SCREENING IS ENOUGH

Anonymous authors
Paper under double-blind review

ABSTRACT

We call query–key relevance *absolute* when its values lie on a fixed bounded scale, depend on neither competing keys nor sequence length, require no sequence-length-dependent calibration, and can all be zero. To realize this notion, we introduce *screening*, whose explicit threshold transforms bounded query–key similarities into relevance values, enabling exact rejection, empty selection, and direct inspection on a common scale. In a controlled comparison of 12 attention mechanisms on a matched Transformer backbone, only screening maintains both low long-context perplexity and robust retrieval beyond the training context; notably, it does so without inference-time scaling. Building on screening, we introduce Multiscreen, a language-model architecture composed of parallel gated screening tiles. Multiscreen retains these long-context gains while achieving greater parameter efficiency, stronger general zero-shot downstream performance, lower training cost at larger scales, and lower model-side time to first token than Transformer baselines. We further develop a normalization design that keeps Multiscreen training stable even at a learning rate of 1 and show that an adapted version likewise stabilizes Transformer at the same learning rate.

1 INTRODUCTION

Handling contexts substantially longer than those seen during training remains a central challenge for large language models (LLMs). Training directly on long sequences is computationally expensive because the cost of Transformer self-attention grows quadratically with sequence length (Vaswani et al., 2017). Models are therefore often trained on relatively short contexts and expected to generalize to much longer ones at inference time, yet increasing the nominal context length does not guarantee that they can effectively use relevant information within it.

A key difficulty lies in how relevant information is selected from the context. In standard softmax attention, scores are unbounded and scale-dependent, while weights redistribute a fixed unit mass, so a key's apparent relevance depends on competing keys. Consequently, softmax cannot directly reject every key, and this competition can dilute relevant signals as contexts grow.

We call query–key relevance *absolute* when its values lie on a fixed bounded scale, depend on neither competing keys nor sequence length, require no sequence-length-dependent calibration, and can all be zero. To realize this notion, we introduce *screening*. Screening maps unit-normalized query–key similarities to nonnegative relevance values and uses an explicit threshold to assign exactly zero relevance to irrelevant keys. Because these values are not normalized across keys, every key can be rejected, making both selected keys and empty selections directly inspectable on a common scale.

Validation loss alone does not reveal whether a model can retrieve and use relevant information from long contexts. To directly evaluate retrieval, we introduce **ABCDigits**, a semantics-free key–value retrieval benchmark designed to isolate retrieval behavior from natural-language semantics and instruction-following effects.

We first isolate screening in a controlled comparison that changes only the attention mechanism on a matched Transformer backbone. Among 12 evaluated variants, only screening maintains both low long-context perplexity and robust retrieval beyond the training context; notably, it requires no inference-time scaling. Nested ablations isolate the contributions of screening's windowing and positional components, while showing that RoPE alone does not explain the abrupt failure of other global mechanisms beyond the training context.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Building on screening, we introduce **Multiscreen**, whose layers comprise parallel gated screening tiles with learned screening windows and minimal positional encoding (MiPE). MiPE becomes inactive for sufficiently large windows, so broad-context tiles do not extrapolate positional rotations beyond those seen during training.

Multiscreen models retain these long-context gains: they maintain low perplexity without inference-time scaling throughout evaluations extending to $32\times$ the training context and show little retrieval degradation as context length increases. In contrast, YaRN delays Transformer degradation but introduces a tradeoff between long-context extrapolation and perplexity at shorter contexts.

Beyond long-context behavior, the fitted scaling trends indicate that Multiscreen reaches similar validation loss with roughly 30% fewer parameters, and the 1.3B Multiscreen model improves general zero-shot downstream performance, outperforming its Transformer counterpart on eight of nine tasks. It also has lower training cost at larger scales and lower model-side time to first token (TTFT) across all evaluated model sizes. At the 1.3B scale, the Transformer has 2.5–$2.9\times$ the TTFT of Multiscreen across evaluated context lengths.

We further develop a normalization design that keeps Multiscreen training stable even at learning rate 1. Adapting this design to Transformer by combining query–key and pre-logit normalization likewise prevents divergence at the same learning rate, showing that this stability is transferable rather than specific to screening.

Our main contributions are as follows:

- We introduce *screening*, a mechanism for absolute query–key relevance, and show through a controlled comparison of 12 attention variants that it alone maintains low long-context perplexity and robust retrieval without inference-time scaling.
- We introduce **ABCDigits**, a semantics-free completion-based retrieval benchmark designed to isolate retrieval behavior without relying on natural-language semantics or instruction-following.
- We build **Multiscreen** from parallel gated screening tiles and show that it retains screening's long-context gains while improving parameter efficiency, general zero-shot downstream performance, training cost, and model-side TTFT relative to Transformer baselines.
- We develop a transferable normalization design that enables stable Multiscreen and Transformer training even at learning rate 1.

2 Related Work

Attention mechanisms and relevance. Alternatives to standard softmax modify its normalization, sparsity, or accessible key set. Normalized variants in our controlled comparison include $\mathrm{Softmax}_1$, which adds a null alternative to the denominator (Miller, 2023); Scalable-Softmax (SSMax), which sharpens attention as sequence length grows (Nakanishi, 2025); Entmax, which produces sparse normalized weights (Peters et al., 2019; Correia et al., 2019; Vasylenko et al., 2026); and Forgetting Attention, which applies data-dependent decay before normalization (Lin et al., 2025a). Sliding-window attention instead restricts the accessible keys (Beltagy et al., 2020). Sigmoid attention and ReLU^2 attention depart from normalized competition by weighting query–key pairs independently (Ramapuram et al., 2025; Hua et al., 2022). These mechanisms relax different constraints on total mass, sparsity, boundedness, pairwise independence, and length dependence; among those evaluated, screening alone combines all properties required for absolute relevance across context lengths. The formal distinctions and complete evaluated configurations are summarized in tables 7 and 8.

Long-context positional representations. Length generalization is also shaped by positional representations, including ALiBi, RoPE, and NoPE (Press et al., 2022; Su et al., 2024; Kazemnejad et al., 2023), as well as RoPE extension methods such as Position Interpolation, YaRN, and LongRoPE (Chen et al., 2023; Peng et al., 2024; Ding et al., 2024). Our controlled comparison includes ALiBi, NoPE as used by Forgetting Attention, and RoPE with YaRN scaling. Multiscreen instead uses minimal positional encoding (MiPE), which rotates only two dimensions for small screening windows and becomes inactive for sufficiently large windows.

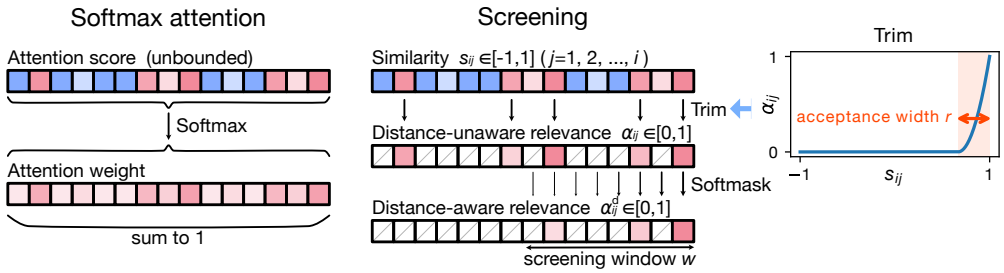


Figure 1: Conceptual comparison between softmax attention and screening. Each cell corresponds to a scalar value associated with an individual key. Colors indicate scalar values (blue: negative, red: positive; intensity indicates magnitude), while hatched cells denote zero values. Softmax attention maps unbounded attention scores to nonnegative attention weights that sum to one, redistributing mass across all keys. In contrast, screening maps bounded query–key similarities to distance-unaware relevance via Trim, which assigns exactly zero relevance to similarities below a threshold, and then applies Softmask to produce distance-aware relevance.

Retrieval evaluation and synthetic benchmarks. Long-context evaluation has shown that models may fail to use relevant information even when it lies within the nominal context window, as in lost-in-the-middle phenomena (Liu et al., 2024). Needle-in-a-haystack and passkey-style benchmarks probe long-context retrieval (Kamradt, 2023; Mohtashami & Jaggi, 2023). For natural-language needle benchmarks, semantic interference from the surrounding context can affect retrieval results (Shi & Penn, 2025). The original passkey evaluation uses a fixed natural-language prompt template, and subsequent work shows that apparent retrieval performance can depend strongly on its prompt and passkey configuration (Yuan et al., 2024). ABCDigits complements these benchmarks with a semantics-free completion task, a fixed key set, and a uniquely determined target value.

3 Model Architecture

Multiscreen realizes absolute query–key relevance through *screening*. Screening begins with unit-normalized query–key dot products, yielding similarities on the fixed interval $[-1, 1]$. Trim and Softmask transform each similarity independently into a relevance value, without normalization across keys or calibration to the sequence length, and these values are used directly for value aggregation. This construction permits exact rejection and empty selection, in contrast to the relative weighting imposed by softmax attention, as illustrated in fig. 1.

The overall architecture is illustrated in fig. 2a. Each layer contains parallel gated screening tiles (see fig. 2b), which project token representations into query, key, value, and gate vectors, apply a screening unit to retrieve relevant context, and gate the result using a GLU-style multiplicative mechanism (van den Oord et al., 2016; Dauphin et al., 2017; Shazeer, 2020) before projecting back to the embedding dimension. In the actual implementation, several operations are fused, and query–key pairs outside finite learned screening windows are skipped for efficiency.

3.1 Overview

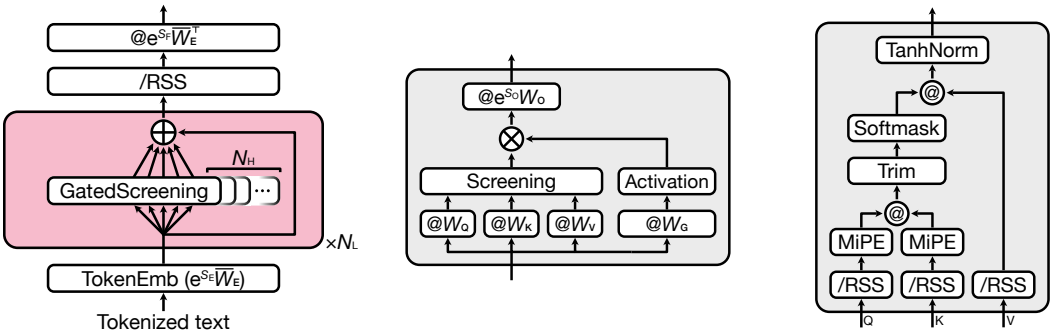
Let $W_{\mathrm{E}} = [\boldsymbol{e}_1; \ldots; \boldsymbol{e}_{|\mathcal{V}|}]$ denote the embedding matrix shared by the input and language-modeling head. We apply *embedding normalization* to each row, defining $\bar{\boldsymbol{e}}_j = \boldsymbol{e}_j / \|\boldsymbol{e}_j\|$. Given a token t_i, its input representation is

$$\boldsymbol{x}_i^{(0)} = \mathrm{e}^{s_{\mathrm{E}}} \bar{\boldsymbol{e}}_{t_i}, \tag{1}$$

where s_{E} is a learned scalar.

Multiscreen then applies N_{L} residual layers, each composed of N_{H} parallel gated screening tiles. Let $\Delta \boldsymbol{x}_i^{(\ell,h)}$ denote the update produced for position i by tile h in layer ℓ. The layer update is

$$\boldsymbol{x}_i^{(\ell)} = \boldsymbol{x}_i^{(\ell-1)} + \sum_{h=1}^{N_{\mathrm{H}}} \Delta \boldsymbol{x}_i^{(\ell,h)}. \tag{2}$$



(a) Multiscreen architecture (b) Gated screening tile (c) Screening unit

Figure 2: (a) Multiscreen comprises N_{L} residual layers with N_{H} parallel gated screening tiles per layer; $\overline{W}_{\mathrm{E}}$ is the row-wise unit-normalized embedding matrix shared by the input and language-modeling head, and $\mathrm{e}^{s_{\mathrm{E}}}$ and $\mathrm{e}^{s_{\mathrm{F}}}$ are learned input and logit scales. Final hidden states are unit-normalized before the head. (b) Each gated screening tile applies query, key, value, and gate projections, screening, gating, and an output projection with learned residual-output scale $\mathrm{e}^{s_{\mathrm{O}}}$. (c) The screening unit applies QKV unit-length normalization, MiPE, Trim, Softmask, and TanhNorm. In the diagrams, "@" denotes matrix multiplication and "/RSS" denotes row-wise unit-length normalization.

After the final layer, we apply *pre-logit normalization* (PLNorm), $\bar{\boldsymbol{x}}_i^{(N_{\mathrm{L}})} = \boldsymbol{x}_i^{(N_{\mathrm{L}})} / \|\boldsymbol{x}_i^{(N_{\mathrm{L}})}\|$, and compute logits using the normalized hidden state and embeddings:

$$z_{ij} = \bar{\boldsymbol{x}}_i^{(N_{\mathrm{L}})} \left(\mathrm{e}^{s_{\mathrm{F}}} \bar{\boldsymbol{e}}_j^{\top}\right), \qquad j \in \{1, \ldots, |\mathcal{V}|\}, \tag{3}$$

where s_{F} is a learned scalar. Neither embedding normalization nor PLNorm includes learned elementwise scales; each only projects its input onto the unit hypersphere. We refer to the combination of these normalizations and the learned input and logit scales $\mathrm{e}^{s_{\mathrm{E}}}$ and $\mathrm{e}^{s_{\mathrm{F}}}$ as the *spherical input–output (SIO) parameterization*. A standard softmax over logits gives next-token probabilities.

3.2 Screening Unit

A screening unit takes projected query, key, and value vectors $\boldsymbol{q}_i, \boldsymbol{k}_i \in \mathbb{R}^{1\times d_{\mathrm{K}}}$ and $\boldsymbol{v}_i \in \mathbb{R}^{1\times d_{\mathrm{V}}}$, and returns a context-dependent representation $\boldsymbol{u}_i \in \mathbb{R}^{1\times d_{\mathrm{V}}}$ (see fig. 2c). Each unit has two learned scalar parameters, s_{w} and s_{r}, which define

$$w = \mathrm{e}^{s_{\mathrm{w}}} + 1, \qquad r = \sigma(s_{\mathrm{r}}), \tag{4}$$

where σ denotes the sigmoid function. Here w is the screening window and r is the acceptance width for similarity. Under our *default inference rule*, if the learned screening window exceeds the maximum sequence length seen during training, we set $w = \infty$. All inference evaluations use this rule unless a finite-window ablation is explicitly identified.

We first apply unit-length normalization to queries, keys, and values:

$$\bar{\boldsymbol{q}}_i = \frac{\boldsymbol{q}_i}{\|\boldsymbol{q}_i\|}, \qquad \bar{\boldsymbol{k}}_i = \frac{\boldsymbol{k}_i}{\|\boldsymbol{k}_i\|}, \qquad \bar{\boldsymbol{v}}_i = \frac{\boldsymbol{v}_i}{\|\boldsymbol{v}_i\|}. \tag{5}$$

In implementation, the denominator in each unit-length normalization is clipped from below by a small positive constant for numerical stability. Normalizing queries and keys bounds query–key similarities in $[-1, 1]$, making thresholding well-defined. Prior work shows that token contributions depend on vector norms as well as attention weights (Kobayashi et al., 2020; Guo et al., 2024). We therefore normalize value vectors to unit length so that differences in their norms do not directly determine their contribution magnitudes.

Before computing similarity, we apply minimal positional encoding (MiPE; see fig. 3) to $\bar{\boldsymbol{q}}_i$ and $\bar{\boldsymbol{k}}_i$. MiPE is a RoPE-like rotation (Su et al., 2024) applied only to the first two dimensions, with a rotation rate modulated by the learned screening window w. For position i, MiPE rotates by angle $\phi(i, w) = i\theta(w)$, where

$$\theta(w) = \begin{cases} \frac{\pi}{w}\left(1 - \frac{w}{w_{\mathrm{th}}}\right), & w < w_{\mathrm{th}}, \\ 0, & w \geq w_{\mathrm{th}}. \end{cases} \tag{6}$$

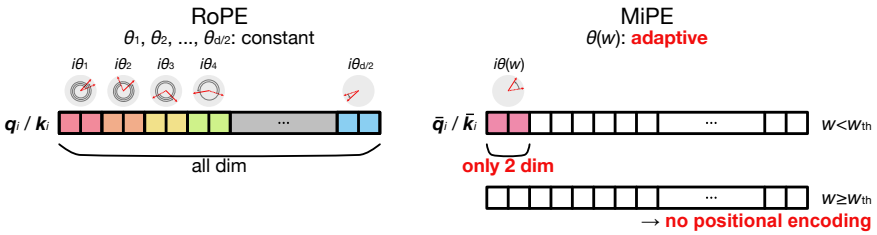


Figure 3: Comparison between RoPE and MiPE. RoPE applies rotations across all dimensions with fixed frequencies. MiPE applies rotation only to two dimensions and adapts the rotation angle based on the learned screening window w, becoming inactive when $w \geq w_{\text{th}}$.

Here w_{th} is a fixed MiPE threshold. Thus MiPE is active only for small screening windows and becomes the identity when $w \geq w_{\text{th}}$.

Let $\tilde{\boldsymbol{q}}_i$ and $\tilde{\boldsymbol{k}}_j$ denote the resulting position-encoded query and key vectors. The bounded similarity is then

$$s_{ij} = \tilde{\boldsymbol{q}}_i \tilde{\boldsymbol{k}}_j^\top, \qquad s_{ij} \in [-1, 1]. \tag{7}$$

We apply a Trim transform (see fig. 1) to obtain distance-unaware relevance:

$$\alpha_{ij} = \left[\max\left(1 - \frac{1 - s_{ij}}{r}, 0\right) \right]^2. \tag{8}$$

This sets relevance exactly to zero whenever $s_{ij} \leq 1 - r$, and smoothly emphasizes similarities close to one.

We next apply a causal distance-aware Softmask parameterized by the screening window w:

$$m_{ij}(w) = \begin{cases} 1 - 2\left(\frac{i-j}{w}\right)^2, & 0 \leq i - j < \frac{w}{2}, \\ 2\left(1 - \frac{i-j}{w}\right)^2, & \frac{w}{2} \leq i - j < w, \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

This piecewise-quadratic form provides a cosine-like taper while avoiding trigonometric evaluation. For $w = \infty$, we take $m_{ij}(w) = 1$ at every causal position. We then define distance-aware relevance

$$\alpha_{ij}^{\text{d}} = \alpha_{ij} m_{ij}(w). \tag{10}$$

A key contributes only if it survives both content-based and distance-based screening.

Finally, the screening unit aggregates the surviving values:

$$\boldsymbol{h}_i = \sum_{j \leq i} \alpha_{ij}^{\text{d}} \bar{\boldsymbol{v}}_j. \tag{11}$$

Because the relevance values $\{\alpha_{ij}^{\text{d}}\}_{j \leq i}$ are not constrained to sum to one, the unit can represent the absence of relevant context. To softly bound the output norm while preserving direction, we define the function

$$\text{TanhNorm}(\boldsymbol{x}) = \frac{\tanh \|\boldsymbol{x}\|}{\|\boldsymbol{x}\|} \boldsymbol{x}. \tag{12}$$

For sufficiently small $\|\boldsymbol{x}\|$, we use the limiting value $\text{TanhNorm}(\boldsymbol{x}) = \boldsymbol{x}$. The screening output is

$$\boldsymbol{u}_i = \text{TanhNorm}(\boldsymbol{h}_i). \tag{13}$$

Overall, screening defines relevance independently for each key, can assign exactly zero relevance to irrelevant keys, and aggregates only the surviving values.

3.3 Gated Screening Tile

A gated screening tile (see fig. 2b) combines screening-based context retrieval with input-dependent feature selection. Given token representations $\boldsymbol{x}_i \in \mathbb{R}^{1\times d_{\mathrm{E}}}$, the tile computes

$$\boldsymbol{q}_i = \boldsymbol{x}_i W_{\mathrm{Q}}, \quad \boldsymbol{k}_i = \boldsymbol{x}_i W_{\mathrm{K}}, \quad \boldsymbol{v}_i = \boldsymbol{x}_i W_{\mathrm{V}}, \quad \boldsymbol{g}_i = \boldsymbol{x}_i W_{\mathrm{G}}, \tag{14}$$

where $W_{\mathrm{Q}}, W_{\mathrm{K}} \in \mathbb{R}^{d_{\mathrm{E}}\times d_{\mathrm{K}}}$ and $W_{\mathrm{V}}, W_{\mathrm{G}} \in \mathbb{R}^{d_{\mathrm{E}}\times d_{\mathrm{V}}}$.

Using the screening unit from section 3.2, the tile retrieves

$$\boldsymbol{u}_i = \mathrm{Screening}\big(\boldsymbol{q}_i, \{\boldsymbol{k}_j\}_{j\le i}, \{\boldsymbol{v}_j\}_{j\le i}, w, r\big)_i. \tag{15}$$

In parallel, it computes a gate via a nonlinear activation

$$\hat{\boldsymbol{g}}_i = \tanh(\mathrm{SiLU}(\boldsymbol{g}_i)), \tag{16}$$

where SiLU is the sigmoid-weighted linear unit (Elfwing et al., 2018). The tile output is

$$\Delta\boldsymbol{x}_i = (\boldsymbol{u}_i \odot \hat{\boldsymbol{g}}_i)\,(\mathrm{e}^{s_{\mathrm{O}}} W_{\mathrm{O}})\,, \tag{17}$$

with $W_{\mathrm{O}} \in \mathbb{R}^{d_{\mathrm{V}}\times d_{\mathrm{E}}}$ and learned scalar s_{O}.

This tile can be viewed as a generalization of GLU-style modules (van den Oord et al., 2016; Dauphin et al., 2017; Shazeer, 2020) and recent gated-attention variants (Qiu et al., 2025), replacing the usual linear or attention-based transformation with screening-based context aggregation. Thus, each tile jointly performs context selection, feature gating, and projection back to the model space. Initialization details are provided in Appendix E. Architectural analyses of MiPE, TanhNorm, and the output gate are provided in Appendices P to R.

4 Experiments

We compare Multiscreen with Transformer baselines across long-context language modeling, retrieval, absolute relevance, optimization stability, parameter efficiency, general zero-shot downstream performance, training cost, and model-side TTFT.

4.1 Experimental Setup

Models in the main experiments are pretrained on SlimPajama (Soboleva et al., 2023) for 2^{38} tokens using sequences of length 2^{12}. Both architectures use AdamW (Loshchilov & Hutter, 2019) without weight decay or gradient clipping. Transformer uses model-size-dependent learning rates, whereas Multiscreen uses the larger learning rate 2^{-4}, enabled by its improved stability. Full model configurations, training hyperparameters, and implementation details are provided in Appendix A.

4.2 Long-Context Perplexity

We evaluate long-context language modeling using position-dependent perplexity. As shown in fig. 4, Multiscreen requires no inference-time scaling and maintains low perplexity up to $32\times$ the training context length. YaRN delays the onset of Transformer degradation, but larger scaling factors increase perplexity at shorter contexts, revealing a tradeoff that is absent from Multiscreen. Additional evaluation details and seed-averaged smaller-model results are provided in Appendix L.

4.3 Retrieval Ability

To isolate retrieval ability, we introduce **ABCDigits**, a synthetic key–value retrieval benchmark in a semantics-free setting. The task presents a shuffled list of mappings from uppercase letters to six-digit integers (e.g., `A=967892`), followed by a query (e.g., `L=`) requiring the model to retrieve the corresponding value. The target mapping appears exactly once in the context, requiring the model to locate a unique occurrence. The number of distinct keys is fixed to avoid key-cardinality confounds, while non-target frequencies are highly skewed so that the unique target does not stand out by rarity alone. Code for generating ABCDigits prompts is included in the anonymized supplementary material. Additional task-construction and evaluation details are provided in Appendix M.

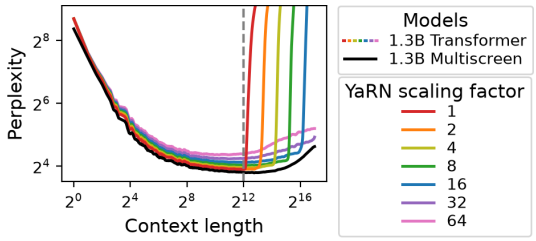
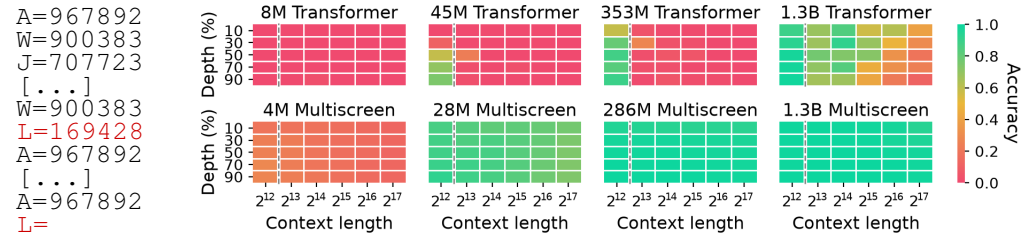


Figure 4: Position-dependent perplexity of the 1.3B models. Colored curves show Transformer with different YaRN scaling factors, whereas the black curve shows Multiscreen, which requires no inference-time scaling. The dashed line marks the pretraining context length (2^{12}).



(a) Prompt example

(b) Retrieval accuracy

Figure 5: (a) Example ABCDigits prompt. (b) ABCDigits exact-match accuracy across model sizes, context lengths, and target depths. Each cell averages 1,000 instances; results are further averaged across three independently trained models, except at 1.3B, where one model is used. Transformer uses context-matched YaRN scaling, while Multiscreen uses our default inference rule. The dashed line marks the pretraining context length (2^{12}).

We evaluate exact-match retrieval under greedy decoding at target context lengths from 2^{12} to 2^{17} and varying target depths. As shown in fig. 5, Multiscreen shows little degradation with increasing context length and maintains high retrieval accuracy beyond the training context length. Notably, even at the training context length, the 286M Multiscreen model outperforms the 1.3B Transformer evaluated with context-matched YaRN scaling when averaged over target depths (98.33% vs. 95.18%), with the gap becoming more pronounced at longer contexts. In contrast, Transformer accuracy degrades increasingly beyond the training context despite using context-matched YaRN scaling.

With roughly one fifth the parameters, the 286M Multiscreen model achieves substantially stronger long-context retrieval than the 1.3B Transformer despite higher validation loss, showing that validation loss alone does not capture retrieval ability. On the passkey retrieval task introduced by Mohtashami & Jaggi (2023), both the 286M and 1.3B Multiscreen models achieve perfect accuracy at every evaluated context length and target depth (Appendix N).

4.4 Isolating the Screening Mechanism

To isolate screening from the remaining Multiscreen architecture, we train a controlled suite that matches the 353M Transformer in all model dimensions (including Q/K/V dimensions), initialization, and training hyperparameters, changing only the attention mechanism. The alternatives include $\mathrm{Softmax}_1$, SSMax, Entmax ($\alpha = 1.5$), Sigmoid, ReLU^2, ALiBi, Forgetting Attention, and sliding-window attention, spanning normalization-based, sparse, noncompetitive, distance-decayed, and fixed-window mechanisms. The Screening variant contains the screening unit—QKV unit-length normalization, MiPE, Trim, Softmask, and TanhNorm—but excludes the output gate and the other components of the gated screening tile. We include two nested ablations: Screening (RoPE) replaces MiPE with RoPE, while Screening (no window) additionally removes the learned window and Softmask.

Screening is the only evaluated mechanism that maintains both low long-context perplexity and robust retrieval across all evaluated context lengths and target depths without inference-time scaling (fig. 6). ALiBi, Forgetting Attention, and sliding-window attention maintain relatively stable long-context

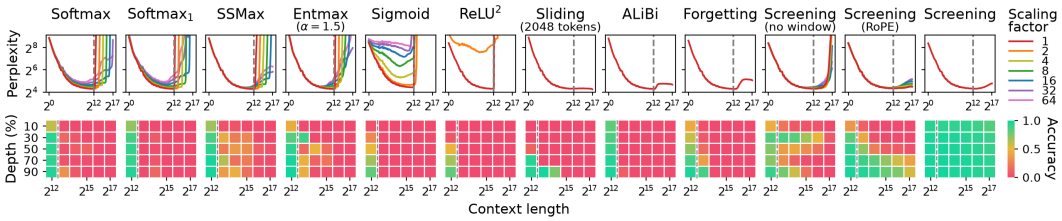
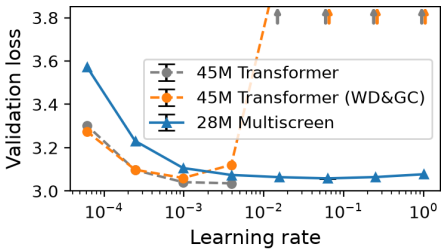


Figure 6: Controlled attention-mechanism comparison on the 353M Transformer backbone with all other settings fixed. The top row shows position-dependent perplexity; the bottom row shows ABCDigits retrieval accuracy. Each variant uses one trained model. In the top row, red denotes no inference-time scaling; for methods requiring scaling, red corresponds to factor 1 and the other colors to the indicated factors. For ABCDigits, methods requiring scaling use a factor equal to the evaluation-to-training context-length ratio. For ReLU^2, scaling factors 4 and above exceed the plotted perplexity range. Dashed lines mark the pretraining context length (2^{12}).



(a) Learning-rate sweep.

Configuration	Outcome
Transformer	Diverged
Transformer + SIO	Diverged
Transformer + PLNorm	Diverged
Transformer + QKNorm	Diverged
Transformer + QKNorm + SIO	Stable
Transformer + QKNorm + PLNorm	Stable
Multiscreen	Stable

(b) Selected outcomes at learning rate 1.

Figure 7: Learning-rate stability of Transformer and Multiscreen. (a) Validation loss across learning rates, with arrows at the upper plot boundary indicating divergence. Each point uses one training run, except Multiscreen at learning rate 2^{-4}, which reuses the three runs from the scaling experiments and shows their mean and standard deviation. (b) Training outcomes at learning rate 1 after 2^{36} tokens (2^{14} steps), where "Stable" indicates completion of this budget without divergence. PLNorm denotes pre-logit normalization. Complete results are provided in Appendix I.

perplexity yet fail to retrieve distant targets, showing that perplexity alone is insufficient to assess long-range retrieval. Even at scaling factor 1, Screening (no window) avoids the abrupt perplexity degradation beyond the training context seen in the other full-context RoPE-based mechanisms, showing that RoPE alone does not explain this failure. Screening (RoPE) markedly outperforms Screening (no window), demonstrating the importance of the learned window and its associated Softmask. Full Screening further improves retrieval by replacing RoPE with MiPE, indicating that MiPE provides additional robustness. Together, these results isolate the long-context behavior to the screening unit itself rather than the output gate or the wider Multiscreen architecture. Formal property comparisons, complete configurations, and implementation details are provided in Appendix O.

4.5 Learning-Rate Stability

We sweep the learning rate for a 45M Transformer and a 28M Multiscreen while keeping all other training settings fixed within each architecture. For comparison, we also evaluate a Transformer control with weight decay 0.1 and gradient clipping at 1.0.

As shown in fig. 7a, both baseline Transformer configurations become unstable and eventually diverge beyond moderate learning rates, whereas Multiscreen remains stable at learning rate 1. This large-learning-rate stability transfers to Transformer when QKNorm is combined with either PLNorm or SIO (fig. 7b). At learning rate 1, QKNorm, PLNorm, and SIO each fail to prevent Transformer divergence when applied alone. These results show that large-learning-rate stability depends on how normalization components are combined and placed. Additional training-loss and gradient-norm analyses are provided in Appendices G and H. Complete Transformer and Multiscreen ablations are provided in Appendix I.

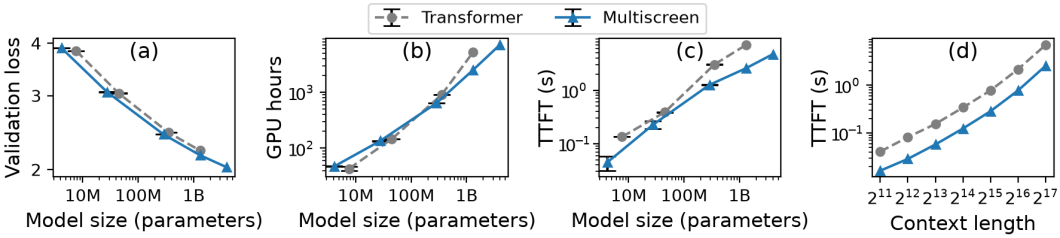


Figure 8: Scaling and computational efficiency. (a) Validation loss and (b) training cost in GPU hours across model sizes, with each model trained on 2^{38} tokens. (c) Model-side TTFT across model sizes at context length 2^{17}. (d) Model-side TTFT across context lengths for the 1.3B models. Error bars indicate standard deviation across independently trained models where multiple runs are available.

4.6 Scaling and Computational Efficiency

Scaling and training cost. We evaluate scaling behavior by comparing models trained on a fixed budget of 2^{38} tokens. For smaller models, we report mean and standard deviation over three runs with different random seeds; larger models are evaluated with a single run. We report training cost in GPU hours, computed as elapsed wall-clock training time multiplied by the number of GPUs used. All training runs use NVIDIA GH200 GPUs, with data-parallel configurations detailed in Appendix A.

As shown in fig. 8a, the fitted scaling trends indicate that Multiscreen requires roughly 30% fewer parameters to reach a similar validation loss. We evaluate general zero-shot downstream performance of the 1.3B models on nine common benchmarks. Multiscreen outperforms Transformer on eight of the nine tasks and improves the unweighted mean accuracy from 52.41% to 53.67%, with per-task results in Appendix B and additional scaling analyses in Appendix F. Figure 8b further shows lower training cost at larger model scales, with a particularly pronounced advantage for the 1.3B models.

Model-side TTFT. We define model-side time to first token (TTFT) as the latency of a full-context forward pass that processes all positions through the model layers but applies the vocabulary projection only at the final position to produce first-token logits. This metric isolates model execution and excludes tokenization, sampling, request queuing, and network or serving overhead. We evaluate TTFT across model sizes at context length 2^{17} and across context lengths from 2^{11} to 2^{17} for the 1.3B models. All measurements use a single NVIDIA GH200 GPU; the complete measurement protocol is provided in Appendix S. As shown in fig. 8c, Multiscreen achieves consistently lower model-side TTFT than Transformer across model sizes. The context-length sweep in fig. 8d likewise shows that Transformer has 2.5–$2.9\times$ the TTFT of Multiscreen at every evaluated input length.

5 Conclusion

In this work, we introduced *screening* to realize absolute query–key relevance on a fixed bounded scale, independently of competing keys and sequence length. Its explicit threshold permits exact rejection and empty selection without sequence-length-dependent calibration, making relevance directly inspectable.

Using ABCDigits and long-context perplexity, a controlled comparison of 12 attention mechanisms showed that only screening maintained robust retrieval and low perplexity beyond the training context without inference-time scaling. Multiscreen, built from parallel gated screening tiles, retained these long-context gains across all evaluated context lengths.

Beyond long-context behavior, Multiscreen improved parameter efficiency, general zero-shot downstream performance, training cost at larger scales, and model-side time to first token. We also developed a normalization design that kept Multiscreen stable even at a learning rate of 1 and whose adaptation likewise stabilized Transformer.

These results show that directly inspectable absolute relevance can coexist with strong long-context behavior, general zero-shot downstream performance, and computational efficiency, offering an alternative to competitive renormalization and length-dependent inference-time calibration.

ETHICS STATEMENT

This work does not involve human participants or the collection of new personal data, and ABCDigits is constructed entirely from synthetic letter–digit mappings. Our models are trained and evaluated on publicly released datasets whose provenance and licenses are documented in Appendix T. Because SlimPajama contains web-derived data, it may reflect biases or harmful content and may contain personal or copyrighted information; we did not independently audit the complete corpus for these issues.

AI USE STATEMENT

Generative AI tools were used for literature retrieval; feedback on experimental design and result interpretation; implementation and debugging of selected baseline and diagnostic methods; and drafting, editing, and presentation of the manuscript. The Multiscreen architecture and its core implementation were developed by the authors without generative AI assistance. All AI-assisted text, code, and scientific suggestions were reviewed by the authors, experimental interpretations were checked against the reported results, and suggested references were verified against the original sources. The authors take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

The Multiscreen architecture and screening mechanism are fully specified in section 3. Appendices A and C to E provide the training setup, model configurations, baseline details, and initialization schemes. Additional scaling and downstream results are provided in Appendices B and F. Learning-rate analyses and detailed protocols for the long-context, retrieval, attention-mechanism comparison, and component-analysis experiments are provided in Appendices G to I and L to R. The TTFT measurement protocol and the datasets and software used in this work are documented in Appendices S and T. Code for generating ABCDigits prompts is included in the anonymized supplementary material.

REFERENCES

Arize AI. Needle in a haystack - pressure testing LLMs, 2023. URL `https://github.com/Arize-ai/LLMTest_NeedleInAHaystack2`. Accessed on Jan 19, 2024.

Simran Arora, Sabri Eyuboglu, Aman Timalsina, Isys Johnson, Michael Poli, James Zou, Atri Rudra, and Christopher Ré. Zoology: Measuring and improving recall in efficient language models. In *International Conference on Learning Representations*, 2024.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, 2023.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. PIQA: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

Sidney Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, Usvsn Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. GPT-NeoX-20B: An open-source autoregressive language model. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, 2022.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*, 2023.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Gonçalo M. Correia, Vlad Niculae, and André F.T. Martins. Adaptively sparse transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

Tri Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations*, 2024.

Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *International Conference on Machine Learning*, 2017.

Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. LongRoPE: Extending LLM context window beyond 2 million tokens. In *International Conference on Machine Learning*, 2024.

Stefan Elfwing, Eiji Uchibe, and Kenji Doya. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural networks*, 107:3–11, 2018.

Nuno M. T. Gonçalves, Hugo Pitorro, Vlad Niculae, Edoardo M. Ponti, Lei Li, André F. T. Martins, and Marcos V. Treviso. AdaSplash-2: Faster differentiable sparse attention. In *International Conference on Machine Learning*, 2026.

Alex Graves, Greg Wayne, and Ivo Danihelka. Neural Turing machines. *arXiv preprint arXiv:1410.5401*, 2014.

Zhiyu Guo, Hidetaka Kamigaito, and Taro Watanabe. Attention score is not all you need for token importance indicator in KV cache reduction: Value also matters. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024.

Alexander Hägele, Elie Bakouch, Atli Kosson, Loubna Ben allal, Leandro Von Werra, and Martin Jaggi. Scaling laws and compute-optimal training beyond fixed training durations. In *Advances in Neural Information Processing Systems*, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021.

Alex Henry, Prudhvi Raj Dachapally, Shubham Shantaram Pawar, and Yuxuan Chen. Query-key normalization for transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020.

Weizhe Hua, Zihang Dai, Hanxiao Liu, and Quoc V. Le. Transformer quality in linear time. In *International Conference on Machine Learning*, 2022.

Greg Kamradt. Needle in a haystack - pressure testing LLMs, 2023. URL `https://github.com/gkamradt/LLMTest_NeedleInAHaystack`. Accessed on Jan 19, 2024.

Amirhossein Kazemnejad, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Payel Das, and Siva Reddy. The impact of positional encoding on length generalization in transformers. In *Advances in Neural Information Processing Systems*, 2023.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. Attention is not only a weight: Analyzing transformers with vector norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020.

Zhixuan Lin, Evgenii Nikishin, Xu He, and Aaron Courville. Forgetting Transformer: Softmax attention with a forget gate. In *International Conference on Learning Representations*, 2025a.

Zhixuan Lin, Johan Obando-Ceron, Xu Owen He, and Aaron Courville. Adaptive computation pruning for the forgetting transformer. In *Second Conference on Language Modeling*, 2025b.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018.

Evan Miller. Attention is off by one. `https://www.evanmiller.org/attention-is-off-by-one.html`, 2023. Accessed: 2026-09-21.

Amirkeivan Mohtashami and Martin Jaggi. Random-access infinite context length for transformers. In *Advances in Neural Information Processing Systems*, 2023.

Ken M. Nakanishi. Scalable-Softmax is superior for attention. *arXiv preprint arXiv:2501.19399*, 2025.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, 2019.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. YaRN: Efficient context window extension of large language models. In *International Conference on Learning Representations*, 2024.

Ben Peters, Vlad Niculae, and André F.T. Martins. Sparse sequence-to-sequence models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

Ofir Press, Noah A. Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations*, 2022.

Zihan Qiu, Zekun Wang, Bo Zheng, Zeyu Huang, Kaiyue Wen, Songlin Yang, Rui Men, Le Yu, Fei Huang, Suozhi Huang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free. In *Advances in Neural Information Processing Systems*, 2025.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019. URL `https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf`.

Jack W. Rae, Anna Potapenko, Siddhant M. Jayakumar, Chloe Hillier, and Timothy P. Lillicrap. Compressive transformers for long-range sequence modelling. In *International Conference on Learning Representations*, 2020.

Jason Ramapuram, Federico Danieli, Eeshan Dhekane, Floris Weers, Dan Busbridge, Pierre Ablin, Tatiana Likhomanenko, Jagrit Digani, Zijin Gu, Amitis Shidani, and Russ Webb. Theory, analysis, and best practices for sigmoid self-attention. In *International Conference on Learning Representations*, 2025.

Vijay Sadashivaiah, Georgios Dasoulas, Judith Mueller, and Soumya Ghosh. Better models, faster training: Sigmoid attention for single-cell foundation models. *arXiv preprint arXiv:2604.27124*, 2026.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. WinoGrande: An adversarial winograd schema challenge at scale. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

Noam Shazeer. GLU variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

Ken Shi and Gerald Penn. Semantic masking in a needle-in-a-haystack test for evaluating large language model long-text capabilities. In *Proceedings of the First Workshop on Writing Aids at the Crossroads of AI, Cognitive Science and NLP (WRAICOGS 2025)*, pp. 16–23, 2025.

Daria Soboleva, Faisal Al-Khateeb, Robert Myers, Jacob R Steeves, Joel Hestness, and Nolan Dey. SlimPajama: A 627B token cleaned and deduplicated version of RedPajama. `https://cerebras.ai/blog/slimpajama-a-627b-token-cleaned-and-deduplicated-version-of-redpajama`, 2023.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Lintang Sutawika, Hailey Schoelkopf, Leo Gao, Baber Abbasi, Stella Biderman, Jonathan Tow, ben fattori, Charles Lovering, farzanehnakhaee70, Jason Phang, Anish Thite, Fazz, Aflah, Niklas, Thomas Wang, sdtblck, nopperl, gakada, researcher2, tttyuntian, Julen Etxaniz, James A. Michaelov, Chris, Chessing234, Hanwool Albert Lee, Janna, Leonid Sinev, Khalid, Kiersten Stokes, and Zdeněk Kasner. Eleutherai/lm-evaluation-harness: v0.4.12, 2026. URL `https://doi.org/10.5281/zenodo.20122284`.

Philippe Tillet, H. T. Kung, and David Cox. Triton: An intermediate language and compiler for tiled neural network computations. In *Proceedings of the 3rd ACM SIGPLAN International Workshop on Machine Learning and Programming Languages*, 2019.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Aaron van den Oord, Nal Kalchbrenner, Oriol Vinyals, Lasse Espeholt, Alex Graves, and Koray Kavukcuoglu. Conditional image generation with PixelCNN decoders. In *Advances in Neural Information Processing Systems*, 2016.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

Pavlo Vasylenko, Hugo Pitorro, André F. T. Martins, and Marcos Treviso. Long-context generalization with sparse attention. In *The Fourteenth International Conference on Learning Representations*, 2026.

Maurice Weber, Daniel Fu, Quentin Anthony, Yonatan Oren, Shane Adams, Anton Alexandrov, Xiaozhong Lyu, Huu Nguyen, Xiaozhe Yao, Virginia Adams, Ben Athiwaratkun, Rahul Chalamala, Kezhen Chen, Max Ryabinin, Tri Dao, Percy Liang, Christopher Ré, Irina Rish, and Ce Zhang. RedPajama: an open dataset for training large language models. In *Advances in Neural Information Processing Systems*, 2024.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, 2017.

Kaiyue Wen, Zhiyuan Li, Jason Wang, David Hall, Percy Liang, and Tengyu Ma. Understanding Warmup-Stable-Decay learning rates: A river valley loss landscape view. In *International Conference on Learning Representations*, 2025.

Kazuki Yano, Shun Kiyono, Sosuke Kobayashi, Sho Takase, and Jun Suzuki. Pre-training LLM without learning rate decay enhances supervised fine-tuning. In *International Conference on Learning Representations*, 2026.

Jiayi Yuan, Hongyi Liu, Shaochen Zhong, Yu-Neng Chuang, Songchen Li, Guanchu Wang, Duy Le, Hongye Jin, Vipin Chaudhary, Zhaozhuo Xu, Zirui Liu, and Xia Hu. KV cache compression, but what must we give in return? a comprehensive benchmark of long context capable approaches. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

A Experimental Setup

This section describes the primary Transformer and Multiscreen pretraining configurations under matched training data and token budgets. Controlled comparisons and ablations document any deviations from this setup in their respective sections.

Tokenization. We adopt the GPT-2 tokenizer (Radford et al., 2019) with a vocabulary of 50,257 tokens.

Training Data. We pretrain the main-experiment models on the SlimPajama (Soboleva et al., 2023) dataset, a cleaned and deduplicated subset of RedPajama (Weber et al., 2024). After tokenization, documents are packed back-to-back into fixed-length input sequences without inserting EOS tokens, while document boundaries are retained as metadata. For every nonfinal token in a document, the prediction target is the next token in that document; for the final token, the target is EOS. For example, documents (a, b, c, d) and (e, f, g, h, i) with sequence length 8 produce input (a, b, c, d, e, f, g, h) and targets $(b, c, d, \langle\mathrm{EOS}\rangle, f, g, h, i)$. The corresponding document IDs are $(0, 0, 0, 0, 1, 1, 1, 1)$ and are used only when document masking is enabled. The tokenized documents contain approximately 627 billion tokens in total. For these experiments, we use 2^{38} tokens (approximately 44% of the dataset), obtained by partitioning the tokenized corpus into contiguous sequences of length 2^{12} and selecting 2^{26} sequences at random. For validation, we use 2^{26} tokens from the SlimPajama validation split, packed into sequences of the same length 2^{12} as in training.

Document Masking. When document masking is enabled, each query can attend only to causal keys with the same document ID. The primary Transformer and Multiscreen runs use this mask. The controlled attention-mechanism comparison disables document masking for all variants because several fused kernels do not support it, as detailed in Appendix O.

Packing Rationale. This construction assigns exactly one meaningful prediction target to every input token. If EOS tokens were instead inserted into the input, each EOS position would either be trained to predict the first token of the next, unrelated document or be excluded from the cross-entropy loss. The former trains the model to predict the first token of an unrelated document at an EOS position, although that token is not a meaningful continuation of the preceding document. The latter makes the number of supervised targets per nominal token batch depend on the number of packed document boundaries. In our construction, document-final tokens predict EOS and all other tokens predict the next token in the same document, so every input token is paired with exactly one meaningful supervised target. With document masking, every prediction conditions only on tokens from the same document. Without it, a token's causal context may include a preceding unrelated document, so the model is trained to predict within the current document in the presence of unrelated preceding context, while document-final tokens still mark document boundaries by predicting EOS.

Model Sizes. To analyze scaling behavior, we train both Transformer and Multiscreen models across multiple parameter scales.

For the Transformer baseline, we adopt a LLaMA-style architecture (Touvron et al., 2023), with architecture hyperparameters based on those used in Pythia (Biderman et al., 2023), and apply weight tying between the input embedding and the language modeling head. We follow Pythia configurations where applicable, including model sizes and peak learning rates. Detailed configurations of the Transformer baseline are provided in Appendix D.

For Multiscreen, the architecture hyperparameters and initialization scheme are provided in Appendices C and E.

Training Setup. All main-experiment models are pretrained with a sequence length of 2^{12} (4,096). We use a global batch size of 2^{22} tokens. Together with the 2^{38}-token budget, this gives 2^{16} optimizer steps per run.

All models are optimized using AdamW (Loshchilov & Hutter, 2019) with $(\beta_1, \beta_2) = (0.9, 0.95)$ and $\epsilon = 10^{-8}$. We linearly warm up the learning rate for 2^{12} steps, after which it is kept constant. This warmup-then-constant schedule keeps the optimization protocol simple and comparable across

architectures under a fixed token budget, and is consistent with recent work studying long stable or decay-free learning-rate phases (Hägele et al., 2024; Wen et al., 2025; Yano et al., 2026). We train using bfloat16 automatic mixed precision.

Neither architecture uses weight decay or gradient clipping in the main experiments. For Transformer, we selected this configuration because it achieved lower validation loss than the Pythia-style setting with weight decay 0.1 and gradient clipping at 1.0. The learning-rate sweep in section 4.5 includes the latter setting as an explicit Transformer control. For Transformer, we use RoPE with $\theta = 10{,}000$ and model-size-dependent learning rates based on Pythia (Biderman et al., 2023). The learning rate is set to the peak value used in Pythia for each model scale: 1×10^{-3} for 8M and 45M models, 3×10^{-4} for 353M, and 2×10^{-4} for 1.3B.

For Multiscreen, we use the same optimizer configuration as the Transformer baseline apart from the learning rate. We use a learning rate of 2^{-4} (i.e., 0.0625), substantially larger than those used for Transformer, enabled by the improved stability of Multiscreen demonstrated in section 4.5.

Training Hardware. All scaling experiments use NVIDIA GH200 GPUs. The 8M, 45M, and 353M Transformer models and the 4M, 28M, and 286M Multiscreen models use 16 GPUs, whereas both 1.3B models and the 4B Multiscreen model use 64 GPUs.

Implementations. Transformer uses `torch.nn.attention.varlen.varlen_attn`, backed by PyTorch's built-in FlashAttention-2 kernel (Dao, 2024), for both training and inference, while Multiscreen uses custom Triton kernels (Tillet et al., 2019) tailored to its screening mechanism. All experiments use PyTorch 2.13.0, CUDA 13.2, and Triton 3.7.1. The hardware and protocol used for TTFT measurements are reported in Appendix S.

B Zero-Shot Downstream Evaluation

We evaluate the 1.3B Transformer and Multiscreen models in the zero-shot setting using version 0.4.12 of the Language Model Evaluation Harness (Sutawika et al., 2026), with `num_fewshot=0` for every task. We report normalized accuracy for HellaSwag (Zellers et al., 2019), PIQA (Bisk et al., 2020), ARC-Easy and ARC-Challenge (Clark et al., 2018), SciQ (Welbl et al., 2017), and OpenBookQA (Mihaylov et al., 2018). We report ordinary accuracy for WinoGrande (Sakaguchi et al., 2020), BoolQ (Clark et al., 2019), and MMLU (Hendrycks et al., 2021). The results are summarized in table 1, with the average computed as the unweighted mean across all nine tasks.

Table 1: Zero-shot downstream accuracy (%) of the 1.3B models, with the better result for each benchmark shown in bold.

	Accuracy (%) ↑	
Benchmark	Transformer	Multiscreen
HellaSwag	55.58	**59.02**
PIQA	72.91	**73.99**
ARC-Easy	52.82	**55.60**
ARC-Challenge	29.35	**29.69**
WinoGrande	58.41	**59.43**
BoolQ	61.80	**64.77**
SciQ	**83.40**	79.40
OpenBookQA	34.00	**35.60**
MMLU	23.42	**25.53**
Average	52.41	**53.67**

C MULTISCREEN ARCHITECTURE AND SCALING

We scale Multiscreen using a single supraparameter Ψ that jointly determines the number of layers, the number of heads, and the embedding dimension. In our experiments, the number of layers and the number of heads are set to $N_{\mathrm{L}} = N_{\mathrm{H}} = \Psi$, and the embedding dimension is set to $d_{\mathrm{E}} = \Psi^2$.

Other architectural hyperparameters are kept fixed across model scales, as summarized in table 2. The resulting model configurations and parameter counts across scales are shown in table 3.

Table 2: Architectural hyperparameters of Multiscreen used in our experiments.

Hyperparameter	**Symbol**	**Value / scaling**
Number of layers	N_{L}	Ψ
Number of heads	N_{H}	Ψ
Embedding dimension	d_{E}	Ψ^2
Query/key dimension	d_{K}	16
Value dimension	d_{V}	64
MiPE threshold	w_{th}	256

Table 3: Architecture configurations and parameter counts for the Multiscreen models used in the scaling experiments.

Quantity	**4M**	**28M**	**286M**	**1.3B**	**4B**
Supraparameter (Ψ)	8	16	32	48	64
Total params	4,134,146	27,546,626	286,347,266	1,304,884,226	3,963,961,346
Non-embedding params	917,698	14,680,834	234,884,098	1,189,092,098	3,758,108,674

D TRANSFORMER BASELINE CONFIGURATIONS

We describe the architecture configurations of the Transformer baseline models used in our experiments. The Transformer baseline adopts a LLaMA-style architecture (Touvron et al., 2023), with architecture hyperparameters largely aligned with those used in Pythia (Biderman et al., 2023). We apply weight tying between the input embedding and the language-modeling head, which differs from the original Pythia configuration, to improve parameter efficiency and strengthen the baseline.

The head dimension is defined as $d_{\mathrm{E}}/N_{\mathrm{H}}$, and the feed-forward dimension is set to $\lfloor \frac{8}{3} d_{\mathrm{E}} \rfloor$. The architecture configurations across model scales are summarized in table 4.

Initialization. For the Transformer baseline, we follow the Pythia/GPT-NeoX initialization scheme (Biderman et al., 2023; Black et al., 2022). Specifically, all weight matrices except residual output projections are initialized using small initialization with standard deviation $\sqrt{2/(5d_{\mathrm{E}})}$, while residual output projections (i.e., the output projections of the attention and feed-forward blocks) are initialized using Wang initialization with standard deviation $2/(N_{\mathrm{L}}\sqrt{d_{\mathrm{E}}})$.

Table 4: Architecture configurations and parameter counts for the Transformer baseline models used in the scaling experiments.

Quantity	**8M**	**45M**	**353M**	**1.3B**
Number of layers (N_{L})	6	6	24	24
Number of heads (N_{H})	4	8	16	16
Embedding dimension (d_{E})	128	512	1,024	2,048
Total params	7,613,312	44,609,024	353,453,056	1,310,935,040
Non-embedding params	1,180,416	18,877,440	301,989,888	1,208,008,704

E MULTISCREEN INITIALIZATION

Except for the gate projection W_{G}, we initialize all projection matrices and the embedding matrix from zero-mean Gaussian distributions whose standard deviations are scaled inversely with the square root of their output dimension (i.e., the second dimension of each weight matrix), as summarized in table 5. The gate projection W_{G} is initialized with a fixed scale independent of dimension. For matrices whose outputs are subsequently unit-normalized, the global initialization scale does not affect the norm of the normalized vectors, but can still affect early optimization dynamics; for projections not followed by such normalization, it directly affects the initial activation scale.

The window parameter s_{w} is initialized independently in each layer with values linearly spaced across heads from 0 to $\log w_{\mathrm{th}}$. The parameter s_{r} is initialized to zero, corresponding to an initial acceptance width of $r = 0.5$. The residual output scale s_{O} is initialized to approximately normalize the total contribution across all tiles in the model, ensuring that the aggregated residual updates remain well-scaled as the number of tiles increases.

The parameters s_{E} and s_{F} control the scale of input embeddings and output logits, respectively, and are initialized to set appropriate scales for these quantities.

Table 5: Parameter shapes and initialization of Multiscreen. All matrix initializations are zero-mean Gaussian, and σ denotes their standard deviation.

Parameter	**Shape**	**Initialization**
W_{Q}	$(d_{\mathrm{E}}, d_{\mathrm{K}})$	$\sigma = 0.1/\sqrt{d_{\mathrm{K}}}$
W_{K}	$(d_{\mathrm{E}}, d_{\mathrm{K}})$	$\sigma = 0.1/\sqrt{d_{\mathrm{K}}}$
W_{V}	$(d_{\mathrm{E}}, d_{\mathrm{V}})$	$\sigma = 0.1/\sqrt{d_{\mathrm{V}}}$
W_{G}	$(d_{\mathrm{E}}, d_{\mathrm{V}})$	$\sigma = 0.1$
W_{O}	$(d_{\mathrm{V}}, d_{\mathrm{E}})$	$\sigma = 0.1/\sqrt{d_{\mathrm{E}}}$
W_{E}	$(\lvert\mathcal{V}\rvert, d_{\mathrm{E}})$	$\sigma = 0.1/\sqrt{d_{\mathrm{E}}}$
s_{w}	per-tile scalar	evenly spaced over $[0, \log w_{\mathrm{th}}]$ per layer
s_{r}	per-tile scalar	0
s_{O}	per-tile scalar	$\log\bigl(1/\sqrt{N_{\mathrm{H}} N_{\mathrm{L}}}\bigr)$
s_{E}	global scalar	0
s_{F}	global scalar	$\log \sqrt{d_{\mathrm{E}}}$

F Additional Scaling Analysis

We further analyze scaling behavior using alternative definitions of model size, as shown in fig. 9.

Non-embedding parameters. In the left panel of fig. 9, we plot validation loss as a function of non-embedding parameters for both Transformer and Multiscreen. The points follow a visually more linear trend on the log–log axes under this parameterization, while the relative advantage of Multiscreen over Transformer is preserved.

Supraparameter Ψ. In the right panel of fig. 9, we analyze scaling behavior using a unified supraparameter Ψ defined for Multiscreen. Plotting against Ψ yields a similarly regular empirical trend, suggesting that Ψ provides a compact and unified characterization of model scaling for Multiscreen.

The empirical trend remains regular through 1.3B, while the 4B point shows a modest deviation in both parameterizations. Because every model is trained on the same fixed token budget, the number of training tokens per parameter decreases with model size. To assess whether incomplete convergence contributes to the 4B deviation, we examine validation loss throughout training in fig. 10. All model sizes continue to improve through 2^{38} training tokens, and the validation-loss reduction over the second half of training increases monotonically with model size. Together, these observations suggest that the 4B deviation partly reflects incomplete convergence under the fixed token budget.

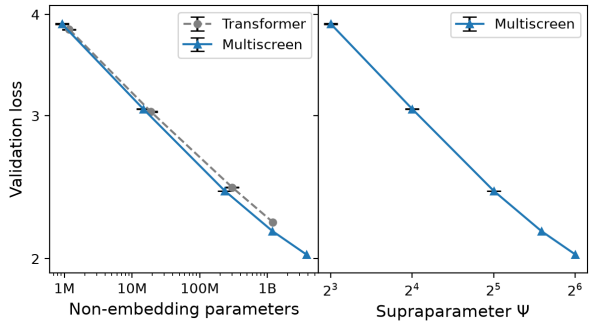


Figure 9: Scaling behavior under alternative definitions of model size. Left: scaling behavior of Transformer and Multiscreen with respect to non-embedding parameters. Right: scaling behavior of Multiscreen with respect to the supraparameter Ψ. Error bars indicate one standard deviation across independently trained models where multiple runs are available.

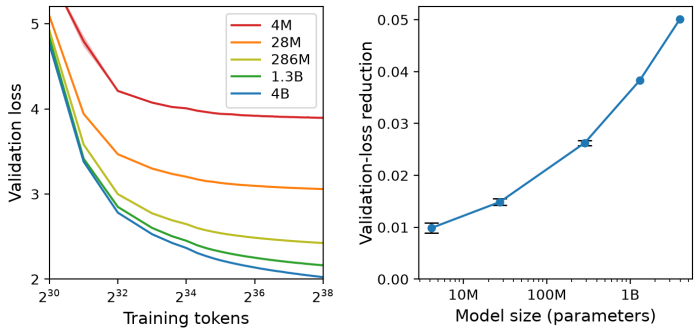


Figure 10: Validation-loss trajectories of Multiscreen models trained under the fixed 2^{38}-token budget. Left: validation loss as a function of training tokens. Right: validation-loss reduction over the second half of training. For the 4M, 28M, and 286M models, curves and shaded regions in the left panel show the mean and one standard deviation across three independently trained models, and error bars in the right panel indicate one standard deviation across the same runs. The 1.3B and 4B results each use a single trained model.

G Training Loss Dynamics

To complement the learning-rate sweep in fig. 7, we visualize training-loss trajectories from the same runs at representative learning rates.

As shown in fig. 11, at higher learning rates, both Transformer configurations either exhibit pronounced loss fluctuations or lie outside the plotted loss range, whereas Multiscreen follows smooth training trajectories and completes training even at learning rate 1.

While fig. 7 summarizes final validation loss and divergence outcomes, the training trajectories shown here illustrate how instability emerges during optimization. To better understand these behaviors, we next analyze the gradient dynamics in Appendix H.

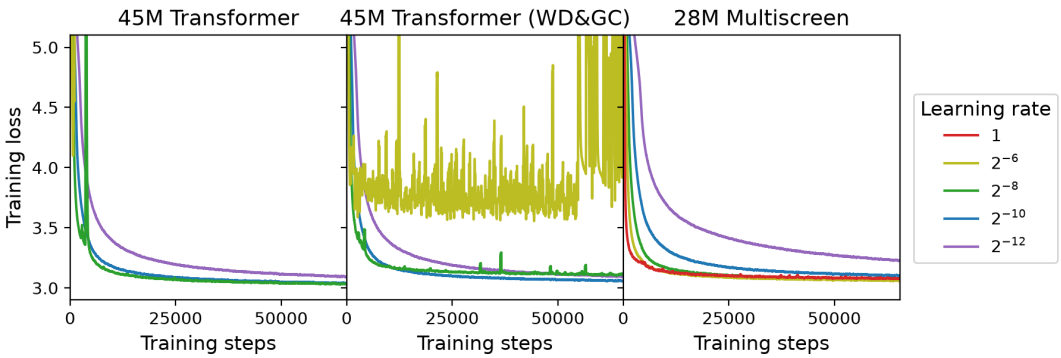


Figure 11: Training loss trajectories from the same runs as in fig. 7, shown for representative learning rates. Left: 45M Transformer without weight decay or gradient clipping; center: the Transformer control with weight decay 0.1 and gradient clipping at 1.0; right: 28M Multiscreen without either. Curves are smoothed using a moving average over 256 training steps. Divergent Transformer runs may lie entirely outside the displayed loss range.

H GRADIENT NORM DYNAMICS

We compare global gradient norms from one 1.3B Transformer run and one 1.3B Multiscreen run used in the scaling experiments (section 4.6). The Transformer and Multiscreen models use learning rates 2×10^{-4} and 2^{-4}, respectively, with Multiscreen remaining stable at the substantially larger value.

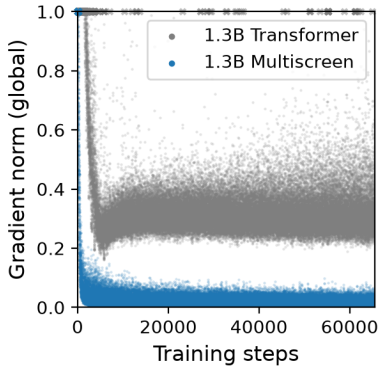


Figure 12: Global gradient-norm dynamics during training for the 1.3B Transformer and Multiscreen models used in the scaling experiments. For visualization, values above 1 are clipped and shown with $\times$ markers.

As shown in fig. 12, Transformer maintains a nonzero lower floor relative to the spread of its gradient norms, whereas Multiscreen has a dense concentration near zero relative to its full gradient-norm range. This contrast motivates testing whether a similar near-zero lower tail appears when controlled Transformer variants are stabilized at learning rate 1.

I Components of Large-Learning-Rate Stability

We conduct controlled diagnostic experiments at learning rate 1 to determine whether the stability of Multiscreen is caused by screening itself or by its normalization and scaling components. Every diagnostic run is trained for 2^{36} tokens (2^{14} steps). For the Transformer variants, "QKNorm" denotes an RMSNorm-based variant of query–key normalization (Henry et al., 2020) with a learned elementwise scale applied to queries and keys. Here, "PLNorm" denotes pre-logit normalization, while "PreNorm" denotes the RMSNorm applied at the beginning of each Transformer layer. Embedding normalization and PLNorm introduce no learned elementwise scales and only project their inputs onto the unit hypersphere. Together with the learned input and logit scales e^{s_E} and e^{s_F}, they form the SIO parameterization defined in section 3. The learned scalars s_E, s_F, and s_O parameterize the input-embedding, output-logit, and residual-output scales, respectively.

Table 6: Training outcomes at learning rate 1 under controlled modifications to Transformer and Multiscreen. "Stable" indicates completion of the 2^{14}-step diagnostic run without divergence. "SIO − PLNorm" denotes the SIO components other than PLNorm.

Configuration	Outcome
Transformer	Diverged
Transformer + SIO	Diverged
Transformer + PLNorm	Diverged
Transformer + SIO − PLNorm	Diverged
Transformer + QKNorm	Diverged
Transformer + QKNorm + SIO	Stable
Transformer + QKNorm + PLNorm	Stable
Transformer + QKNorm + SIO − PLNorm	Stable
Transformer − PreNorm + QKNorm + SIO	Diverged
Transformer − PreNorm + QKNorm + PLNorm	Diverged
Transformer − PreNorm + QKNorm + SIO − PLNorm	Diverged
Multiscreen	Stable
Multiscreen − SIO	Diverged
Multiscreen − PLNorm	Stable
Multiscreen − (SIO − PLNorm)	Diverged
Multiscreen − s_O	Diverged
Multiscreen + PreNorm	Diverged

The clearest shared pattern in table 6 is the interaction between Q/K normalization and SIO. For Transformer, QKNorm and SIO each diverge when introduced alone, whereas their combination is stable at learning rate 1. Multiscreen likewise incorporates both Q/K unit-length normalization and SIO by design; the standard model is stable at learning rate 1, whereas removing SIO causes divergence. Together, these results suggest that pairing a form of Q/K normalization with SIO contributes strongly to large-learning-rate stability across both architectures. Q/K unit-length normalization is not independently ablated in Multiscreen because it establishes the fixed $[-1, 1]$ input scale assumed by Trim and is therefore inseparable from the screening-unit design.

The remaining controls refine this shared pattern. For Transformer, QKNorm also stabilizes both PLNorm and SIO without PLNorm when PreNorm is retained. Notably, PreNorm plays opposite roles across the two architectures: removing it from any of the three stable Transformer configurations causes divergence, whereas adding it to Multiscreen also causes divergence. For Multiscreen, removing PLNorm alone preserves stability, whereas retaining only PLNorm from SIO causes divergence. Removing s_O also causes Multiscreen to diverge. Thus, the precise supporting components and normalization placement remain architecture dependent.

The five selected stable configurations exhibit broadly similar validation-loss trajectories in this diagnostic run (fig. 13). Although their absolute scales and dispersions differ substantially, all five exhibit a dense lower tail near zero relative to their overall gradient-norm ranges, whereas the standard Transformer at learning rate 10^{-3} maintains a visibly nonzero lower floor (fig. 14). This controlled

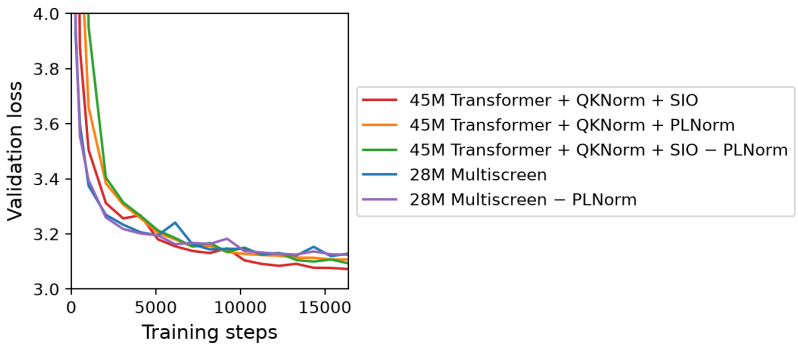


Figure 13: Validation-loss trajectories for five selected configurations that remain stable at learning rate 1. PLNorm denotes pre-logit normalization, and each curve shows one 2^{14}-step training run.

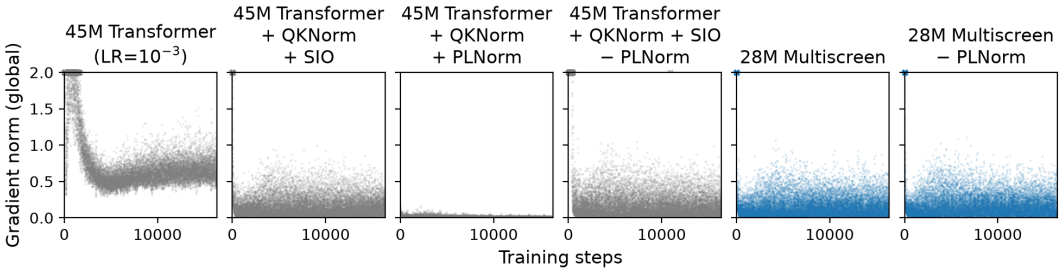


Figure 14: Global gradient norms for five selected configurations that remain stable at learning rate 1, together with the standard 45M Transformer at learning rate 10^{-3} as a low-learning-rate reference. PLNorm denotes pre-logit normalization, and each panel shows one 2^{14}-step training run. For visualization, values above 2 are clipped and shown with $\times$ markers.

comparison reproduces the contrast in fig. 12 and reveals a common gradient-norm signature of large-learning-rate stability across architectures.

Taken together, these experiments identify Q/K normalization paired with SIO as a transferable design pattern for large-learning-rate stability, while showing that its supporting component interactions depend on the backbone.

J VISUALIZATION OF DISTANCE-AWARE RELEVANCE

We visualize distance-aware relevance maps α_{ij}^{d} for all tiles in a 4M Multiscreen model used in section 4.6, as shown in fig. 15. These visualizations highlight the presence of learned screening windows and the extent of sparsity within them.

The input sequence is a 228-token natural-language passage (the abstract of this paper). All layers and heads are shown, arranged in a grid layout where rows correspond to layers and columns correspond to heads.

Within each map, dark gray regions indicate causally inaccessible positions or positions outside the learned screening window, while color intensity represents the magnitude of α_{ij}^{d} over the remaining positions. Each tile is annotated with its layer and head indices, the learned screening window w, the acceptance width r, and the fraction of nonzero relevance values among causally accessible positions within the window, $\Pr(\alpha_{ij}^{\mathrm{d}} > 0)$.

Different tiles cover different context ranges, with some focusing on local neighborhoods and others covering broader portions of the sequence. Many maps contain a substantial number of zero entries, with the degree of sparsity varying across tiles.

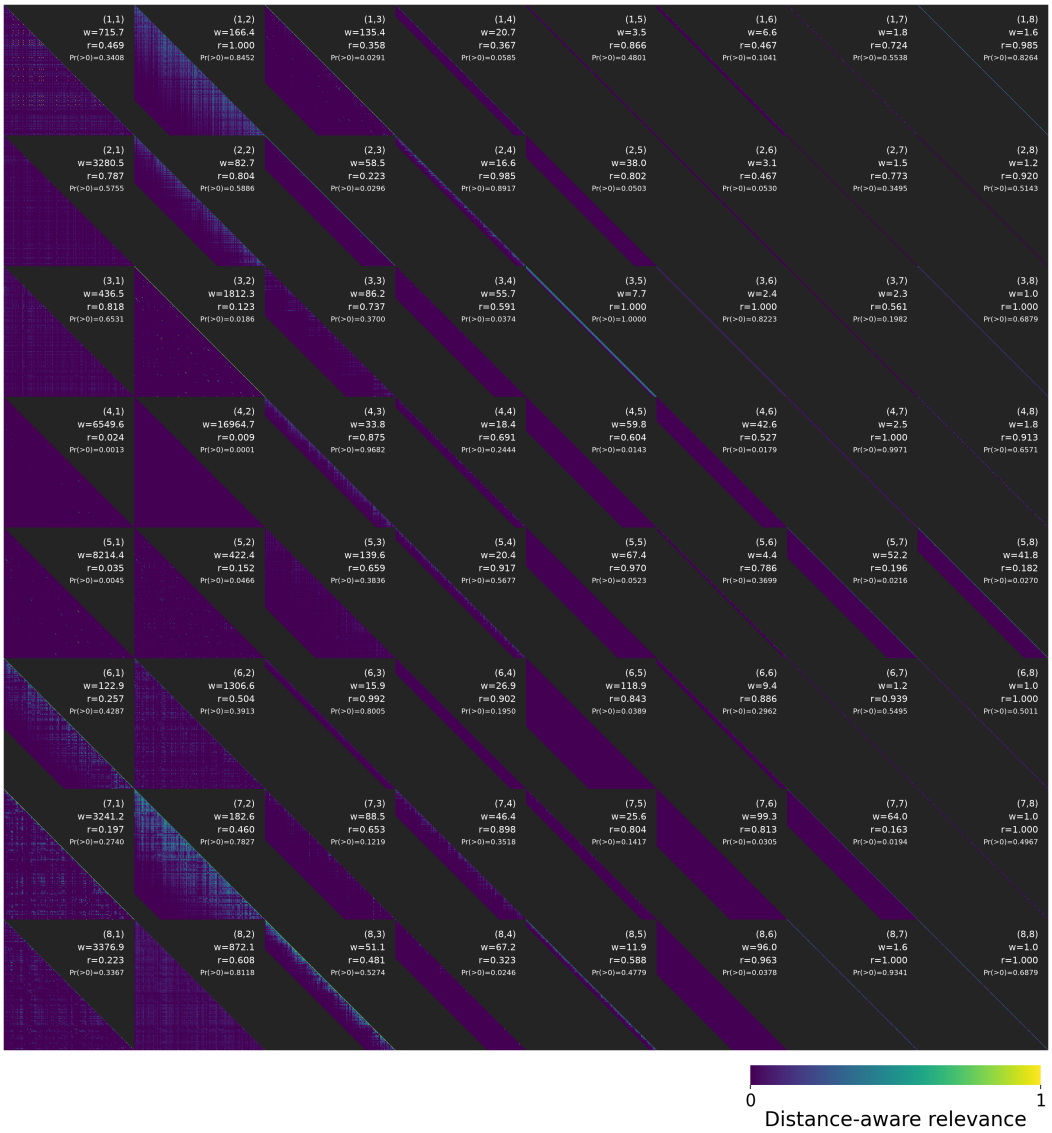


Figure 15: Distance-aware relevance maps across layers and heads. Each map shows the distance-aware relevance α^{d}_{ij}, with rows indexing query positions and columns indexing key positions. Dark gray regions indicate causally inaccessible positions or positions outside the learned screening window. Each tile is annotated with its layer and head indices, the learned screening window w, the acceptance width r, and the fraction of nonzero relevance values $\Pr(\alpha^{\mathrm{d}}_{ij} > 0)$ among causally accessible positions within the window, providing a summary of its sparsity and coverage.

K Distribution of Learned Screening Parameters

Each screening unit has independently learned scalar parameters s_{w} and s_{r}, which determine the screening window w and acceptance width r. To better understand how these parameters are used after training, we visualize the resulting values of w and r across all layers and heads of the 4B Multiscreen model. The figure reports the raw finite values of w before applying the default inference rule, which replaces values exceeding the pretraining context length with $w = \infty$.

As shown in fig. 16, learned screening windows span several orders of magnitude. In this model, most units remain local or medium-range, while 6.0% learn window values exceeding the pretraining context length and are therefore treated as having $w = \infty$ under the default inference rule. The learned acceptance widths are often close to one, but also vary substantially across units, indicating that different screening units specialize to different degrees of selectivity. This diversity suggests that Multiscreen combines many local screening units with a smaller number of broad-context units.

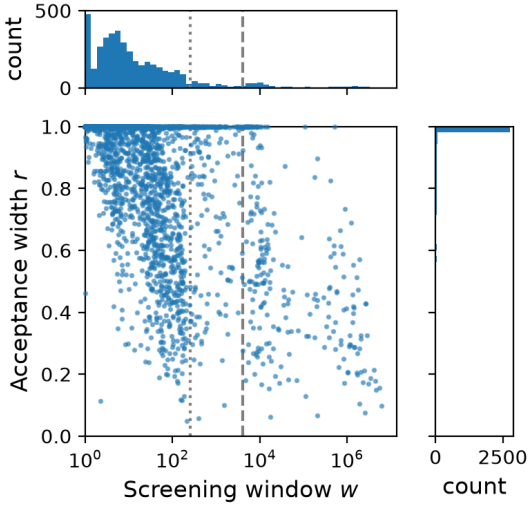


Figure 16: Distribution of learned screening parameters in the 4B Multiscreen model. Each point corresponds to one screening unit, with horizontal position indicating the learned screening window w and vertical position indicating the learned acceptance width r. The horizontal axis is shown on a logarithmic scale. Marginal histograms show the corresponding distributions of w and r. The dotted vertical line marks the MiPE threshold ($w_{\mathrm{th}} = 2^8$), at or above which MiPE becomes inactive, and the dashed vertical line marks the pretraining context length (2^{12}), beyond which the default inference rule sets $w = \infty$. Raw learned windows span several orders of magnitude, with many local or medium-range units and a smaller number exceeding the pretraining context length.

L LONG-CONTEXT PERPLEXITY: EVALUATION DETAILS

We provide additional details for the long-context perplexity evaluation in section 4.2, along with robustness results using smaller models trained with multiple random seeds.

Evaluation setup. We evaluate long-context language modeling using position-dependent perplexity over long sequences. At each reported position, perplexity is averaged over a trailing window spanning the preceding 10% of the context length and ending at that position to reduce local variance.

We use the PG-19 dataset (Rae et al., 2020), consisting of books published before 1919 from Project Gutenberg.[1] We pool the training, validation, and test splits and use all 5,747 documents whose tokenized length exceeds 2^{17}. The test split alone contains too few documents of this length to provide stable position-dependent estimates at the longest contexts. PG-19 was not explicitly deduplicated against SlimPajama, so overlap with the pretraining corpus may remain. We therefore use this evaluation as a controlled comparison between architectures rather than as a held-out estimate of absolute generalization performance. Every compared model encountered exactly the same pretraining tokens in the same order and is evaluated on the same segments, preventing differential data exposure between Transformer and Multiscreen. For each document, we extract a contiguous segment of $2^{17} + 1$ tokens centered at the midpoint of the document. This construction ensures that predictions are evaluated in the middle of long contexts rather than near document boundaries.

YaRN scaling. For Transformer, we apply YaRN (Peng et al., 2024) with multiple scaling factors to assess extrapolation beyond the training context length. For base models trained with sequence length 2^{12}, we evaluate scaling factors $\times 1$, $\times 2$, $\times 4$, $\times 8$, $\times 16$, $\times 32$, and $\times 64$. Multiscreen has no RoPE component, so YaRN scaling does not apply.

Seed-averaged smaller-model results. In addition to the 1.3B results in fig. 4, we evaluate the 353M Transformer and 286M Multiscreen models used in the scaling experiments. For each architecture, we use three independently trained models with different random seeds and report the mean and one standard deviation.

As shown in fig. 17, the same qualitative behavior as in section 4.2 is observed. Multiscreen maintains stable perplexity substantially beyond the training context, whereas Transformer with smaller YaRN scaling factors exhibits sharp degradation beyond the corresponding scaled range. Increasing the YaRN scaling factor delays this degradation but increases perplexity at shorter contexts, including within the training range.

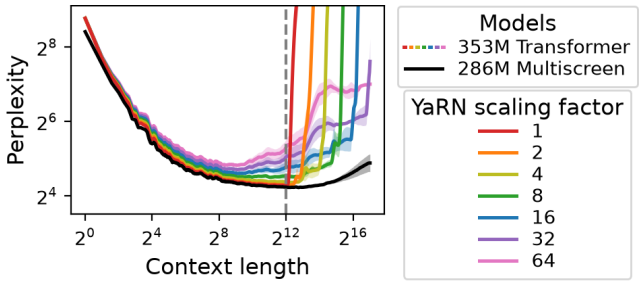


Figure 17: Long-context perplexity comparison for 353M Transformer and 286M Multiscreen models. Curves show mean perplexity over three independently trained models, and shaded regions indicate one standard deviation. The horizontal axis indicates the context length at which position-dependent perplexity is evaluated, and the vertical axis is perplexity. The dashed vertical line marks the pretraining context length (2^{12}). Colored curves show Transformer with different YaRN scaling factors, while the black curve shows Multiscreen, which has no RoPE component and therefore requires no YaRN scaling. Multiscreen maintains stable perplexity substantially beyond the training context, while Transformer exhibits a tradeoff between delayed breakdown and increased perplexity at shorter contexts.

[1] https://huggingface.co/datasets/Geralt-Targaryen/pg19

M ABCDIGITS: TASK CONSTRUCTION AND EVALUATION DETAILS

We provide additional construction and evaluation details for ABCDigits, the synthetic key–value retrieval benchmark evaluated in fig. 5b and discussed in section 4.3.

Motivation. ABCDigits builds on synthetic associative recall and key–value retrieval benchmarks studied in prior work (Graves et al., 2014; Olsson et al., 2022; Liu et al., 2024; Arora et al., 2024). We cast retrieval as a structured completion problem that directly requires recovering the associated value. This provides a direct probe of retrieval behavior while reducing confounding effects from natural-language semantics, instruction following, and prompt-specific factors.

Task construction. Each ABCDigits instance presents a shuffled list of equations mapping uppercase letters to six-digit integers, such as `A=967892`. A query is formed by appending a target letter followed by an equals sign, such as `L=`, and the model must complete the corresponding integer. The target mapping appears exactly once in the context, requiring the model to locate this unique occurrence. Since each letter is consistently mapped to a single integer across the context, the task implicitly enforces a one-to-one key–value correspondence without requiring explicit instructions.

A key design choice is that the number of distinct keys remains fixed at 26 uppercase letters regardless of context length. This removes the bias that retrieval becomes harder simply because the number of keys increases, allowing us to better isolate length-dependent retrieval behavior.

To ensure that extremely low-frequency equations are not anomalous within the context, we construct the non-target portion of the context in two stages. First, we include exactly one instance of each of the 25 non-target equations. We then fill the remaining context by sampling additional non-target equations from a highly skewed categorical distribution with weights proportional to $2^0, 2^1, \ldots, 2^{24}$, where the assignment from letters to weights is randomized for each instance. This avoids making infrequent key–value pairs appear anomalous, ensuring that a range of frequencies is naturally represented within the context. All non-target equations are then shuffled. Finally, the unique target equation, such as `L=169428`, is inserted at a specified depth, and the query suffix `L=` is appended. A concrete example of the resulting prompt is shown in fig. 5a in the main text.

Evaluation protocol. Following the visualization protocol commonly used in needle-in-a-haystack evaluations (Kamradt, 2023; Arize AI, 2023), we measure retrieval accuracy across a grid of context lengths and target depths. Target context lengths range from 2^{12} to 2^{17}, and depths are set to 0.1, 0.3, 0.5, 0.7, and 0.9. Because each equation is retained in full, the realized token count may differ slightly from its target. Depth is implemented at the equation level: a depth of d means that the target equation is inserted after approximately a fraction d of the equations in the context, before the final query.

For each context length and depth, we generate 1,000 independent ABCDigits instances. Every model is evaluated on the same instances at each setting. We evaluate exact-match accuracy under greedy decoding with temperature 0. An instance is counted as correct only if the generated integer exactly matches the target value.

Models and inference settings. The main ABCDigits evaluation compares Transformer models with 8M, 45M, 353M, and 1.3B parameters against Multiscreen models with 4M, 28M, 286M, and 1.3B parameters. For all models below 1.3B parameters, results are averaged over three independently trained models, with 1,000 instances evaluated per model and setting. For the 1.3B models, we evaluate a single trained model for each architecture due to computational constraints.

For Transformer, we apply YaRN (Peng et al., 2024) with a scaling factor matched to each evaluation context length. Specifically, for evaluation context length L and training context length $L_{\text{train}} = 2^{12}$, we set the YaRN scaling factor to L/L_{train}; for example, evaluation at $L = 2^{15}$ uses a scaling factor of 8. Multiscreen has no RoPE component, so YaRN scaling does not apply. Multiscreen's default inference rule is fixed across context lengths: after training, any learned screening window exceeding the maximum sequence length seen during training is converted once to $w = \infty$, and all windows then remain unchanged.

N PASSKEY RETRIEVAL RESULTS

We evaluate retrieval performance on the passkey retrieval task introduced by Mohtashami & Jaggi (2023), as a complementary test to ABCDigits.

Setup. The experimental setup follows that of ABCDigits (Appendix M). We use the same range of context lengths (2^{12} to 2^{17}), target depths (0.1, 0.3, 0.5, 0.7, and 0.9), and number of trials (1,000 instances per setting). As in the original task, each passkey is an integer sampled from 1 to 50,000; unlike its random-insertion protocol, we place the passkey at each fixed target depth in our evaluation grid. We count an output as correct only when it produces the passkey immediately at the answer position, rather than searching the first 100 generated tokens for the first integer. We use the same model scales and architecture-specific inference settings; only the prompt format is changed to follow the passkey retrieval task.

Results. The results are shown in fig. 18. We observe the same qualitative trend as in ABCDigits: Multiscreen is substantially more robust to increasing context length than Transformer, and its retrieval accuracy improves strongly with model size. Under our default inference rule, the 286M and 1.3B Multiscreen models achieve 100% retrieval accuracy at every evaluated context length and target depth.

In contrast, Transformer performance degrades rapidly as context length increases. Even with context-matched YaRN scaling, Transformer performance deteriorates sharply and can collapse at longer contexts. As in ABCDigits, Multiscreen outperforms the corresponding Transformer baseline at every scale even at the training context length, despite using fewer parameters. These results show that the improved retrieval ability of Multiscreen is not specific to the ABCDigits construction.

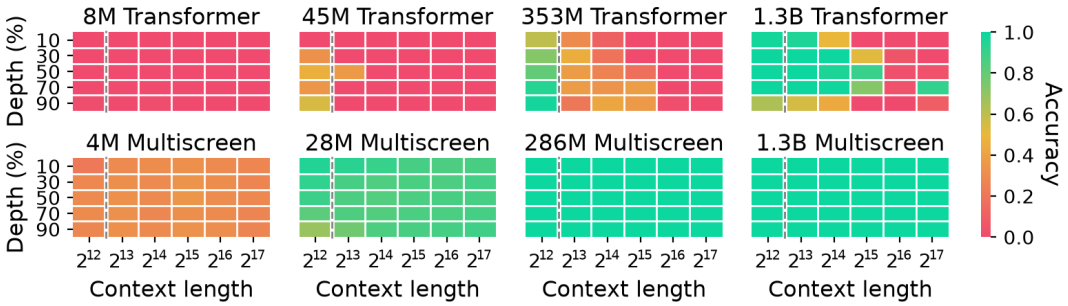


Figure 18: Passkey retrieval accuracy across model sizes for Transformer and Multiscreen. The setup follows that of ABCDigits, with identical context lengths, target depths, and evaluation protocol. Top row: Transformer baselines. Bottom row: Multiscreen models using our default inference rule. For Transformer, the YaRN scaling factor is set to the evaluation context length divided by the training context length. Multiscreen is more robust to increasing context length than Transformer, with the 286M and 1.3B Multiscreen models achieving 100% accuracy across the grid.

O CONTROLLED ATTENTION-MECHANISM COMPARISON

We compare screening with a broad set of attention mechanisms using the architecture and training configuration of the 353M Transformer, subject to the common document-masking exception described below. Each model is trained for 2^{38} tokens using the same learning rate and remaining training hyperparameters as the 353M Transformer. Each variant is trained once with a single random seed. In the main comparison, only the attention mechanism is replaced; model dimensions, including the query, key, and value dimensions, and initialization otherwise match the 353M Transformer. Throughout this section, L_{eval} denotes the evaluation context length, while $L_{\text{train}} = 2^{12}$ denotes the pretraining context length.

Document masking. We disable document masking for every model in this comparison because the publicly available fused kernels used for several variants, including Entmax, Sigmoid, and Forgetting Attention, do not support document-level masking. This common setting ensures that every variant is trained with the same visibility across documents within a packed sequence. The standard 353M Transformer trained with document masking is reported separately in figs. 5b and 17 and exhibits the same qualitative length-extrapolation failure as the Softmax control in this comparison.

Fused implementations. We use the publicly available fused implementations for Forgetting Attention,[2] Sigmoid attention through TritonSigmoid version 0.1.2 (Sadashivaiah et al., 2026),[3] and Entmax ($\alpha = 1.5$) through AdaSplash version 0.2.2 (Gonçalves et al., 2026).[4] For Forgetting Attention, we enable adaptive computation pruning with `adaptive_threshold="auto"` (Lin et al., 2025b).

Compared mechanisms. The full Screening variant consists only of the screening unit: unit-length normalization of queries, keys, and values, MiPE, Trim, Softmask, and TanhNorm. It does not include the output gate or any other modification from the gated screening tile, and all components outside attention remain identical to those of the Transformer backbone. The Screening (no window) variant removes the learned screening window w and its dependent components, omitting MiPE and Softmask while retaining unit-length normalization of queries, keys, and values, Trim, and TanhNorm, and uses ordinary RoPE. The Screening (RoPE) variant retains the learned window and Softmask while replacing MiPE with ordinary RoPE. The exact configurations are summarized in table 7.

The parameter counts differ by at most 0.12% across the comparison, keeping model capacity closely matched. Softmax_1 adds one to the Softmax denominator, giving $\text{softmax}_1(z)_i = \exp(z_i)/(1 + \sum_j \exp(z_j))$ and allowing the total attention weight to approach zero (Miller, 2023). In the table, YaRN denotes the full method, which combines NTK-by-parts RoPE-frequency interpolation with pre-softmax attention-logit scaling, whereas YaRN* retains the RoPE-frequency interpolation but omits the attention-logit scaling (Peng et al., 2024). Throughout this comparison, s denotes the common scaling factor used to parameterize both YaRN positional scaling and method-specific length corrections. The method-specific corrections use L_{train} as their base length, so context-matched scaling with $s = L_{\text{eval}}/L_{\text{train}}$ corresponds to the evaluation length L_{eval}. A dash indicates that no inference-time scaling is applied. The "method-specific" designation indicates that a variant additionally modifies its own length-dependent term.

For Sigmoid attention, the base logit bias is $b = -\log L_{\text{train}}$ (Ramapuram et al., 2025). At scaling factor s, we replace it with $b_s = -\log(sL_{\text{train}}) = b - \log s$, shifting every attention logit by an additional $-\log s$. Following Hua et al. (2022), the unscaled ReLU^2 variants compute attention weights as $A = \text{ReLU}^2(QK^\top/L_{\text{train}} + B_{\text{rel}})$, where B_{rel} denotes the relative-position bias used in both variants. For each layer and head, we construct this bias from learned query-side and key-side vectors $\boldsymbol{q}_{\text{rel}}, \boldsymbol{k}_{\text{rel}} \in \mathbb{R}^{128}$, matching the original formulation. Its entries are

$$(B_{\text{rel}})_{ij} = \text{RoPE}(\boldsymbol{q}_{\text{rel}}, i)\,\text{RoPE}(\boldsymbol{k}_{\text{rel}}, j)^\top, \tag{18}$$

where the two vectors are rotated at the query and key positions, respectively. These vectors account for the additional parameters of the ReLU^2 variants in table 7. For scaling factor s, we instead

[2] https://github.com/zhixuan-lin/forgetting-transformer

[3] https://github.com/MSDLLCpapers/triton-sigmoid

[4] https://github.com/deep-spin/adasplash

Table 7: Configurations in the controlled attention-mechanism comparison. All listed variants use the same query, key, and value dimensions. The Context column describes formal support, so mechanisms with distance-dependent attenuation but no hard support boundary are classified as Full. Parameter counts are reported in millions and rounded to two decimal places.

Variant	Pos. enc.	Context	Scaling	# Params (M)
Softmax	RoPE	Full	YaRN	353.45
Softmax_1	RoPE	Full	YaRN	353.45
SSMax	RoPE	Full	YaRN*	353.45
Entmax ($\alpha = 1.5$)	RoPE	Full	YaRN*	353.45
Sigmoid	RoPE	Full	YaRN* + method-specific	353.45
ReLU^2	RoPE	Full	YaRN* + method-specific	353.55
Sliding window	RoPE	2048	–	353.45
ALiBi	ALiBi	Full	–	353.45
Forgetting	NoPE	Full	–	353.85
Screening (no window)	RoPE	Full	YaRN*	353.45
Screening (RoPE)	RoPE	Learned w	YaRN*	353.45
Screening	MiPE	Learned w	–	353.45

use $A_s = \text{ReLU}^2((QK^\top/L_{\text{train}} + B_{\text{rel}})/s)$, scaling the complete pre-activation, including the relative-position bias, by $1/s$. The original GAU formulation normalizes by the sequence length but does not specify an additional adjustment for evaluation beyond the training context; we use the above rule as its method-specific extrapolation treatment.

For Forgetting Attention, ALiBi, the sliding-window baseline, and full Screening, we apply no inference-time scaling. Following its original specification, Forgetting Attention uses no positional encoding (Lin et al., 2025a). We classify both ALiBi and Forgetting Attention as full-context at the level of their underlying relevance rules because their distance-dependent attenuation does not impose a hard support boundary. Forgetting Attention's adaptive computation pruning can omit sufficiently small contributions computationally; we treat these implementation-level zeros as an efficiency approximation rather than a formal property of the underlying relevance rule.

Additional GAU-style parameterization. We also evaluate ReLU^2 with the original GAU-style query/key parameterization, in which learned elementwise scales and shifts are applied to a shared projected representation (Hua et al., 2022). Doubling its query/key dimension approximately matches the parameters of the separate W_Q and W_K projections, yielding a 353.75M-parameter variant. Because this variant underperforms the separate-projection ReLU^2 variant on both evaluations, we report the stronger and more tightly controlled separate-projection result in fig. 6.

Additional sliding-window sizes. We also evaluate sliding-window attention with fixed window sizes of 512 and 1024 tokens. Because these narrower windows further restrict accessible context without changing the qualitative conclusion, the main comparison reports the 2048-token variant as the most favorable sliding-window control for long-range retrieval.

Properties of the relevance rules. Table 8 summarizes the rule-level properties and the cross-length scaling applied to each evaluated mechanism. "Variable total mass" means that the relevance values are not constrained to sum to one, and "Pairwise independent" means that the relevance assigned to one query–key pair does not depend on the other available keys. "Exact rejection" requires that an accessible key can receive exactly zero relevance based on its content, excluding zeros imposed only by causal or fixed-window masks, while "Empty selection" requires that all accessible keys can simultaneously receive exactly zero relevance.

"No explicit length term" means that the relevance rule contains no explicit sequence-length term, whereas "No inference scaling" means that the evaluated mechanism is used across context lengths without a context-length-dependent adjustment at inference. Although inference-time scaling can preserve boundedness and pairwise independence at any fixed factor, it changes the calibration of

Table 8: Properties relevant to absolute relevance across context lengths for the mechanisms evaluated in the controlled comparison. "Bounded" indicates that each relevance value has a finite upper bound. Structural zeros introduced by causal or fixed-window masks are excluded from "Exact rejection."

Variant	Variable total mass	Pairwise independent	Bounded	Exact rejection	Empty selection	No explicit length term	No inference scaling
Softmax	–	–	✓	–	–	✓	–
Softmax$_1$	✓	–	✓	–	–	✓	–
SSMax	–	–	✓	–	–	–	–
Entmax ($\alpha = 1.5$)	–	–	✓	✓	–	✓	–
Sigmoid	✓	✓	✓	–	–	–	–
ReLU^2	✓	✓	–	✓	✓	–	–
Sliding window	–	–	✓	–	–	✓	✓
ALiBi	–	–	✓	–	–	✓	✓
Forgetting	–	–	✓	–	–	✓	✓
Screening (no window)	✓	✓	✓	✓	✓	✓	–
Screening (RoPE)	✓	✓	✓	✓	✓	✓	–
Screening	✓	✓	✓	✓	✓	✓	✓

relevance values across context lengths. A check mark indicates that a property holds, while a dash indicates that it does not.

We call a relevance mechanism absolute across context lengths when it assigns each accessible query–key pair a relevance value independently on a bounded, length-invariant scale, permits exact rejection and empty selection, does not constrain the total relevance mass to one, and requires no inference-time scaling. Under this definition, the listed properties are requirements for absolute relevance across context lengths, but satisfying them does not by itself guarantee better empirical performance. For screening, at fixed query and key representations, adding other keys does not mechanically renormalize existing relevance values, and zero has the explicit meaning of rejection, making the resulting relevance signal easier to inspect and interpret. We do not directly evaluate this potential interpretability benefit in the present work.

Softmax$_1$ illustrates why variable total mass and empty selection are distinct: its total mass can approach zero but remains strictly positive for finite logits. Likewise, Sigmoid attention has variable total mass and pairwise-independent weights, but its strictly positive outputs preclude exact rejection and empty selection. Entmax can assign exact zeros to individual keys, but its unit-sum constraint prevents empty selection. ReLU^2 permits exact rejection and empty selection but has unbounded relevance values and an explicit sequence-length term. Screening combines variable total mass, pairwise independence, bounded relevance, exact rejection, and empty selection without an explicit sequence-length term or inference-time scaling.

Evaluation protocol. We evaluate each model on long-context perplexity and ABCDigits retrieval using the protocols described in Appendix L and Appendix M, respectively. The datasets, sample counts, and all other evaluation settings exactly match the main evaluations. For variants that use inference-time scaling in the ABCDigits evaluation, we set the scaling factor to $L_{\mathrm{eval}}/L_{\mathrm{train}}$. For example, evaluation at $L_{\mathrm{eval}} = 2^{15}$ uses a scaling factor of 8. The dashed vertical lines in all panels mark the pretraining context length, $L_{\mathrm{train}} = 2^{12}$.

Screening is the only evaluated mechanism that maintains both low long-context perplexity and robust retrieval across all evaluated context lengths and target depths without inference-time scaling (fig. 6). As shown in the top row, ALiBi, Forgetting Attention, and sliding-window attention maintain

relatively stable perplexity at long contexts. However, the bottom row shows that all three lose retrieval accuracy as context length increases: the distance-dependent attenuation in ALiBi and Forgetting Attention suppresses distant information, while sliding-window attention cannot access targets outside its window. Thus, stable long-context perplexity alone does not establish access to distant information. Softmax and the remaining global-attention alternatives likewise lose retrieval accuracy rapidly beyond the training context despite their respective extrapolation treatments.

Even at scaling factor 1, Screening (no window) avoids the abrupt perplexity degradation beyond the training context seen in the other full-context RoPE-based mechanisms. With context-matched scaling, it also compares favorably with most alternatives on ABCDigits beyond the training context, but its long-context perplexity behavior is the more distinctive result. Because Screening (no window) and the failing global mechanisms all use RoPE, this contrast shows that RoPE alone cannot explain their abrupt degradation beyond the training context.

Screening (RoPE) substantially improves both evaluations over Screening (no window), isolating the benefit of the learned w-dependent Softmask construction while holding RoPE fixed. Replacing RoPE with MiPE while retaining Softmask produces a striking further gain: Full Screening achieves perfect ABCDigits retrieval across every evaluated context length and target depth. Because this replacement is the only difference from Screening (RoPE), the comparison isolates MiPE's additional contribution. Together, these results isolate the long-context behavior to the screening unit itself rather than the output gate or the wider Multiscreen architecture.

P THE ROLE OF TANHNORM

TanhNorm softly bounds the norm of the screening aggregation while preserving its direction, which may be especially important because screening relevance values are not constrained to sum to one.

To isolate this role, we compare the standard 286M Multiscreen configuration with a matched variant trained without TanhNorm while holding all remaining model, training, and evaluation settings fixed. Because TanhNorm has no trainable parameters, removing it leaves the parameter count unchanged. All results in this section use checkpoints at 2^{36} training tokens rather than the 2^{38}-token checkpoints used in the main experiments. The TanhNorm baseline uses the same three independently trained 286M Multiscreen runs as the model-scaling experiments in section 4.6, and we average their results under both inference-window treatments. The variant without TanhNorm is evaluated using a single independently trained run.

We evaluate both models under two inference-window treatments. The default rule converts any learned screening window exceeding the maximum sequence length seen during training to $w = \infty$ once after training and then keeps every window fixed across evaluation lengths. The as-learned condition instead retains every learned window value unchanged. We use the long-context perplexity and ABCDigits protocols described in Appendices L and M, respectively.

As shown in fig. 19, the two conditions with TanhNorm nearly coincide and maintain low perplexity across the evaluated context lengths. Without TanhNorm, both conditions remain stable beyond the pretraining context length rather than failing immediately at its boundary, but their perplexity eventually increases sharply. The default rule fails earlier than the as-learned condition, showing that finite learned windows can delay this instability but do not eliminate it. Thus, TanhNorm substantially extends the range of stable length extrapolation under both window treatments and is particularly important when broad-window units are converted to $w = \infty$.

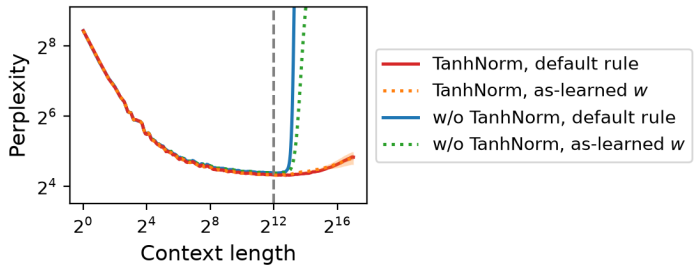


Figure 19: Long-context perplexity for 286M Multiscreen with and without TanhNorm under the default inference rule and with as-learned w. For the TanhNorm baseline, curves show the mean across three independently trained runs, and shaded regions indicate one standard deviation; the variant without TanhNorm uses a single run. The gray dashed line marks the pretraining context length (2^{12}).

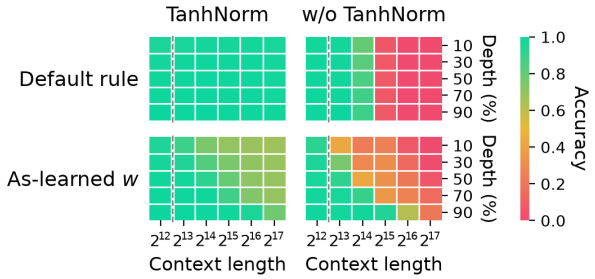


Figure 20: ABCDigits retrieval accuracy for the same 286M Multiscreen variants and inference-window treatments as in fig. 19. Results with TanhNorm are averaged over three independently trained runs, whereas results without TanhNorm use a single run. The default rule applies the post-training conversion to $w = \infty$, whereas the as-learned condition leaves every learned window unchanged. The gray dashed lines mark the pretraining context length (2^{12}).

The retrieval results in fig. 20 reveal distinct failure modes for the two window treatments. With TanhNorm, the default rule maintains near-perfect accuracy across the evaluated context lengths and target depths, whereas retaining as-learned windows introduces a gradual distance-dependent decline at longer contexts. Without TanhNorm, the default rule remains effective beyond the pretraining context length but then collapses across target depths once the context becomes sufficiently long. Retaining as-learned windows avoids this simultaneous collapse, but retrieval instead becomes increasingly local: targets nearer the query remain accessible longer than distant targets. Even without TanhNorm, screening retains substantial length-extrapolation ability in both perplexity and retrieval. TanhNorm extends this robustness, stabilizing long-context aggregation and preserving global retrieval when the default inference rule converts learned windows beyond the training context length to $w = \infty$.

Q THE ROLE OF MIPE

MiPE couples positional rotation to each unit's learned screening window and becomes inactive for sufficiently broad windows. To isolate its role, we compare the standard 286M Multiscreen configuration with variants that replace MiPE with RoPE or NoPE while retaining the learned window and Softmask. These replacements introduce no trainable parameters, so all variants have the same parameter count, and all remaining model, training, and evaluation settings are fixed.

All results in this section use checkpoints at 2^{36} training tokens. The MiPE baseline uses the same three independently trained 286M Multiscreen runs as the model-scaling experiments in section 4.6, whereas each positional-encoding ablation uses a single independently trained run. We evaluate all variants using Multiscreen's default inference rule for w and the long-context perplexity and ABCDigits protocols in Appendices L and M.

As shown in fig. 21, replacing MiPE with RoPE or NoPE changes long-context perplexity only modestly. MiPE and NoPE maintain low perplexity without inference-time positional scaling, while the RoPE variant remains comparable under smaller YaRN scaling factors but worsens under larger factors at the longest contexts. Thus, long-context perplexity alone provides little indication of the large retrieval differences observed below.

Retrieval sharply separates the positional-encoding variants despite their similar perplexity behavior (fig. 22). NoPE retains strong retrieval, showing that the core long-context retrieval ability does not arise from positional encoding alone. Nevertheless, MiPE outperforms NoPE in all 30 evaluated context-length–depth combinations, increasing mean accuracy across the grid from 90.71% to 98.76%. After averaging over target depths at each context length, MiPE maintains 98.22%–99.16% accuracy, whereas NoPE ranges from 89.44% to 93.80%. In contrast, the RoPE variant loses most retrieval accuracy beyond the pretraining context length both with context-matched YaRN and when unscaled.

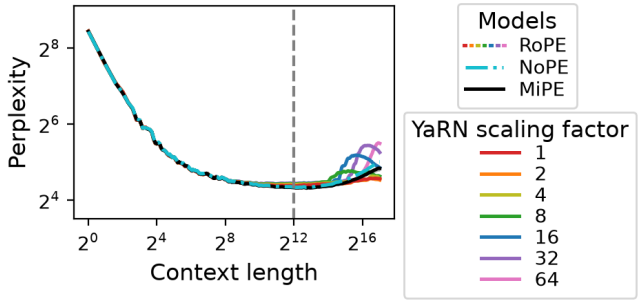


Figure 21: Long-context perplexity for 286M Multiscreen using MiPE, NoPE, or RoPE. The MiPE curve shows the mean across three independently trained runs, and its shaded region, largely obscured by the mean curve, indicates one standard deviation; the NoPE and RoPE curves each use a single run. MiPE and NoPE use no inference-time positional scaling; the colored RoPE curves use the indicated fixed YaRN scaling factors. The gray dashed line marks the pretraining context length (2^{12}).

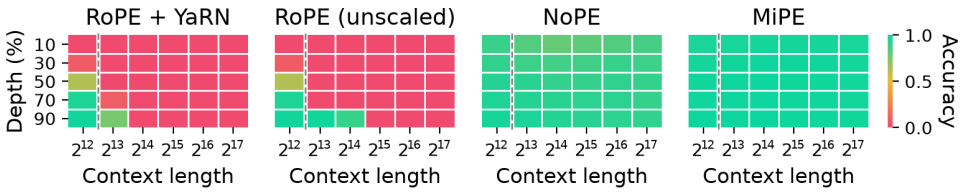


Figure 22: ABCDigits retrieval accuracy for the same 286M Multiscreen positional-encoding variants. The first panel applies context-matched YaRN scaling to RoPE, whereas the second uses unscaled RoPE; MiPE and NoPE require no inference-time positional scaling. The MiPE results are averaged over three independently trained runs, whereas the NoPE and RoPE results each use a single run. The gray dashed lines mark the pretraining context length (2^{12}).

These results show that MiPE does more than avoid the failure of conventional positional rotations: it supplies positional structure that consistently improves retrieval over NoPE while preserving scaling-free length extrapolation. Together with the strong Screening (no window) result in fig. 6, the RoPE ablation should not be interpreted as showing that RoPE inherently causes length-extrapolation failure; rather, it indicates that RoPE is poorly matched to the learned-window Softmask construction in this setting.

R THE ROLE OF THE OUTPUT GATE

The output gate modulates the screening output before its projection back to the model dimension. To assess its role, we compare the standard 286M Multiscreen configuration with a closely capacity-matched variant without the gate.

Removing the gate reduces the parameter count, so we increase both the number of heads and the number of layers from 32 to 38 in the no-gate control. The resulting models contain 286.35M and 288.05M parameters, respectively, making the no-gate control approximately 0.60% larger than the standard model. All other training and evaluation settings are unchanged, and both models are evaluated at 2^{36} training tokens. The standard Multiscreen results average the same three independently trained runs used in the model-scaling experiments in section 4.6, whereas the no-gate control uses a single independently trained run.

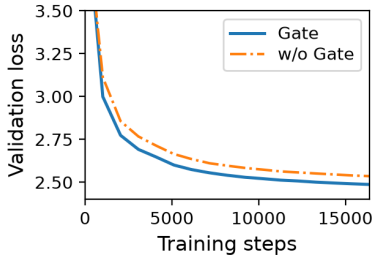


Figure 23: Validation loss over 2^{16} training steps (2^{36} tokens) for standard Multiscreen and the closely capacity-matched variant without the output gate. The Multiscreen curve averages three independently trained runs, and its shaded region, largely obscured by the mean curve, indicates one standard deviation; the no-gate curve uses a single run.

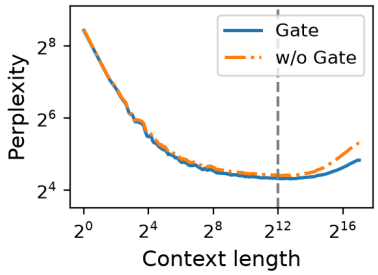


Figure 24: Long-context perplexity for standard Multiscreen and the closely capacity-matched variant without the output gate. The standard Multiscreen curve shows the mean across three independently trained runs, and its shaded region, largely obscured by the mean curve, indicates one standard deviation; the no-gate curve uses a single run. The gray dashed line marks the pretraining context length (2^{12}).

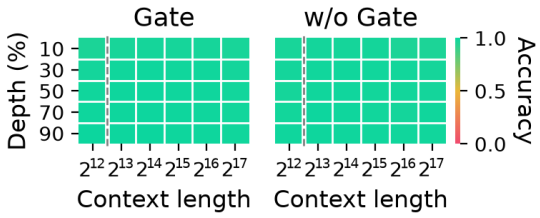


Figure 25: ABCDigits retrieval accuracy for standard Multiscreen and the closely capacity-matched variant without the output gate. Results for standard Multiscreen are averaged over three independently trained runs, whereas the no-gate results use a single run. The gray dashed lines mark the pretraining context length (2^{12}).

As shown in fig. 23, standard Multiscreen reaches consistently lower validation loss throughout training. At 2^{36} training tokens, its validation loss is 2.4871 ± 0.0005, compared with 2.5356 for the no-gate control. Because the no-gate control is slightly larger, this advantage cannot be attributed to a larger parameter count in the gated model. The output gate therefore provides a consistent improvement in overall language-modeling quality.

The no-gate control exhibits consistently higher perplexity, in line with its validation-loss disadvantage, while retaining the same qualitative length-extrapolation behavior as standard Multiscreen (fig. 24). Both variants also maintain near-perfect ABCDigits accuracy across the evaluated context lengths and target depths, with little discernible difference between them (fig. 25). Together, these results indicate that the output gate primarily improves language-modeling quality rather than providing the core mechanism for long-context extrapolation or retrieval.

S TTFT MEASUREMENT PROTOCOL

All model-side TTFT measurements were performed on a single NVIDIA GH200 GPU with batch size 1, causal masking, and mixed precision with bfloat16 matrix multiplications. We used the same Transformer and Multiscreen kernel implementations as in training. Each trained model was measured 100 times, and we report the mean TTFT.

T EXISTING ASSETS AND LICENSES

We summarize the existing datasets, tokenizers, and software packages used in this work, together with their licenses and terms of use.

Datasets. We use SlimPajama (Soboleva et al., 2023) for pretraining. SlimPajama is a cleaned and deduplicated dataset derived from RedPajama (Weber et al., 2024), and is released under the Apache License 2.0 by Cerebras.[5] RedPajama itself aggregates multiple data sources with source-specific licenses and terms of use.

For long-context perplexity evaluation, we use PG-19 (Rae et al., 2020). The PG-19 repository is released under the Apache License 2.0[6], and the benchmark consists of books extracted from Project Gutenberg that were published before 1919. Project Gutenberg states that the vast majority of its eBooks are in the public domain in the United States, while users outside the United States should verify the copyright status under their local laws.[7]

Tokenization. We use the GPT-2 tokenizer (Radford et al., 2019). The GPT-2 code and tokenizer assets are released by OpenAI under a modified MIT license.[8]

Software. Our implementation uses PyTorch (Paszke et al., 2019) for neural-network training and inference. PyTorch is distributed under a BSD-style license.[9] For Transformer baselines, we use PyTorch's `torch.nn.attention.varlen.varlen_attn` implementation with its built-in FlashAttention-2 backend (Dao, 2024). For Multiscreen kernels, we use Triton (Tillet et al., 2019), which is released under the MIT license.[10] We access the downstream benchmarks through version 0.4.12 of the Language Model Evaluation Harness (Sutawika et al., 2026), which is released under the MIT License,[11] and cite the original benchmark sources in Appendix B. The controlled attention-mechanism comparison uses the forgetting-transformer implementation[12] and TritonSigmoid version 0.1.2,[13] both released under the MIT License, and AdaSplash version 0.2.2, released under the BSD 3-Clause License.[14]

[5] https://www.cerebras.ai/blog/slimpajama-a-627b-token-cleaned-and-deduplicated-version-of-redpajama

[6] https://github.com/google-deepmind/pg19

[7] https://www.gutenberg.org/policy/permission.html

[8] https://github.com/openai/gpt-2

[9] https://github.com/pytorch/pytorch

[10] https://github.com/triton-lang/triton

[11] https://github.com/EleutherAI/lm-evaluation-harness

[12] https://github.com/zhixuan-lin/forgetting-transformer

[13] https://github.com/MSDLLCpapers/triton-sigmoid

[14] https://github.com/deep-spin/adasplash