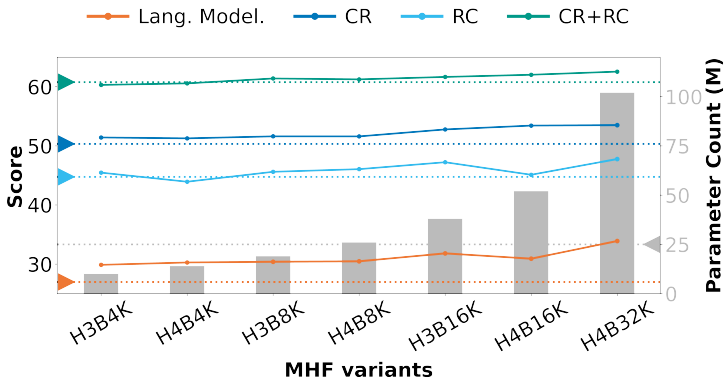
Four IDs beat one under matched parameter counts

LAMBADA accuracy (%)	B4K	B8K	B16K
H1: one ID	4.29	8.77	14.30
H4: four IDs	**30.27**	**30.47**	**30.91**

- H1 repeats its single value across four coordinates to match H4's parameters.
- Adding buckets alone leaves the token-identity ambiguity unresolved.

Paper, Sec. 6 and Table 3; PDF p. 8. Ablation uses 1B models trained up to 20B tokens.

More buckets bring diminishing returns



- Gray bars: embedding parameters; colored lines: mean task-group scores.
- H4B16K is the authors' chosen parameter/performance compromise.

Paper, Fig. 3 and Sec. 6; PDF p. 8. 1B, approximately 20B training tokens.

Character-trigram LSH adds no consistent gain

LSH : MMH3 heads	0:4	1:3	2:2	3:1	4:0
LAMBADA accuracy (%)	30.91	30.60	30.80	31.40	31.36

- The four-head configuration is held fixed while MMH3 heads are replaced.
- Performance does not rise monotonically with morphological hashing.

Paper, Sec. 6 and Table 4; PDF p. 8. 1B analysis setting.

The rank argument is suggestive, not a measured cause

- A standard LM's logit matrix factors through dimension d, so its rank is at most d.
- MHF sums multiple coordinate-specific terms $C_i E_i^\top \Pi_i$; distinct terms may increase the attainable rank.
- Appendix A's printed $\min(B, Hd)$ bound does not follow directly from that sum with different Π_i.
- The main 1B/3B configurations have one training seed each; Figure 2 shows mixed task-level expansion results.

Paper, Appendix A; PDF pp. 14–15. Limitations, PDF p. 9; Fig. 2, PDF p. 7.

Take-home messages

1. Full signatures solve token identity: shared individual buckets can still support deterministic token recovery.
2. The interface is useful at the tested larger scales: 1B/3B improve many tasks; 100M is less favorable.
3. Vocabulary parameters stay fixed under this expansion: inference still scores registered tokens, and the unique-signature map must be maintained.

Paper, Secs. 3, 5, 6 and Appendix H; PDF pp. 3–8, 17.

Paper and primary evidence

- Xue, Yamaguchi, and Aletras. *MultiHashFormer: Hash-based Generative Language Models*. EMNLP 2026; arXiv:2606.28057v2.
- Architecture: Fig. 1, PDF p. 2. Main benchmarks: Table 1, PDF p. 6.
- Rare words and vocabulary expansion: Table 2/Fig. 2, PDF p. 7.
- Ablations: Tables 3–4/Fig. 3, PDF p. 8. Theory and extra checks: Appendix A/F/H, PDF pp. 14–18.

Backup: what rank bound follows from the decomposition?

$$S = \sum_{i=1}^{H} C_i E_i^\top \Pi_i, \qquad C_i \in \mathbb{R}^{N \times d}, \quad E_i^\top \Pi_i \in \mathbb{R}^{d \times |\mathcal{V}|}.$$

$$\operatorname{rank}(S) \le \min\{N, |\mathcal{V}|, H \min(d, B)\}.$$

- Different Π_i allow the terms to occupy different token subspaces.
- Row-wise softmax normalization adds a row-constant term; an upper bound alone does not show the attained rank.

Reader's algebra from Paper, Appendix A.2; PDF p. 15.

Backup: decoding cost and source-text discrepancies

- Valid-token normalization scans $|\mathcal{V}_{\mathrm{actl}}|$ signatures. On one MI300X, Appendix H reports 97.38 ms/sample for 1B H4B16K versus 96.42 for Standard on LAMBADA.
- Sec. 5.1 attributes COPA 71.00 to H3B10K; Table 1 assigns it to H4B16K. The deck uses Table 1.
- Sec. 4.1 gives H3B10K $B = 10{,}240$; Appendix E Table 9 gives 10,624.
- Appendix G says the T-FREE comparison uses 20B tokens; Table 11's caption says 12B.

Paper, Table 1/PDF p. 6; Sec. 4.1/PDF p. 5; Appendix E/G/H/PDF pp. 17–18.