

[Why HTML?](#)[Report Issue](#)[Back to Abstract](#)[Download PDF](#)

License: CC BY 4.0
arXiv:2605.09751v1 [cs.CL] 10 May 2026

Language Models Without a Trainable Input Embedding Table: Learning from Fixed Minimal Binary Token Codes

Andrey Bochkov

Corresponding author: andrey.bochkov@gmail.com

Abstract

Trainable input embedding tables are a standard component of modern language models. We ask whether they are actually necessary at the input interface. For a vocabulary of size V , exact token identity requires only $K = \lceil \log_2 V \rceil$ bits. We replace the usual trainable $V \times d_{\text{model}}$ input embedding matrix with fixed minimal binary token codes and a zero-parameter lift to model width. In our main setting, $V = 65,536$, so $K = 16$, and tokens are represented by fixed 16-dimensional binary codes tiled to $d_{\text{model}} = 1024$. We also evaluate a fully table-free variant in which codes are generated from token IDs on the fly and randomly recoded by an invertible affine transform over \mathbb{F}_2^K . Across matched 32-layer decoder-only models trained on approximately 17B tokens and evaluated over three independent training seeds, fixed minimal codes achieve comparable held-out validation perplexity to a standard learned-input baseline while removing 67.1M trainable input parameters. The fixed-code runs have a lower mean validation perplexity in our experiments, 2.36 versus 2.44, but the observed gap is within the measured seed-to-seed variation of 4.8%; we therefore interpret the result as evidence that the trainable input table is not necessary, rather than as a statistically resolved superiority claim. The table-free affine-recoded variant remains close at 2.39 despite a slightly shorter training run. These results show that, in this regime, a

trainable input embedding table is not necessary for useful language modeling. The output projection remains standard and trainable. Modern language models often handle inputs whose tokenization provides weak or fragmented lexical cues: unusual casing, spelling variation, rare strings, code fragments, or character-like decompositions. This suggests that linguistic interpretation cannot be understood solely as retrieval of a semantic vector for each token. Rather, meaning is constructed by the contextual computation performed by the model. Our experiment isolates this idea in an extreme form by removing the learned input lookup entirely while preserving exact token identity.

1. Introduction

Modern language models almost universally begin with a trainable input embedding table. A token ID $t \in \{0, \dots, V-1\}$ is mapped to a learned vector from

$$E_{\text{in}} \in \mathbb{R}^{V \times d_{\text{model}}},$$

and this vector becomes the starting point for all subsequent computation. This design is so standard that the input embedding table is often treated as an indispensable component of the architecture.

From the standpoint of token identity, however, this interface is highly overparameterized. A vocabulary of size V contains V discrete symbols, and uniquely specifying one of them requires only

$$K = \lceil \log_2 V \rceil$$

bits. For $V = 65,536$, the exact minimum is $K = 16$. This raises a simple question: *does a Transformer language model need a trainable input embedding table, or is exact token identity enough?*

We study this question directly. We replace the learned input lookup with a fixed minimal binary code

$$c(t) \in \{0, 1\}^K, \quad K = \lceil \log_2 V \rceil,$$

followed by a fixed zero-parameter lift

$$x(t) = R c(t) \in \mathbb{R}^{d_{\text{model}}}.$$

In our implementation, \mathbf{R} is a deterministic tiled lift: the K -bit code is repeated until the model width is reached. Thus the model begins not from a learned continuous lexical vector, but from an injective symbolic code containing only exact token identity. Here “binary” refers to the code alphabet $\{0, 1\}$, not to low-precision quantization.

We also test a fully table-free variant. Instead of storing any input table, the model computes bits directly from token IDs and optionally applies an invertible affine recoding over \mathbb{F}_2^K :

$$\tilde{c}(t) = A c(t) \oplus b, \quad A \in GL(K, 2), \quad b \in \{0, 1\}^K.$$

For $V = 2^K$, this recoding remains a bijection over the full binary hypercube. It therefore preserves exact token identity while removing any special role of the canonical binary ordering of token IDs.

This construction is intentionally severe. Before the first trainable layer, all token inputs lie in a subspace of dimension at most K , even though the model width is much larger. In our main experiments, a 32-layer decoder-only Transformer with $d_{\text{model}} = 1024$ receives token inputs of effective rank only 16. If such a model performs comparably to a standard learned-input baseline, then the trainable input embedding table is not an architectural necessity in this regime.

We evaluate matched 32-layer decoder-only models trained from scratch on approximately 17B tokens. All runs use the same tokenizer, vocabulary size, architecture, output head, data mixture, and training recipe; only the input parameterization changes. The learned-input baseline uses a trainable $65,536 \times 1024$ input table. Our fixed-code model removes this 67.1M-parameter trainable input block entirely.

Empirically, fixed minimal binary codes match the learned-input baseline while removing the input table. Across three independent seeds, the fixed-code model has a lower mean validation perplexity in our runs, 2.36 versus 2.44 for the learned-input baseline, but the gap is within the measured seed-to-seed variation of 4.8%. We therefore use these results to support a non-necessity claim rather than a statistically significant improvement claim. A fully table-free affine-

recoded variant remains close at 2.39. These results indicate that the standard trainable input table can be removed without degrading held-out next-token modeling quality in this setting.

Our claim is intentionally narrow. We do not argue that learned input embeddings are never useful, that binary codes are universally optimal, or that output vocabulary projections are unnecessary. The output head remains standard and trainable throughout. We show instead that, under matched conditions in this regime, a large trainable *input* embedding table is not required for useful language modeling.

The surprising aspect of this result is not that the code is compact. It is that the model never learns an input lexical lookup at all. In the standard architecture, the input embedding table is the only trainable mechanism that converts token identity into the continuous computational space of the Transformer. Our fixed-code models remove this mechanism entirely: token identity enters only as an injective minimal binary code, and all continuous representation learning must occur inside the Transformer stack. Thus, non-degradation relative to a learned-input baseline is itself the central result.

Our goal is not to show that fixed binary codes are the best possible input parameterization. The experiment is instead a constructive counterexample to a stronger implicit assumption: that a trainable semantic input embedding table is necessary for language modeling. For such a necessity claim, the relevant evidence is not a benchmark leaderboard improvement, but the existence of a nontrivial language model that trains, converges, and models held-out text without any trainable input lookup. The matched learned-input baseline is used to show that this counterexample is not merely degenerate: removing the input table preserves comparable held-out likelihood under the same tokenizer, architecture, data, and training recipe. For reproducibility, code, configuration files, and supplementary artifacts are available at <https://huggingface.co/E6E831728>.

Contributions.

- We introduce a minimal-code input parameterization that replaces the trainable input embedding table with fixed binary token codes of width $K = \lceil \log_2 V \rceil$ and a zero-parameter lift to model width.
- We study a fully table-free affine-recoded variant that generates token codes directly from token IDs and applies an invertible transform over \mathbb{F}_2^K .
- In matched 32-layer experiments over three independent seeds, fixed minimal codes remove 67.1M trainable input parameters while matching the learned-input baseline. The fixed-code runs have lower mean validation perplexity in our experiments, 2.36 versus 2.44, but this gap lies within the measured 4.8% seed-to-seed variation.
- We isolate the input-interface question from output-layer changes: the output projection remains standard and trainable, while the trainable input parameter count drops from 67.1M to zero.

2. Related Work

Learned embeddings.

Dense learned word vectors such as word2vec and GloVe established the view that lexical items can be represented by learned continuous vectors (Mikolov et al., [2013](#); Pennington et al., [2014](#)). Transformer language models inherited this convention: token IDs are mapped through a learned input table before contextual processing (Vaswani et al., [2017](#); Radford et al., [2019](#); Brown and others, [2020](#)). Our work revisits this default design and asks whether the trainable *input* table itself is necessary when exact token identity is already available.

Embedding compression and sharing.

Many methods reduce the cost of vocabulary-dependent parameters while retaining a learned lexical interface. ALBERT factorizes embeddings through a smaller learned table and projection (Lan et al., [2020](#)); adaptive inputs allocate different capacities across vocabulary bands (Baevski and Auli, [2019](#)); DeFINE uses deep factorized input embeddings (Mao and others, [2019](#)). Other

approaches share parameters across tokens, including LightRNN (Li et al., [2016](#)), hash embeddings (Svenstrup et al., [2017](#)), and compositional code learning (Shu and Nakayama, [2018](#)). These methods compress or share learned input representations. In contrast, we remove the trainable input embedding table entirely and replace it with a fixed minimal injective code. Hash embeddings still learn component vectors and combine them through a trainable parameterization; our table-free variant does not learn any input-side token vectors or projections.

Output-layer efficiency.

Hierarchical and adaptive softmax methods reduce the cost of predicting over large vocabularies (Morin and Bengio, [2005](#); Mnih and Hinton, [2009](#); Grave et al., [2017](#)). Our question is orthogonal: we leave the output projection standard and trainable, and study only whether the *input* embedding table is necessary.

Alternative token granularities.

Character-, byte-, and tokenization-free models change the units presented to the model (Kim et al., [2016](#); Xue et al., [2022](#); Clark et al., [2022](#); Tay et al., [2022](#); Yu and others, [2023](#)). These approaches improve robustness or open-vocabulary behavior by changing the tokenization interface, but they typically still use learned input embeddings or learned front ends. We keep the tokenizer fixed and change only how token IDs are represented at the input.

Fixed input representations.

Frozen or externally constructed token representations have been used in several NLP settings, but such methods usually still provide a vocabulary-sized table of token-specific vectors. Our strongest variant is table-free: token codes are generated algorithmically from token IDs using $K = \lceil \log_2 V \rceil$ bits, optionally affinely recoded over \mathbb{F}_2^K , and lifted to model width with no trainable input parameters. This distinguishes our work from both embedding compression and frozen embedding lookup methods.

Our work is also related in spirit to the broader use of random or fixed feature maps, where a non-trainable input transformation is followed by learned downstream layers. The distinction is that our code is not a random projection of

continuous inputs: it is an exact minimal injective code for discrete token identity, and the main comparison is to a standard trainable input table in a language model.

3. Method

We replace the standard trainable input lookup $t \mapsto E_{\text{in}}[t]$ with a deterministic binary coding interface. For a vocabulary of size V , exact injective token identity requires at least

$$K = \lceil \log_2 V \rceil$$

bits. We therefore assign token ID $t \in \{0, \dots, V-1\}$ the canonical minimal code

$$c(t) = \text{bin}_K(t) \in \{0, 1\}^K, \quad c(t)_j = \left\lfloor \frac{t}{2^j} \right\rfloor \bmod 2.$$

When $V = 2^K$, this is a bijection between token IDs and the full hypercube $\{0, 1\}^K$.

To test whether performance depends on canonical token-ID ordering, we also use invertible affine recodings over \mathbb{F}_2^K :

$$\tilde{c}(t) = A c(t) \oplus b, \quad A \in GL(K, 2), \quad b \in \{0, 1\}^K.$$

This preserves injectivity while changing the token-to-code assignment.

The binary code is mapped to model width by a fixed zero-parameter lift

$$x(t) = R \tilde{c}(t), \quad R \in \mathbb{R}^{d_{\text{model}} \times K}.$$

In our experiments, d_{model} is divisible by K and R is a tiled identity:

$$R = \begin{bmatrix} I_K \\ I_K \\ \vdots \\ I_K \end{bmatrix}.$$

Thus the K -bit code is repeated until the model width is reached. For $V = 65,536$, $K = 16$ and $d_{\text{model}} = 1024$, so each token is represented by a 16-dimensional binary code tiled 64 times.

The map $t \mapsto x(t)$ is collision-free whenever A is invertible over \mathbb{F}_2 and R has full column rank. Moreover, all input vectors lie in a subspace of dimension at most K , despite the model width being $d_{\text{model}} \gg K$. Proofs of minimality, injectivity, full-hypercube coverage, and effective input rank are given in Appendix [D](#).

The canonical fixed-code variant can be implemented either as a frozen deterministic lookup or by computing bits on the fly. When the frozen lookup stores exactly the vectors generated by the deterministic rule, it is functionally identical to the table-free implementation: replacing the lookup by on-the-fly bit extraction produces the same input vector for every token. Thus the canonical fixed-code experiment should be interpreted as an input-parameterization experiment, not as evidence for the necessity of storing a table. The affine-recoded run uses the table-free implementation directly. The affine-recoded variant is fully table-free: it extracts bits from token IDs, applies the fixed affine transform, casts the result to floating point, and applies the tiled lift. In all variants, the trainable input-embedding parameter count is zero.

4. Experimental Setup

Study design.

Each full-scale model variant is trained with three independent random seeds. Seeds affect model initialization and data-order randomness; the tokenizer, architecture, data mixture, validation split, optimizer hyperparameters, and input parameterization are fixed within each variant. We report mean validation metrics across seeds and the observed seed-to-seed variation.

Architecture.

All models are 32-layer decoder-only Transformers with $d_{\text{model}} = 1024$, 32 attention heads, rotary positional encodings, GELU activations, context length 1024, and an untied trainable output projection. The vocabulary size is $V = 65,536$, so the

minimal code width is $K = 16$. A standard learned-input baseline therefore has $65,536 \times 1024 = 67.1\text{M}$ trainable input parameters; the fixed-code variants have zero trainable input parameters. The output projection remains standard and trainable in all models.

Input parameterizations.

We compare three variants:

1. **Learned input table:** the standard trainable lookup $E_{\text{in}} \in \mathbb{R}^{V \times d_{\text{model}}}$.
2. **Fixed minimal binary code:** $x(t) = R \text{bin}_{16}(t)$ with a fixed tiled lift to width 1024.
3. **Affine-recoded table-free code:** $x(t) = R(A \text{bin}_{16}(t) \oplus b)$, where $A \in GL(16, 2)$ and $b \in \{0, 1\}^{16}$. This variant contains no input table and computes codes from token IDs at runtime.

Since $d_{\text{model}} / n_{\text{head}} = 32 = 2K$, each attention head receives two complete copies of the 16-bit code under the tiled lift.

Data and training.

All models are trained from scratch on the same FineWeb-Edu (Penedo and others, 2024) and Cosmopedia (Ben Allal and others, 2024) mixture for approximately 17B tokens (we sample FineWeb-Edu with probability 0.8 and Cosmopedia with probability 0.2), with a held-out validation split from the same preprocessing pipeline. The validation split is constructed at the document level before tokenization and is shared across all model variants. The learned-input baseline and canonical fixed-code models are trained with three independent seeds to 17.093B tokens per seed. Full optimizer, batch-size, learning-rate, hardware, and checkpointing details are reported in Appendix A.1. The tokenizer is a deterministic 65,536-token Unicode-compatible tokenizer constructed before training and kept fixed for all runs. Optional SFT continuations are run from the corresponding base checkpoints using the same input parameterization and architecture. All SFT continuations use the same supervised mixture, learning-rate schedule, and evaluation protocol. Because all models use the same tokenizer, validation perplexity is used only for within-tokenizer comparisons

between input parameterizations. We do not interpret these perplexities as directly comparable to models trained with different tokenizers.

Evaluation.

We focus the main paper on held-out language-modeling quality under matched training conditions. For each three-seed comparison, we report the arithmetic mean across seeds. We also report the relative seed range, defined as $(\max_s m_s - \min_s m_s) / \text{mean}_s(m_s)$. We do not assume normality and do not use these ranges as confidence intervals.

5. Results

5.1. Main result: held-out language modeling does not require a trainable input table

Table 1: Main held-out language-modeling comparison at full 32-layer scale. All models use the same tokenizer, architecture, output head, and training pipeline; only the input parameterization differs. Lower is better for validation loss and perplexity. Relative seed variation is computed as $(\max - \min) / \text{mean}$ across three seeds.

Model variant	Tokens/seed	Val loss	Val PPL, mean	Rel. seed range
Learned input table	17.1 (B)	0.893	2.44	4.8%
Fixed minimal binary code	17.1 (B)	0.859	2.36	4.6%
Affine-recoded minimal code (table-free)	16.3 (B)	0.871	2.39	4.5%

[Table 1](#) reports the primary comparison of the paper. Under matched architecture, tokenizer, and training recipe, replacing the standard learned input embedding table with fixed minimal binary token codes does not harm held-out language modeling. Across three independent seeds, the canonical fixed-code model has lower mean validation perplexity than the learned-input baseline in

our runs, 2.36 versus 2.44 at the same final token count of 17.093B per seed. However, the observed difference is within the measured seed-to-seed variation of 4.8%, so we do not claim a statistically significant improvement.

This result is the central empirical finding of the paper: the learned input table can be removed without degrading held-out language-modeling quality in this regime. The learned baseline uses a trainable input embedding matrix of size

$$65,536 \times 1024 = 67,108,864$$

parameters, whereas the fixed-code model uses zero trainable input parameters. Thus, in this regime, removing the trainable input embedding table entirely from the optimization problem does not degrade language-model quality in our experiments; the fixed-code runs have a slightly lower mean perplexity, but this difference is not statistically resolved.

The table-free affine-recoded model also remains competitive, achieving validation perplexity 2.39 with zero trainable input parameters. This is important because it removes any special role of the canonical binary ordering of token IDs: the model still trains successfully when the minimal binary codebook is randomly recoded by an invertible affine transform over \mathbb{F}_2^{16} before entering the Transformer. We therefore do not attribute the result to a privileged token-ID ordering or to a hand-designed semantic geometry at the input.

Interpretation.

The main takeaway from [Table~1](#) is not that fixed minimal binary codes are universally optimal, but that a large learned continuous input table is *not necessary* for useful language modeling in this setting. Exact token identity, supplied through a fixed minimal binary interface, is sufficient for the downstream Transformer stack to learn effective internal representations.

5.2. Robustness to code assignment

The affine-recoded experiment provides a direct robustness check on the code assignment itself. If the canonical fixed-code model were competitive only because the raw binary expansion of token IDs introduced a special or

exploitable structure, then a random invertible affine recoding of the full codebook would be expected to degrade performance sharply. We do not observe such a collapse.

The affine-recoded table-free model remains close to the canonical fixed-code model and is also competitive with the learned-input baseline in this run (2.39 versus 2.44), despite seeing slightly fewer training tokens. Because this run terminates at a different token count, we interpret it primarily as a robustness check rather than as a strict superiority claim. We interpret this as evidence that the core requirement is a stable injective token-identity interface, not a particular continuous geometry or a particular ordering of token IDs. At the same time, the canonical and affine-recoded variants are not identical, and we do not claim full invariance to code assignment. Rather, the result shows that performance remains viable and competitive under substantial codebook recoding.

Why the benchmark evidence is secondary.

For completeness, Appendix C reports optional SFT evaluations. We treat those results as secondary because the central claim of the paper is established by the matched held-out language-modeling comparison in [Table 1](#).

5.3. Summary of findings

Across all full-scale comparisons, the most stable result is that fixed minimal binary input codes remain comparable to the standard learned-input baseline while eliminating the trainable input table. The canonical fixed-code model achieves the lowest mean validation loss and perplexity in our runs while eliminating 67.1M trainable input-embedding parameters, but the margin is within measured seed-to-seed variation. The fully table-free affine-recoded model remains close. Optional SFT continuation evaluations in the appendix show that the fixed-code interface remains usable after a second training stage, but they are not used to establish the main claim.

Taken together, these experiments support the paper's central claim: in this regime, a trainable input embedding table is not necessary for useful language

modeling. Exact token identity delivered through fixed minimal binary codes is sufficient for a Transformer to learn competitive internal representations.

Supplementary matched-depth ablations at 9 and 16 layers in Appendix B show the same non-degradation pattern, but are not used as the primary evidence for the paper.

6. Discussion

The main result is a sufficiency statement: under the matched conditions studied here, a trainable *input* embedding table is not necessary for useful language modeling. A fixed minimal binary token code provides exact token identity, and the Transformer stack can build effective internal representations from that stable symbolic interface. This does not imply that learned input embeddings are never useful, nor that minimal binary codes are universally optimal. It shows that the standard learned input lookup is not an architectural requirement in this regime.

The affine-recoded experiment helps clarify the role of code assignment. The canonical binary code could in principle benefit from accidental token-ID ordering. By applying an invertible affine transform over \mathbb{F}_2^{16} , we preserve injectivity while changing the coordinate interpretation of the codebook. The resulting table-free model remains competitive, suggesting that the key ingredient is a stable collision-free token-identity interface rather than a privileged learned or hand-designed input geometry. However, we test only one affine recoding at full scale and do not claim invariance to all possible code assignments.

Finally, the mechanism behind the fixed-code model's slight perplexity advantage remains open. Possible explanations include regularization from removing a large trainable input block, reduced parameter drift at the input interface, or easier separation between token identity and contextual representation learning. Identifying the causal mechanism is an important direction for future work.

A non-degradation result is meaningful here because the experiment tests a necessity claim. If the trainable input table were essential, removing it and restricting the input to a 16-dimensional fixed code should have produced a clear degradation. It did not.

The result should be interpreted as a constructive counterexample to necessity, not as a benchmark-optimization claim. A single well-controlled model that learns useful language modeling without a trainable input embedding table is already informative: it shows that such a table is not required as the place where lexical or semantic structure must first be learned. Additional seeds and matched baselines are useful for ruling out implementation artifacts and estimating variability, but the conceptual point is the existence of a working language model whose input interface contains only a fixed injective token-identity code.

The result also clarifies a distinction that is often blurred in discussions of embeddings. A token representation at the input need not be a semantic vector; it may simply be a stable address. Natural language models routinely process inputs whose tokenization provides weak lexical cues, such as unusual casing, rare strings, code fragments, or character-like decompositions. In such cases, meaning cannot be explained as direct retrieval of a single pretrained semantic vector. It must be constructed by contextual computation. Our fixed-code experiment isolates this principle in an extreme form: the input code contains only token identity, yet the model still learns useful language modeling.

7. Limitations

Several limitations remain. First, the full-scale comparison uses one tokenizer family, and one main vocabulary size. Although the method extends directly to arbitrary V via $K = \lceil \log_2 V \rceil$, broader experiments across tokenizers, vocabulary sizes, byte-level settings, and multilingual corpora are needed. Second, although we train each main full-scale variant with three independent seeds, the observed seed-to-seed variation is still large enough that we do not claim statistically significant superiority of fixed codes over learned input embeddings. We therefore emphasize the non-degradation and parameter-removal result rather than a strict performance improvement claim. Third, we use a shared training recipe rather than extensive hyperparameter retuning for each input

parameterization. A larger sweep, especially over learning rate and warmup for the learned-input baseline, could change the small observed perplexity margins. Our conclusion therefore relies on non-degradation under a shared recipe rather than on optimality of the fixed-code parameterization.

8. Conclusion

We revisited the necessity of trainable input embedding tables in Transformer language models. Replacing the standard learned $V \times d_{\text{model}}$ input lookup with fixed minimal binary token codes of width $K = \lceil \log_2 V \rceil$ removes all trainable input-embedding parameters while preserving exact token identity.

In matched 32-layer experiments with $V = 65,536$, fixed 16-dimensional binary codes tiled to $d_{\text{model}} = 1024$ match the learned-input baseline while eliminating 67.1M trainable input parameters. The fixed-code runs have lower mean held-out validation perplexity in our experiments, 2.36 versus 2.44, but this gap is within the measured 4.8% seed-to-seed variation. A fully table-free affine-recoded variant remains close at 2.39.

These results support a narrow but important conclusion: in the regime studied here, useful language modeling does not require a trainable input embedding table. A stable, injective, minimal token-identity code is sufficient for the Transformer stack to learn effective internal representations.

9. Broader Impact

This work studies the input interface of decoder-only Transformer language models. The main potential benefit is scientific: fixed minimal binary token codes provide a simpler and more controlled way to study what information must be present at the input of a language model. The method may also modestly reduce optimizer-state requirements associated with the input layer by removing trainable input-embedding parameters.

The risks are similar to those of language-model training more generally. Any method that simplifies model construction could incrementally lower the barrier to building systems that may be misused for spam, misinformation, or other harmful text generation. The method does not introduce a new class of capabilities, and the models studied here are small research models, but standard documentation and responsible release practices remain appropriate for any released code or checkpoints.

The work does not introduce new personally identifying datasets or new data collection procedures. Replacing the input interface does not address standard language-model issues such as bias, hallucination, unsafe generations, or training-data artifacts.

References

A. Baevski and M. Auli (2019)

Adaptive input representations for neural language modeling.

In International Conference on Learning Representations,
Cited by: [§2](#).

L. Ben Allal et al. (2024)

Cosmopedia.

Hugging Face dataset.

Note:

<https://huggingface.co/datasets/HuggingFaceTB/cosmopedia>

Cited by: [§4](#).

T. B. Brown et al. (2020)

Language models are few-shot learners.

In Advances in Neural Information Processing Systems,

Vol. **33**, pp. 1877–1901.

Cited by: [§2](#).

J. H. Clark, D. Garrette, I. Turc, and J. Wieting (2022)

CANINE: pre-training an efficient tokenization-free encoder for language representation.
Transactions of the Association for Computational Linguistics **10**, pp. 73–91.

Cited by: [§2](#).

É. Grave, A. Joulin, M. Cissé, D. Grangier, and H. Jégou (2017)

Efficient softmax approximation for GPUs.

In Proceedings of the 34th International Conference on Machine Learning, pp. 1302–1310.

Cited by: [§2](#).

Y. Kim, Y. Jernite, D. Sontag, and A. M. Rush (2016)

Character-aware neural language models.

In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, pp. 2741–2749.

Cited by: [§2](#).

Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut (2020)

ALBERT: a lite BERT for self-supervised learning of language representations.

In International Conference on Learning Representations,

Cited by: [§2](#).

X. Li, T. Qin, J. Yang, X. Hu, and T. Liu (2016)

LightRNN: memory and computation-efficient recurrent neural networks.

In Advances in Neural Information Processing Systems, Vol. **29**, pp. 4385–4393.

Cited by: [§2](#).

Y. Mao et al. (2019)

DeFINE: deep factorized input token embeddings for neural sequence modeling.

In International Conference on Learning Representations,

Cited by: [§2](#).

T. Mikolov, K. Chen, G. Corrado, and J. Dean (2013)

Efficient estimation of word representations in vector space.
arXiv preprint arXiv:1301.3781.
Cited by: [§2](#).

A. Mnih and G. E. Hinton (2009)

A scalable hierarchical distributed language model.
In Advances in Neural Information Processing Systems,
Vol. **21**, pp. 1081–1088.
Cited by: [§2](#).

F. Morin and Y. Bengio (2005)

Hierarchical probabilistic neural network language model.
In Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics,
pp. 246–252.
Cited by: [§2](#).

G. Penedo et al. (2024)

The fineweb datasets: decanting the web for the finest text data at scale.
arXiv preprint arXiv:2406.17557.
Cited by: [§4](#).

J. Pennington, R. Socher, and C. D. Manning (2014)

GloVe: global vectors for word representation.
In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP),
pp. 1532–1543.
Cited by: [§2](#).

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever (2019)

Language models are unsupervised multitask learners.
Technical report
OpenAI.
Note: Technical report
Cited by: [§2](#).

R. Shu and H. Nakayama (2018)

Compressing word embeddings via deep compositional code learning.

In International Conference on Learning Representations,
Cited by: [§2](#).

D. Svenstrup, J. Hansen, and O. Winther (2017)

Hash embeddings for efficient word representations.

In Advances in Neural Information Processing Systems,
Vol. **30**, pp. 4928–4936.
Cited by: [§2](#).

Y. Tay, M. Dehghani, V. Aribandi, J. Gupta, P. Pham, Z. Qin, D. Bahri, and D. Metzler (2022)

Charformer: fast character transformers via gradient-based subword tokenization.

In International Conference on Learning Representations,
Cited by: [§2](#).

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin (2017)

Attention is all you need.

In Advances in Neural Information Processing Systems,
Vol. **30**, pp. 5998–6008.
Cited by: [§2](#).

L. Xue, A. Barua, N. Constant, M. Kale, R. Al-Rfou, A. Roberts, and C. Raffel (2022)

ByT5: towards a token-free future with pre-trained byte-to-byte models.

Transactions of the Association for Computational Linguistics **10**,
pp. 291–306.
Cited by: [§2](#).

L. Yu et al. (2023)

MEGABYTE: predicting million-byte sequences with multiscale transformers.

arXiv preprint arXiv:2305.07185.
Cited by: [§2](#).

Appendix A Technical appendices and supplementary material

A.1. Training details

All full-scale base runs use the same 32-layer decoder-only Transformer architecture with $d_{\text{model}} = 1024$, $n_{\text{head}} = 32$, context length 1024, rotary positional embeddings, GELU activations and an untied output projection head. The vocabulary size is 65,536.

Each main variant is trained with three independent random seeds. The random seed controls parameter initialization and data-order randomness. The tokenizer, validation split, architecture, optimizer hyperparameters, context length, and input parameterization are fixed across seeds. Unless otherwise stated, reported validation metrics are means across the three seeds. The observed relative seed-to-seed deviation in validation perplexity is 4.8%.

Training uses AdamW with weight decay 0.1, gradient clipping 1.0, micro-batch size 16 for base pretraining, gradient accumulation 200, and FP32 precision. The learning-rate schedule is cosine decay with warmup; The learned-input baseline and canonical fixed-code model are trained to 17.093B tokens. The affine-recoded table-free run is trained to 16.276B tokens due to scheduling constraints; we report exact token counts in the main results table.

Base pretraining was performed on $2 \times$ H100 80GB accelerators per run, with observed throughput of approximately 19.8k tokens/s. SFT continuations used the same architecture and corresponding input parameterization; observed SFT throughput was approximately 5.5k tokens/s.

Validation loss is computed on a held-out split drawn from the same preprocessing pipeline as the pretraining corpus. Downstream logit-based evaluations score each answer option by conditional log-likelihood and select the highest-scoring option.

The validation split is constructed before tokenization at the document level: documents assigned to validation are excluded from the training stream. We do not create validation examples by slicing contiguous chunks from documents that

also appear in training. Validation loss is computed with the same tokenizer and context length as training.

Table 2: Training hyperparameters for the main runs.

Run	Precision	Optimizer	LR	Warmup steps	Weight decay	Grad clip
Base pretraining	FP32	AdamW	4.0×10^{-4}	150	0.1	1.0
SFT continuation	FP32	AdamW	2.0×10^{-5}	100	0.01	1.0

Table 3: Batching and compute configuration.

Run	GPUs	Global micro-batch	Grad accum	Effective tokens/update
Base pretraining	2 × H100 80GB	16	200	3.28M
SFT continuation	1 × H100 80GB	6	100	0.61M

The micro-batch size is reported globally across GPUs; the per-device micro-batch is 8 for the 2-GPU base runs.

A.2. Tokenizer details

All experiments use the same deterministic tokenizer with vocabulary size 65,536. The tokenizer is constructed before training and then kept fixed for all model variants. The vocabulary contains reserved special tokens, direct Unicode-compatible token assignments, and a fixed set of frequent multi-character tokens extracted before model training. The exact tokenizer construction script and vocabulary file are included in the supplementary material. Because all

comparisons in the paper use the same tokenizer, validation perplexities should be interpreted as within-tokenizer comparisons rather than as cross-tokenizer language-modeling scores.

Appendix B Additional matched-depth ablations

We additionally ran smaller matched-depth comparisons at 9 and 16 Transformer layers. These experiments use the same tokenizer and compare the same two input interfaces: a standard learned input table and fixed minimal binary token codes. They are included only as supplementary sanity checks; the main claim of the paper is established by the 32-layer comparison in [Table 1](#).

Table 4: Supplementary matched-depth ablations. Lower validation perplexity is better. These runs are not intended as a depth-scaling study; they only check whether the fixed-code interface remains competitive with a learned-input baseline at matched depth.

Depth	Learned input table Val PPL	Fixed minimal code Val PPL	Relative change
9 layers	3.77	3.76	-0.3%
16 layers	2.42	2.38	-1.7%
32 layers	2.44	2.36	-3.3%

Across these additional matched-depth checks, replacing the learned input table with fixed minimal binary codes does not degrade validation perplexity. We do not interpret the table as evidence for a depth scaling law, since hyperparameters were not independently optimized for each depth.

Appendix C SFT evaluations

To test whether the fixed-code interface remains usable under a second-stage adaptation regime, we also ran optional supervised fine-tuning continuations of all three base pretrained models. Because these runs change the objective and data mixture relative to the main pretraining comparison, we treat them as secondary.

Table 5: Optional supervised fine-tuning (SFT) continuation results. These runs are secondary because they alter the data distribution and objective relative to the base pretraining comparison. Higher is better except for validation perplexity and SQuAD gold-answer perplexity, where lower is better.

Model variant	Val PPL	ARC-C	HellaSwag	BoolQ	SQuAD gold PPL
Learned input table + SFT	2.42	29.77	29.40	64.06	3.48
Fixed minimal binary code + SFT	2.39	28.09	26.70	61.33	3.33
Affine-recoded minimal code + SFT	2.44	31.44	26.80	64.06	3.65

The main conclusion from [Table 5](#) is compatibility rather than dominance: all three models can be further adapted successfully, and the fixed-code variants remain competitive after SFT. The canonical fixed-code model again achieves the best validation perplexity among the SFT continuations and the best SQuAD gold-answer perplexity, while the affine-recoded variant achieves the strongest ARC-C score by a narrow margin. The learned-input baseline remains strongest on HellaSwag and ties for the strongest BoolQ score. As in the base setting, the picture is task-dependent rather than uniformly one-sided.

These SFT results reinforce the main message of the paper. Fixed minimal binary token codes are not merely sufficient for pretraining convergence; they also provide a usable interface for continued downstream adaptation. However, because SFT introduces additional confounds, we do not treat these continuation results as the main evidence for the paper's core claim.

Appendix D Proofs and additional method details

We study decoder-only language models in which the standard trainable input embedding table is replaced by a deterministic binary coding interface. Let the vocabulary be

$$\mathcal{V} = \{0, 1, \dots, V - 1\},$$

and let the model width be $d = d_{\text{model}}$. A conventional language model learns an input embedding matrix

$$E_{\text{in}} \in \mathbb{R}^{V \times d},$$

and initializes token $t \in \mathcal{V}$ with the learned vector $E_{\text{in}}[t]$. We instead compute the input representation as

$$x(t) = R \tilde{c}(t) \in \mathbb{R}^d,$$

where: (i) $\tilde{c}(t) \in \{0, 1\}^K$ is a fixed binary code assigned to token t , (ii) $R \in \mathbb{R}^{d \times K}$ is a fixed zero-parameter lift from code space to model width, and (iii) no trainable input embedding parameters are used.

The rest of the Transformer architecture is unchanged. In particular, positional encoding, self-attention, feed-forward layers, normalization, and the output projection remain standard. Our claim in this paper concerns the *input* interface only; the output projection is still trainable.

D.1. Minimal binary token codes

The central observation is that exact token identity does not require a d -dimensional learned vector. For a vocabulary of size V , the smallest binary code length that can uniquely identify all tokens is

$$K = \lceil \log_2 V \rceil.$$

Proposition 1 (Minimal injective binary code length). *Any injective map from a vocabulary of size V into $\{0, 1\}^K$ requires*

$$K \geq \lceil \log_2 V \rceil.$$

Proof. The space $\{0,1\}^K$ contains exactly 2^K distinct binary vectors. An injective assignment of V tokens to binary codes is possible only if $2^K \geq V$. Taking base-2 logarithms yields

$$K \geq \log_2 V.$$

Since K must be an integer, the minimal admissible value is

$$K \geq \lceil \log_2 V \rceil.$$

■

We therefore use the exact minimal width

$$K = \lceil \log_2 V \rceil.$$

For token IDs represented as integers $t \in \{0, \dots, V-1\}$, the canonical minimal code is the K -bit binary expansion of t :

$$c(t) = \text{bin}_K(t) \in \{0,1\}^K,$$

with components

$$c(t)_j = \left\lfloor \frac{t}{2^j} \right\rfloor \bmod 2, \quad j = 0, \dots, K-1.$$

Here $j=0$ corresponds to the least significant bit. When $V=2^K$, this map is a bijection between token IDs and the full binary hypercube $\{0,1\}^K$. When $V < 2^K$, the map remains injective and leaves $2^K - V$ codes unused.

Although the construction trivially extends to any code width $K' \geq \lceil \log_2 V \rceil$, we focus on the exact minimal case because it gives the strongest test of whether a trainable input embedding table is actually necessary.

D.2. Affine recoding over \mathbb{F}_2

To test whether performance depends on the canonical ordering of token IDs or on a special geometry of the raw binary expansion, we also consider affine recodings of the codebook over the finite field \mathbb{F}_2 . Let

$$A \in GL(K, 2)$$

be an invertible $K \times K$ binary matrix and let

$$b \in \{0, 1\}^K$$

be a binary shift vector. We define the affine-recoded token code as

$$\tilde{c}(t) = A c(t) \oplus b,$$

where \oplus denotes addition modulo 2, and all multiplications by A are performed over \mathbb{F}_2 .

This family is strictly richer than simple coordinate permutations: a dense invertible matrix A can mix multiple input bits into each output bit while preserving invertibility. In the special case $A = I$ and $b = \mathbf{0}$, we recover the canonical code.

After the affine recoding step, the binary vector $\tilde{c}(t)$ is cast to floating point and used as a standard real-valued input to the fixed lift described next.

Proposition 2 (No collisions under affine recoding and full-rank lift). Assume that $c : \mathcal{V} \rightarrow \{0, 1\}^K$ is injective, that $A \in GL(K, 2)$ is invertible over \mathbb{F}_2 , and that $R \in \mathbb{R}^{d \times K}$ has full column rank:

$$\text{rank}(R) = K.$$

Then the full input map

$$x(t) = R \tilde{c}(t) = R(A c(t) \oplus b)$$

is injective as a map from \mathcal{V} into \mathbb{R}^d . In particular, distinct tokens never collide at the model input.

Proof. Suppose $x(t_1) = x(t_2)$ for two tokens $t_1, t_2 \in \mathcal{V}$. Then

$$R \tilde{c}(t_1) = R \tilde{c}(t_2),$$

so

$$R(\tilde{c}(t_1) - \tilde{c}(t_2)) = \mathbf{0}.$$

Because $\text{rank}(R) = K$, the nullspace of R is trivial, hence

$$\tilde{c}(t_1) = \tilde{c}(t_2)$$

as vectors in \mathbb{R}^K , and therefore also as binary vectors. Since

$$\tilde{c}(t) = A c(t) \oplus b$$

and A is invertible over \mathbb{F}_2 , equality of the affine-recoded codes implies

$$c(t_1) = c(t_2).$$

Finally, because c is injective, this yields $t_1 = t_2$. Therefore the input map $x(\cdot)$ is injective. ■

The proposition is important operationally: in our setup, replacing the learned input table does *not* collapse multiple tokens onto the same continuous input vector. Distinct token identities remain distinct all the way to the first trainable layer.

Corollary 1 (Full hypercube coverage when $V = 2^K$). *If $V = 2^K$ and $c(t) = \text{bin}_K(t)$, then the set of affine-recoded codes*

$$\{\tilde{c}(t) : t \in \mathcal{V}\}$$

is exactly the full hypercube $\{0, 1\}^K$.

Proof. When $V = 2^K$, the canonical map $c(t) = \text{bin}_K(t)$ is a bijection from \mathcal{V} to $\{0, 1\}^K$. Any affine map

$$z \mapsto Az \oplus b$$

with $A \in GL(K, 2)$ is a bijection on $\{0, 1\}^K$. Therefore the composition $t \mapsto \tilde{c}(t)$ is also a bijection onto the full hypercube. ■

A useful consequence is that in the $V = 2^K$ case, the codebook is perfectly balanced: each bit takes value $\mathbf{1}$ for exactly 2^{K-1} tokens and value $\mathbf{0}$ for exactly 2^{K-1} tokens. Likewise, for any distinct code coordinates $i \neq j$, each of the four patterns $(\mathbf{0}, \mathbf{0})$, $(\mathbf{0}, \mathbf{1})$, $(\mathbf{1}, \mathbf{0})$, and $(\mathbf{1}, \mathbf{1})$ occurs exactly 2^{K-2} times. Thus affine recoding changes the token-to-code assignment, but does not introduce trivial global frequency imbalances into the codebook.

D.3. Zero-parameter lift to model width

The binary code $\tilde{c}(t) \in \{0, 1\}^K$ must be mapped into the model width d . We do this with a fixed matrix

$$R \in \mathbb{R}^{d \times K},$$

yielding the continuous input vector

$$x(t) = R \tilde{c}(t).$$

Any fixed full-column-rank lift is admissible. In the main experiments, d is chosen to be divisible by K , so we write

$$d = sK$$

for an integer s . We then use the tiled lift

$$R = \begin{bmatrix} I_K \\ I_K \\ \vdots \\ I_K \end{bmatrix} \in \{0, 1\}^{d \times K},$$

where I_K is the $K \times K$ identity matrix repeated s times vertically. Equivalently,

$$x(t) = \begin{bmatrix} \tilde{c}(t) \\ \tilde{c}(t) \\ \vdots \\ \tilde{c}(t) \end{bmatrix} \in \mathbb{R}^d,$$

i.e., the K -dimensional code is tiled deterministically until the full model width is reached.

This lift has zero trainable parameters and full column rank. It also preserves the complete token code in every consecutive block of K hidden dimensions. In our main setting, $V = 65,536$, $K = 16$, and $d = 1024$, so the token identity presented to the model is an exact minimal 16-dimensional binary code tiled to width 1024.

Proposition 3 (Effective input rank). *Let*

$$\mathcal{X} = \{x(t) : t \in \mathcal{V}\} \subseteq \mathbb{R}^d$$

be the set of all possible input vectors. Then

$$\dim(\text{span}(\mathcal{X})) \leq \text{rank}(R) \leq K.$$

If $V = 2^K$ and $\text{rank}(R) = K$, then

$$\dim(\text{span}(\mathcal{X})) = K.$$

Proof. By construction, every input vector has the form

$$x(t) = Rz_t$$

for some binary vector $z_t = \tilde{z}(t) \in \{0, 1\}^K$. Therefore

$$\text{span}(\mathcal{X}) \subseteq \text{col}(R),$$

which implies

$$\dim(\text{span}(\mathcal{X})) \leq \text{rank}(R) \leq K.$$

Now assume $V = 2^K$ and $\text{rank}(R) = K$. By [Corollary~1](#), the set $\{z_t : t \in \mathcal{V}\}$ equals the full hypercube $\{0, 1\}^K$. The span of the full hypercube in \mathbb{R}^K is all of \mathbb{R}^K , so

$$\text{span}(\{z_t\}) = \mathbb{R}^K.$$

Hence

$$\text{span}(\mathcal{X}) = R \mathbb{R}^K = \text{col}(R),$$

whose dimension is $\text{rank}(R) = K$. ■

This proposition highlights how severe the intervention is: before the first trainable layer, all token inputs lie in a subspace of dimension at most K , even though the model width is $d \gg K$. In our main experiments, this means that a 1024-dimensional Transformer receives inputs whose effective linear rank is only 16.

D.4. Frozen-lookup and table-free implementations

The same deterministic input map can be realized in two equivalent ways.

Frozen lookup.

One may precompute

$$E_{\text{fix}}[t] = x(t) \in \mathbb{R}^d$$

for all $t \in \mathcal{V}$ and store the resulting matrix

$$E_{\text{fix}} \in \mathbb{R}^{V \times d}$$

as a non-trainable lookup table. This is convenient in ablations because it can be implemented using the same software path as a standard embedding layer, except that the weights are fixed.

Table-free implementation.

Alternatively, one may compute the code directly from the integer token ID at runtime:

$$c(t)_j = \left\lfloor \frac{t}{2^j} \right\rfloor \bmod 2, \quad j = 0, \dots, K-1,$$

apply the affine recoding

$$\tilde{c}(t) = A c(t) \oplus b,$$

and then form

$$x(t) = R \tilde{c}(t).$$

This version has no input embedding table at all.

For a fixed choice of (c, A, b, R) , the frozen-lookup and table-free implementations are functionally identical: they produce exactly the same input vector $x(t)$ for every token t . The difference is purely computational, not statistical.

D.5. Parameterization and memory implications

A standard learned input embedding table contains

$$Vd$$

trainable parameters. By contrast, our proposed input interface contains **zero** trainable input parameters. In the table-free affine variant, the fixed specification

consists only of: (i) the minimal code width K , (ii) an invertible binary matrix $A \in GL(K, 2)$, (iii) a binary shift vector $b \in \{0, 1\}^K$, and (iv) the fixed lift R .

Thus the trainable input parameter count drops from Vd to 0 , while the fixed metadata scales as $O(K^2)$ rather than $O(Vd)$. In our main setting with $V = 65,536$ and $d = 1024$, a conventional learned input embedding table contains

$$65,536 \times 1024 = 67,108,864$$

trainable parameters, whereas the table-free minimal-code interface uses no trainable input parameters at all.

Summary.

Our method replaces the standard learned input lookup

$$t \mapsto E_{\text{in}}[t]$$

with the deterministic map

$$t \mapsto x(t) = R(A \text{bin}_K(t) \oplus b), \quad K = \lceil \log_2 V \rceil.$$

This preserves exact token identity, guarantees no collisions when R has full column rank, uses the minimal possible binary code width, and removes the trainable input embedding table entirely in the table-free implementation.

Experimental support, please [view the build logs](#) for errors. Generated by [L^AT_EX_ML](#).

Instructions for reporting errors

We are continuing to improve HTML versions of papers, and your feedback helps enhance accessibility and mobile support. To report errors in the HTML that will help us improve conversion and rendering, choose any of the methods listed below:

- Click the "Report Issue" (✖) button, located in the page header.

Tip: You can select the relevant text first, to include it in your report.

Our team has already identified [the following issues](#). We appreciate your time reviewing and reporting rendering errors we may not have found yet. Your efforts will help us improve the HTML versions for all readers, because disability should not be a barrier to accessing research. Thank you for your continued support in championing open access for all.

Have a free development cycle? Help support accessibility at arXiv! Our collaborators at LaTeXML maintain a [list of packages that need conversion](#), and welcome [developer contributions](#).