Cracks in the Foundation:
Seemingly Minor Architectural Choices Impact Long Context Extension

Anonymous Authors[1]

Abstract

One might imagine that architectural variations within the dense transformer paradigm have a limited effect on accuracy. However, we demonstrate that this is not the case in the long context setting. Specifically, we show that a set of four minor architectural decisions— all made by at least one of the Olmo, Llama, and Qwen dense model families— have a compoundingly negative effect on long context extensibility. Any one of these choices alone has a minor impact on long context performance, but combining three or more can drop the performance downstream by up to 47%. Furthermore, these differences are not detectable from short-context loss or validation datasets. We demonstrate this with controlled ablations that hold data, tokenizer, and extension recipe fixed while varying normalization, GQA, pretraining context length, and sliding window attention. We show that much of the variation in long context ability across model families is driven by these architectural features and detectable from applying context extension early in pretraining. After over 170,000 GPU hours of training, we release the resulting set of models as ArchPool, a set of 26 comparable 7B models with checkpoints before and after long-context extension. This pool includes several architectures that outperform the Llama 3 architecture on long context extensibility. In an analysis of our ablation models, we identify patterns in attention sink behavior and attention distributions across context that are attributable to specific architectural differences.

1. Introduction

Pretraining large language models is an expensive and time-consuming process. From architectural choices to data selection, many design choices carry the potential to shift the behavior of the resulting downstream model— and these factors can interact in complex ways. Because running experiments at full scale is prohibitively expensive, a critical question is how to validate design decisions early in training or at smaller scale. This challenge is especially acute for capabilities that are elicited later in the development cycle— such as reasoning (Wang et al., 2025) or agentic behavior (Qin et al., 2025)— since architectural choices must be made long before these capabilities can be directly observed.

Long-context processing is an important instance of this problem. Context length is typically extended by modifying positional embeddings and continuing to pretrain at longer context lengths during a midtraining phase at the end of pretraining (Xiong et al., 2024). Because this phase comes late in the development cycle, practitioners must commit to architectural decisions before they can observe how those decisions affect long-context behavior. Compounding this, most long-context extension recipes are developed on a small set of base models: the majority of works focus on extending Llama family models (Fu et al. (2024); Gao et al. (2025); Lu et al. (2024b); Chen et al. (2024); Peng et al. (2026), *inter alia*), with relatively few considering other architectures (e.g. Ding et al. (2024); Zhao et al. (2024); Hu et al. (2024)). As a result, it is unclear how broadly existing recipes transfer, or whether the base architecture itself is a decisive factor in downstream long-context performance, even when comparing only transformer models.[1]

In this work, we demonstrate that a small set of cross-model-family architectural variations account for substantial variation in downstream long context performance by performing a set of data- and optimization-controlled pretraining experiments. We first show that short context performance is not

---

[1] Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

[1] In parallel, recent work on architectures outside of the standard transformer has been motivated by improving long context performance (e.g Gu & Dao (2024); Yang et al. (2025c); Peng et al. (2023)); we focus here on variations *within* the transformer family.

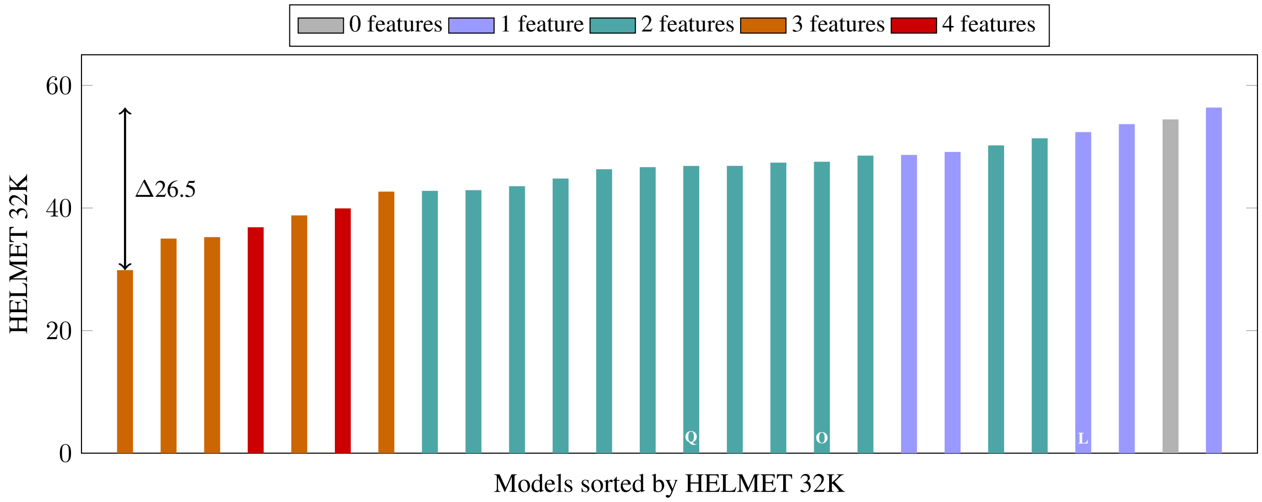

Figure 1. HELMET 32K scores across all ArchPool models with identical data and context extension strategy, sorted worst to best. The colors indicate the count of (individually minor) choices made that downweight long context performance; the combination of these features can dramatically degrade performance, up to 26.5 points on HELMET. (Q), (O), and (L) indicate the Qwen 3, Olmo 3, and Llama 3 architectures; none of these is optimal.

sufficient to predict long context performance by training a sweep of models with the same short context behavior but dramatically variable long context behavior; then, we study how normalization decisions, the use of GQA, sliding window attention, and the pretraining context length shift long context performance. We select only values for these four factors that have been used in Llama 2, Llama 3, Qwen 3, or Olmo 3, and pretrain a sweep of 24 models, ArchPool, which represent varying choices for each factor. We show that even interpolating in this narrow design space can result in dramatically divergent long context performance (e.g. in Figure 1). We identify minimal pairs of architectural changes that cause downstream long context performance to change, characterize the compounding impact of applying multiple of these architectural changes at once, and analyze the attention patterns of these model pairs.

2. Setting

We perform a set of controlled pretraining experiments to construct ArchPool, a set of 26 comparable models in the 7-8B parameter range. Appendix A provides full descriptions.

2.1. Architectural choices ablated

We consider four primary architectural design decisions: normalization strategy, grouped-query attention, the use of sliding windows, and pretraining context length. These features were selected because they differ across several major recent model releases and have explicit connections to the attention mechanism or context length.

Normalization The two normalization factors we consider are layernorm ordering and the presence of QK norm. The observation that QK norm can limit long context performance was first made by Yang et al. (2025b); we replicate this finding in our setting and further explore how specific variants of QK norm and norm order play a role.

QK norm is often implemented layerwise, as a normalization step applied before the concatenated query matrix is split for specific attention heads. This may be implemented as an RMS norm:

$$\hat{Q} = \frac{Q}{\mathrm{RMS}(Q)}\,\gamma^{Q}, \qquad \hat{K} = \frac{K}{\mathrm{RMS}(K)}\,\gamma^{K} \tag{1}$$

Where γ^Q, γ^K are normalization parameters learned for each layer. This *layerwise* implementation of QK norm is used in Olmo 2 and 3 (OLMo et al., 2025; Olmo Team et al., 2025). We also consider the headwise variant of QK norm, used by Qwen 3 (Yang et al., 2025a), Gemma 3 (Gemma Team et al., 2025), and Marin 32B (Marin Community, 2025). Headwise QK norm applies normalization separately to each attention head's queries and keys, learning per-head values γ_h^Q, γ_h^K for $h = 1, \ldots, H$:

$$\hat{Q}_h = \frac{Q_h}{\mathrm{RMS}(Q_h)}\,\gamma_h^{Q}, \qquad \hat{K}_h = \frac{K_h}{\mathrm{RMS}(K_h)}\,\gamma_h^{K} \tag{2}$$

A related question is whether to place the layernorm before or after the sublayer (i.e. prenorm or post-sublayer-norm[2]).

---

[2] Closely related to perinorm, which normalizes the inputs and outputs of the sublayer (Kim et al., 2025)

Applying post-sublayer-norm without QK norm can lead to training divergence because models that apply normalization after the sublayer are more sensitive to gradient instability caused by large or high-variance attention logits.[3]

Grouped query attention (GQA). GQA (Ainslie et al., 2023) increases inference efficiency by reusing the same key-value matrices for multiple query heads in the same layer, reducing the size of the key-value cache. Typical GQA models share 8 key-value heads for 32 query heads (Grattafiori et al., 2024; Yang et al., 2025a). Note that because GQA shares some W_K and W_V parameters, it reduces the total capacity of the network; when this occurs, we adjust the intermediate size slightly to result in the same total parameter count. This should benefit models with GQA in our comparisons.

Sliding window attention (SWA). Modern sliding window attention implementations generally intersperse layers of local window attention with layers of full attention in a many-to-one pattern (Olmo Team et al., 2025; Gemma Team et al., 2025). We use the Olmo 3 configuration, which has 3 local attention layers of 4096 context for every 1 global attention layer.

Pretrained context length. Zhao et al. (2024) observe that long context ability is impacted by the pretraining context length, with models trained at a longer context length able to support longer post-extension context length as well. We pretrain Olmo and Llama variants at 4096 context length, matching the context length of the prior generation for each model. This allows us to compare models trained at 4K directly with models pretrained with a 4K sliding window (but 8K total context).

2.2. Training

We pretrain each model for 140B tokens (the Chinchilla-optimal amount of data (Hoffmann et al., 2022b)). Then, we adjust the RoPE theta for context extension (Xiong et al., 2023) and continue pretraining for 10B tokens on 64K context data from the Longmino mix (Olmo Team et al., 2025), annealing the learning rate to 0. We hold the data selection and ordering, the learning rate and schedule, and the tokenizer constant across all models in ArchPool.

Ideally, we would also standardize initialization. However, because not all models have the same parameters (e.g. if GQA or QK norm are added), we cannot exactly synchronize the initializations. Where possible, we reuse the same initalization across runs; if there is a minor difference in parameterization, we use the remainder of the same initialization and only re-initialize the new parameters. Appendix A identifies the initialization for each model, and we further discuss the impact of model initialization in Section 4.

2.3. Evaluation

We evaluate all models downstream on 3 popular measures of long context performance: RULER (Hsieh et al., 2024), which represents a sweep of synthetic Needle-in-a-Haystack (NIAH) style tasks of increasing complexity; HELMET (Yen et al., 2025), which additionally considers in-context learning, reranking, and question-answering tasks; and LongPPL (Fang et al., 2025), a variant of perplexity that only considers tokens that require long-range dependencies to predict. Because we are working with weak models early in pretraining, we consider only the subtasks from HELMET that do not require generating long spans of text to evaluate with an LM judge. We observe that all three measures correlate closely; for readability, we primarily report HELMET at 32K in the main text, unless trends differ across the three measures. To compare model architectures, we construct multiple paired comparisons, holding constant all other factors, wherever possible. We also fit linear regressions from architectural features, short-context metrics, and observed attention distributions to measure the predictive power of each feature across ArchPool.

3. Short context metrics are not always predictive of long context performance

The models in ArchPool demonstrate that seemingly small changes in the model recipe have a dramatic downstream effect on long context extensibility. The performance of these models ranges from 29.9 to 56.4 on HELMET at 32K, or 44.7 to 67.7 on RULER at 32K.

Can this downstream effect be predicted from pretraining metrics? We observe that short context metrics are surprisingly poor predictors of long-context behavior. We consider a sweep of measures to try to predict long context performance downstream, demonstrating that standard pretraining metrics are insufficient to predict long context performance.

Intrinsic metrics. We measure the loss for each training run at the end of pretraining and at the end of long context extension and compute the correlation between these values and the resulting HELMET score. Training loss correlates only weakly with downstream long context performance ($R^2 = 0.29$); surprisingly, the loss during pretraining is *more predictive* of downstream score than the loss during context extension ($R^2 = 0.06$). We measure perplexity of the pre-context-extension model over 11 held-out text samples from differing domains (Magnusson et al., 2024); these scores range from completely uncorrelated

---

[3] We confirm this experimentally as well; in a run with post-sublayer-norm and no QK norm, training consistently diverged before 140B tokens.

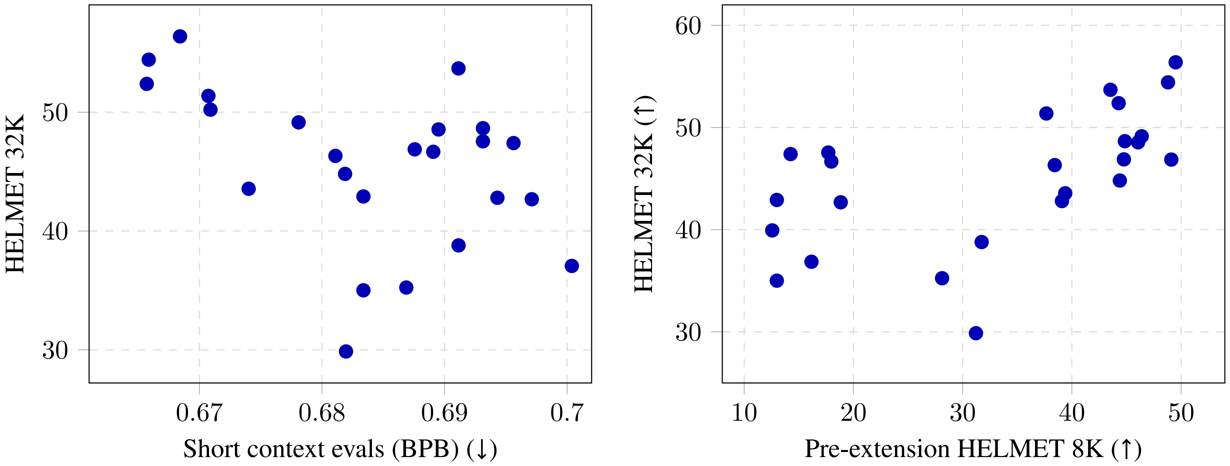

Figure 2. Benchmark scores pre-extension largely fail to predict long-context benchmark scores post-extension, even when evaluating on the shorter-context version of the same benchmark. Each point is a model from ArchPool.

to weakly correlated with downstream long context scores. The best-correlated splits do not align with common understanding of the types of data that require long-range dependencies: a sample of data from WikiText (Merity et al., 2016) is the most correlated ($R^2 = 0.39$), with perplexity of samples from academic texts and code showing little to no correlation with long context behavior downstream ($0.01 \le R^2 \le 0.20$).[4]

Downstream evaluations. During pretraining, models are often periodically evaluated on a set of development benchmarks. We evaluate on a set of 16 in-loop benchmarks, detailed in Appendix D.3; because early pretraining checkpoints are often inconsistent at answering in multiple-choice format (Bunn et al., 2025), we score by the bits-per-byte on the correct answer for each benchmark question instead of accuracy (Heineman et al., 2025). Figure 2 (left) shows the average of these evaluations graphed against downstream HELMET score; there is a slight correlation between the two ($R^2 = 0.17$), but this fails to exceed the predictiveness of the best perplexity measures. Note that all scores are very close together: in our setting, where models are trained with the same data and very similar architectures, little difference is observable between models pre-context-extension. Clearly, standard in-loop evaluations are not sufficient to provide signal for downstream long context extensibility.

Is this merely an issue of choosing the wrong benchmarks? We consider a benchmark more directly related to our evaluations for performance downstream: the shortest context split of HELMET, which, at 8K, is possible for most of our models to process without context extension. Recall that these models are being evaluated pre-anneal and early in training, so we expect scores to be quite poor. Figure 2 (right) graphs these pre-extension scores against the post-extension performance; while HELMET scores pre-extension are both more variable and slightly more predictive ($R^2 = 0.32$) than other short-context evaluations, they still fail to predict double-digit swings in HELMET performance downstream.

4. Impact of architectural choices

Clearly, the differing performance of these models is not solely attributable to how well each model has fit the pretraining corpus. We consider the individual effect of each of the four factors we have identified in turn and show that their behavior is best modeled as an additive impairment to long context: the individual features selected do not matter nearly as much as the number of long-context-inhibiting features present. In Appendix B, we also discuss the impact of additional features that we considered: pretraining with linear layers quantized to float8 and changing the pretraining run's random initialization.

All four features reduce long context capability downstream. In paired comparison runs, each architectural feature results in some degradation of long context performance downstream. Pretraining at 4096 instead of 8192 context length and training with sliding window layers result in modest average degradations of 1-2 points on HELMET at 32K. Figure 3 shows that pretraining with GQA configurations that use increasingly aggressive degrees of query sharing (i.e. decreasing the number of KV heads) degrades performance from the Llama 3 configuration, and pretraining with *more* KV heads than the Llama 3 architecture improves performance. But the largest individual performance impact by far arises from normalization choice. On the Olmo

---

[4] These trends hold even if we evaluate correlation with LongPPL, which is a more intrinsic metric of long context quality. See a full per-dataset breakdown in Appendix D.3.

architecture, changing Olmo 3's choice of QK norm and post-sublayer-norm to prenorm results in a 6 point gain on HELMET; conversely, adding these features to Llama 3's architecture results in a 3.8 point drop.[5]

Minor effects of architecture compound. With the exception of QK norm, most individual architectural features have relatively minor effects in controlled comparisons where this is the only potentially detrimental feature tested. However, when paired together, these feaures can have much more significant effects. For instance, sliding window has a minor negative effect when it is applied to a model without GQA (-1.1 points from the full attention configuration). However, in multiple comparison runs, adding sliding window layers to a configuration that also has GQA results in a dramatic drop in performance: -9 points on average. The worst-scoring runs in ArchPool combine two or more features that limit the expressivity of attention— for instance, the single worst configuration combines GQA, sliding windows, and headwise QK norm, for a total effect on performance far worse than the sum of these individual effects.

We find that downstream long context performance can be well-estimated by simply counting how many of these four architectural choices are present in a ArchPool model— this single numerical feature is the single most predictive feature for downstream long context performance (in-sample $R^2 = 0.67$, LOO $R^2 = 0.61$), outperforming even a linear regression over the four individual axes of architectural variation. It is not any single feature which results in catastrophically poor long context extensibility, but the combination of several features that each reduce the expressivity of attention.

Figure 1 also shows that combining all four features is not necessarily worse than combining three features in our setting. We hypothesize that this may be because at least some of these features impact long context abilities in overlapping ways: for instance, sliding window attention restricts the model to only 4K context at some layers during pretraining, and pretraining at 4K restricts the model to only 4K context at *all* layers during pretraining. However, the predictor that counts the presence of any of the four features remains more predictive than any predictor that collapses two of the features into the same category.

Llama 3 is a particularly good architecture for long context. While it was previously unclear whether the ease of extensibility for Llama 3 was due to architectural or data factors (since the pretraining data for Llama 3 was not publicly disclosed), we find evidence that this is primarily an architectural phenomenon. The Llama 3 architecture model is one of the best models in the design space. This suggests that context extension recipes developed with this model may require additional effort to apply to other architectures— for instance, our results show that same context extension recipe applied to the Llama 3, Qwen 3, and Olmo 3 architectures is far more effective for the Llama-like model, even when the other models were pretrained for the same duration and data as the Llama-like model. This also helps explain our empirical observation that Olmo 3 Base is more challenging to context extend than Llama 3 Base.

5. Analysis

5.1. Is this measuring token efficiency or a fundamental capability gap?

Our main experiments compare models after the same 10B token context extension. However, it's possible that extending with a much longer context extension phase would wash out these differences. To test this, we choose three representative models at different quality points in ArchPool: the models with the architectures of Llama 3, Olmo 3, and the worst-scoring architecture on HELMET downstream. We refer to these models as L, O, and W. On this trio of models, we conduct 1B[6] and 50B context extensions and compare performance with increasing amounts of data. Figure 4 shows the performance of these models with increasing amounts of data. As expected, all models see better long context performance with 50B token extensions; however, this effect is not enough to compensate for architecture-driven differences in base quality. Even after 50B tokens, the worst architecture does not reach the same performance as the Llama architecture achieves after 1B tokens, and the differences between the three architectures remains relatively stable. While it's possible that context extensions at a more extreme scale could overpower these architectural effects, we see no evidence of this occurring up to the 50B token scale, where the context extension phrase represents 26% of the total tokens seen by the model.

5.2. Do these differences persist in longer pretraining runs?

To be able to measure many architectural configurations, we extend models after 140B tokens of pretraining. However, most modern models are trained far longer, with long context extension often applied after the model has seen trillions of tokens (Olmo Team et al., 2025). Would these effects wash out with a longer pretraining run?

We measure the difference between two configurations of

---

[5] We find that the choice of headwise versus layerwise QK norm and the ordering of prenorm vs post-sublayer-norm have some effect as well, although the decision to apply QK norm at all dominates; for more details, see Appendix B.

[6] Note that because we perform each context extension run as an anneal to 0 learning rate, the 1B extension is not merely an earlier checkpoint in the 10B extension run

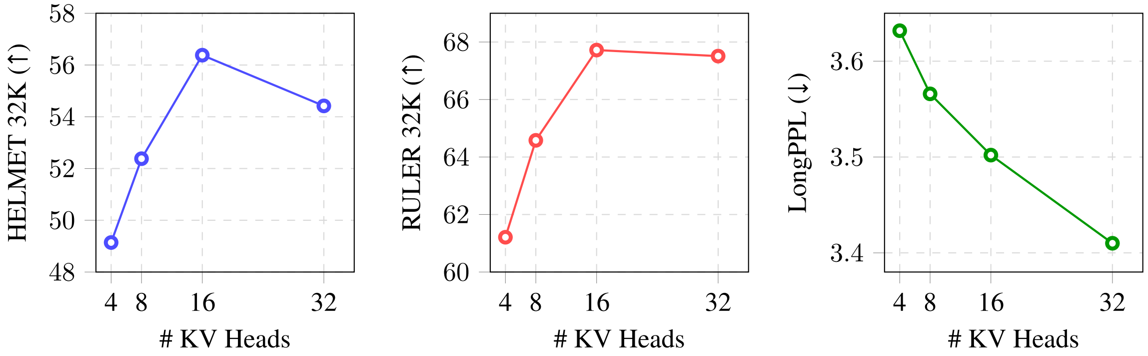

Figure 3. GQA is harmful to long context performance, even when adjusting for comparable parameter count. Variants of Llama that increase the number of KV heads improve performance on long context benchmarks. 8 KV heads is the Llama GQA configuration, and 32 KV heads indicates no GQA.

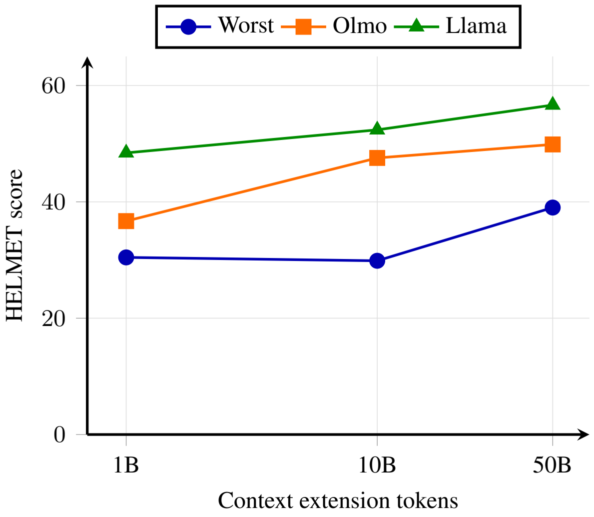

Figure 4. Performance on HELMET after 1B, 10B, or 50B token extension for three representative runs (the worst architecture, the Olmo 3 architecture, and the Llama 3 architecture). Longer context extension fails to wash out architectural differences.

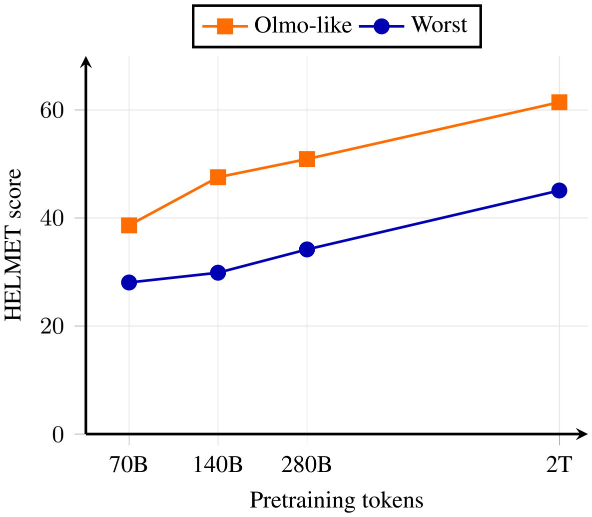

Figure 5. HELMET score at extensions performed after progressively more pretraining tokens for two training runs. The difference in long context behavior is consistent, especially from 140B onwards.

interest on much longer pretraining runs. At 70B, 140B, 280B, and 2T tokens into a longer pretraining run, we adjust the RoPE theta and perform a 10B anneal on long context (64K) data. Figure 5 shows the relative behavior of these models on downstream long context evaluations at each point. We show that nontrivial long context performance can be recovered from extensions at least as early as 70B, and the relative performance of the two architectures remains fairly consistent over the course of pretraining. Because we observe that the two models are closer in performance at 70B than at any point 140B or later, we train all other ArchPool models to 140B tokens before extending to ensure we are not missing differences across models.

5.3. Attention behaviors in ArchPool

Armed with a set of models that differ in downstream performance, we now seek to understand the finegrained differences between these models. We compute a number of statistics of the attention distribution across all models, measured over a set of 100 long documents selected at random from an even split of Project Gutenberg books, government reports (Huang et al., 2021), FineWeb PDFs (Penedo et al., 2024), and legal texts (Henderson* et al., 2022). We measure the entropy of the attention distribution at each attention head at positions 1K, 4K, 16K, and 32K. We make separate comparisons for layers using full attention and (where applicable) sliding window attention layers. We also measure the percentage of attention mass in the attention sink (defined as the first 100 tokens of visible context) and the local context (defined as the 100 tokens of context immediately preceding the current token).

High entropy and attention sinks are positive indicators—and QK norm dampens both effects. Yang et al. (2025b) observe that models with QK-norm have higher entropy in the attention distribution (i.e. less peaky attention), which they suggest makes it more challenging for the model to attend over longer contexts. We also observe this effect and note a specific downstream consequence. Attention entropy is also influenced by the presence of attention sinks (Xiao et al., 2024): positions early in the context window that consistently receive substantial attention mass. The presence of these sinks is theorized to be a result of models attempting to "discard" excess attention weight (Bondarenko et al., 2023; Qiu et al., 2026), and believed to make models more difficult to quantize (Ye et al., 2025).

Figure 6 visualizes both sink attention and attention entropy. While the presence of attention sinks is generally considered negative (e.g. Qiu et al. (2025) names reduction of attention sinks as a core benefit of their approach), the presence of attention sinks here correlates with improved long context performance ($R^2 = 0.38$). In the absence of another mechanism such as differential attention (Ye et al., 2025) or gating (Qiu et al., 2025), attention sinks appear to be the default strategy learned by QK-norm-less transformers to compensate for excess attention. Thus, attention sink behavior corresponds with long context abilities in ArchPool.

Retrieval heads Wu et al. (2024) propose the existence of *retrieval heads*, specialized attention heads that are primarily responsible for retrieving information from prior context in both short and long context processing. They demonstrate that these heads are disproportionately responsible for the ability to perform long context fact retrieval. We measure this by running an analysis on the same set of documents with a needle-in-a-haystack (NIAH) task injected. We compute, at the end of prefill, the percentage of attention that is placed on the needle tokens. Then, we generate a completion for each example; if the model successfully generates the needle text, we measure how much each individual attention head attended back to the needle text during generation and use this to compute the retrieval score for each head (Wu et al., 2024).

At prefill time, we observe no more than a slight correlation between the ability to place more attention on the needle tokens and performance on downstream evaluations. Models with QK norm (headwise *or* layerwise) uniformly place less attention on the needle tokens than models without QK norm. However, during generation we observe little difference in retrieval head behavior across models in ArchPool, with very low retrieval head scores across models. These models may be too weak to reliably identify retrieval heads. Alternatively, some other mechanism may govern long context abilities in models early in training.

6. Related work

Architectural choices and long-context extensibility. Prior works in this area primarily focus on positional embeddings or modifications to the extension method itself. The closest work to our setting, Yang et al. (2025b), compares RoPE, NoPE, and QK-normalized RoPE models and shows that these three variants can differ substantially on long-context evaluation despite similar standard performance. More broadly, work on positional design and extrapolation shows that even a single architectural axis can have a large effect on longer-range behavior: Kazemnejad et al. (2023), Press et al. (2022), and Wang et al. (2024) find substantial differences across positional embedding variants, and Lu et al. (2024a) find that sparse attention generally lags full attention for long context. Other work studies how to improve extrapolation by changing attention or positional embedding settings during context extension (Sun et al., 2023; Chen et al., 2023). A separate line of long-context work changes the architecture, for example by modifying the attention mechanism (Zimerman & Wolf, 2023), through recurrence and memory (Dai et al., 2019), or newer architectures designed for effectively unbounded context (Ma et al., 2024); see Huang et al. (2023) for a survey of this space. In contrast, we focus on choices made before pretraining that may *unintentionally* determine downstream long-context extension outcomes.

Performance prediction from smaller scales. A separate methodological literature asks how to make expensive model-development decisions using smaller or earlier experiments. Controlled pretraining suites such as Pythia (Biderman et al., 2023), open-science efforts such as OLMo and LLM360 (Groeneveld et al., 2024; Liu et al., 2023), and small-experiment prediction work such as DataDecide (Magnusson et al., 2025) show that many choices can be studied scientifically before full-scale deployment; scaling-law work shows that even expensive allocation decisions can be forecast from smaller runs (Hoffmann et al., 2022a). Some recent work has also looked at integrating features of the data distribution or architecture into scaling law predictions (Liu et al., 2026). Magnusson et al. (2024); Fang et al. (2025) argue that easy scalar proxies such as perplexity can miss important behavior during training runs, while recent work on evaluation reliability shows that the signal-to-noise properties of benchmarks themselves can strongly affect how useful small experiments are for model-development decisions (Heineman et al., 2025). More general work on capability prediction emphasizes that downstream behavior is often harder to forecast from pretraining signals alone (Schaeffer et al., 2025). Our setting is far cheaper than full-scale long context training, but more expensive than

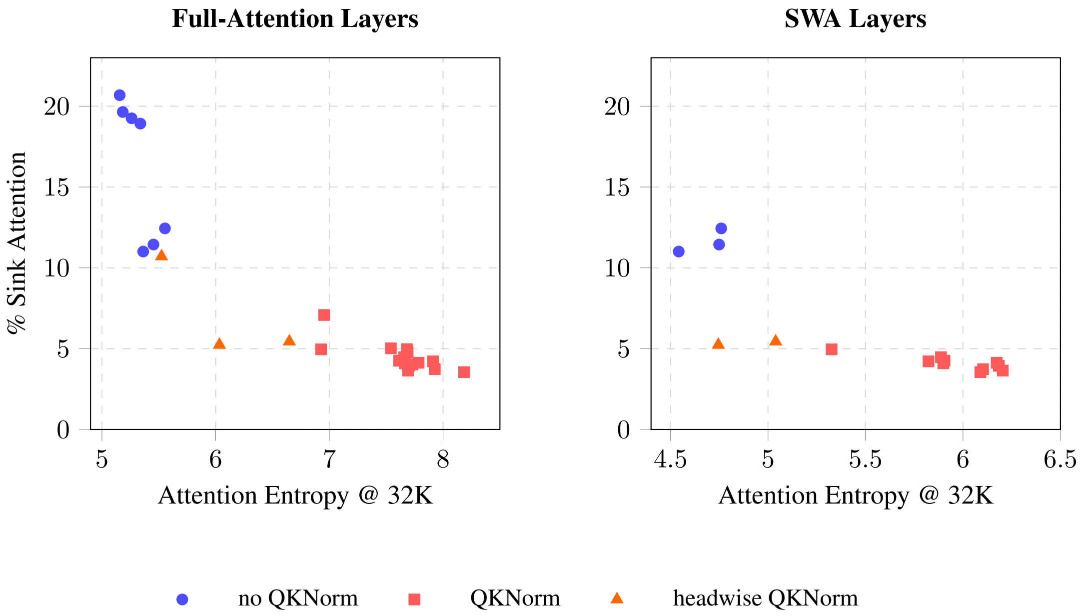

Figure 6. Attention entropy and the presence of a a strong attention sink both cluster by presence of QK norm. Models with sliding window attention have less attention sink behavior on full attention layers; these represent the lower cluster of blue dots.

measuring pretraining signals alone; ArchPool also serves as a set of models with empirical downstream results to evaluate future performance prediction metrics.

7. Conclusion

We demonstrate that a series of small, individually reasonable architectural perturbations, well grounded in the literature, can result in dramatically reduced long context capabilities. We show that this degradation is difficult to detect in short context metrics but detectable from context extension runs very early into pretraining. This suggests two interesting directions for future research: evaluating the minimal size or token budget at which these effects can be reliably measured, to further reduce the cost of architectural experimentation; and devising better proxy metrics for short context models to estimate long context performance without performing a context extension. We also believe there may be more to understand mechanistically about the differences between ArchPool models. Finally, ArchPool's parallel runs, traversing a similar optimization problem with slightly different architectures, may be useful for research into other phenomena in early pretraining. To this end, we release 38 checkpoints for each model, representing the full pretraining and long context extension.

Each feature we ablate has some clear benefit— stability for normalization, pretraining efficiency for context length, and inference efficiency for sliding window and GQA— and individually may be justified because of these other factors. Yet the combination of these features results in unacceptable long context extensibility. By exposing the interplay between these factors in a controlled setting, we hope to enable model developers to make more informed choices about their architecture design and to spur future research into alternatives that better navigate these tradeoffs.

Training cost

Pretraining ablations carry a heavy computational cost. We estimate that the cost of training the 26 models for this paper was approximately 170,000 H100 hours, in addition to the costs of initial experimentation, evaluation, and additional ablation runs. Using the calculations from Morrison et al. (2025), this is equivalent to approximately 42.63 metric tons of CO2 emitted, if our hardware was of equivalent power usage. In releasing all artifacts including intermediate checkpoints, we hope that this cost can be amortized by the reuse of ArchPool.

References

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. Gqa: Training generalized multi-query transformer models from multi-head checkpoints, 2023. URL `https://arxiv.org/abs/2305.13245`.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program synthesis with large language models. *arXiv* *preprint arXiv:2108.07732*, 2021.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory 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 *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2397–2430. PMLR, 2023. URL `https://proceedings.mlr.press/v202/biderman23a.html`.

Yelysei Bondarenko, Markus Nagel, and Tijmen Blankevoort. Quantizable transformers: Removing outliers by helping attention heads do nothing, 2023. URL `https://arxiv.org/abs/2306.12929`.

Alec Bunn, Sarah Wiegreffe, and Ben Bogin. Fine-tune on the format: First improving multiple-choice evaluation for intermediate LLM checkpoints. In Ofir Arviv, Miruna Clinciu, Kaustubh Dhole, Rotem Dror, Sebastian Gehrmann, Eliya Habba, Itay Itzhak, Simon Mille, Yotam Perlitz, Enrico Santus, João Sedoc, Michal Shmueli Scheuer, Gabriel Stanovsky, and Oyvind Tafjord (eds.), *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pp. 511–521, Vienna, Austria and virtual meeting, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-261-9. URL `https://aclanthology.org/2025.gem-1.46/`.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models, 2025. URL `https://arxiv.org/abs/2405.09818`.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code, 2021. URL `https://arxiv.org/abs/2107.03374`.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation, 2023. URL `https://arxiv.org/abs/2306.15595`.

Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. Longlora: Efficient fine-tuning of long-context large language models, 2024. URL `https://arxiv.org/abs/2309.12307`.

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.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2978–2988, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1285. URL `https://aclanthology.org/P19-1285/`.

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, 2024. URL `https://arxiv.org/abs/2402.13753`.

Lizhe Fang, Yifei Wang, Zhaoyang Liu, Chenheng Zhang, Stefanie Jegelka, Jinyang Gao, Bolin Ding, and Yisen Wang. What is wrong with perplexity for long-context language modeling? In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=fL4qWkSmtM`.

Yao Fu, Rameswar Panda, Xinyao Niu, Xiang Yue, Hannaneh Hajishirzi, Yoon Kim, and Hao Peng. Data engineering for scaling language models to 128k context, 2024. URL `https://arxiv.org/abs/2402.10171`.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The Pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

Tianyu Gao, Alexander Wettig, Howard Yen, and Danqi Chen. How to train long-context language models (effectively), 2025. URL `https://arxiv.org/abs/2410.02660`.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. URL `https://arxiv.org/abs/2503.19786`.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan

Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal

Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL `https://arxiv.org/abs/2407.21783`.

Sidney Greenbaum and Gerald Nelson. The International Corpus of English (ICE) project. *World Englishes*, 15(1): 3–15, 1996. doi: 10.1111/j.1467-971X.1996.tb00088.x.

Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, William Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah A. Smith, and Hannaneh Hajishirzi. OLMo: Accelerating the science of language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15789–15809. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.841. URL `https://aclanthology.org/2024.acl-long.841/`.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces, 2024. URL `https://arxiv.org/abs/2312.00752`.

David Heineman, Valentin Hofmann, Ian Magnusson, Yuling Gu, Noah A. Smith, Hannaneh Hajishirzi, Kyle Lo, and Jesse Dodge. Signal and noise: A framework for reducing uncertainty in language model evaluation. In *Advances in Neural Information Processing Systems*, 2025. URL `https://openreview.net/forum?id=sAFottNlra`.

Peter Henderson*, Mark S. Krass*, Lucia Zheng, Neel Guha, Christopher D. Manning, Dan Jurafsky, and Daniel E. Ho. Pile of law: Learning responsible data filtering from the law and a 256gb open-source legal dataset, 2022. URL `https://arxiv.org/abs/2207.00220`.

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

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models, 2022a. URL `https://arxiv.org/abs/2203.15556`.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models, 2022b. URL `https://arxiv.org/abs/2203.15556`.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, and Boris Ginsburg. RULER: What's the real context size of your long-context language models? In *First Conference on Language Modeling*, 2024. URL `https://openreview.net/forum?id=kIoBbc76Sy`.

Zhiyuan Hu, Yuliang Liu, Jinman Zhao, Suyuchen Wang, Yan Wang, Wei Shen, Qing Gu, Anh Tuan Luu, See-Kiong Ng, Zhiwei Jiang, and Bryan Hooi. Longrecipe: Recipe for efficient long context generalization in large language models, 2024. URL `https://arxiv.org/abs/2409.00509`.

Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. Efficient attentions for long document summarization, 2021. URL `https://arxiv.org/abs/2104.02112`.

Yunpeng Huang, Jingwei Xu, Zixu Jiang, Junyu Lai, Zenan Li, Yuan Yao, Taolue Chen, Lijuan Yang, Zhou Xin, and Xiaoxing Ma. Advancing transformer architecture in long-context large language models: A comprehensive survey, 2023. URL `https://arxiv.org/abs/2311.12351`.

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*, volume 36, 2023. URL `https://openreview.net/forum?id=Drrl2gcjzl`.

Jeonghoon Kim, Byeongchan Lee, Cheonbok Park, Yeontaek Oh, Beomjun Kim, Taehwan Yoo, Seongjin Shin, Dongyoon Han, Jinwoo Shin, and Kang Min Yoo. Peri-ln: Revisiting normalization layer in the transformer architecture, 2025. URL https://arxiv.org/abs/2502.02732.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. Solving quantitative reasoning problems with language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 3843–3857, 2022.

Emmy Liu, Amanda Bertsch, Lintang Sutawika, Lindia Tjuatja, Patrick Fernandes, Lara Marinov, Michael Chen, Shreya Singhal, Carolin Lawrence, Aditi Raghunathan, Kiril Gashteovski, and Graham Neubig. Not-just-scaling laws: Towards a better understanding of the downstream impact of language model design decisions, 2026. URL https://arxiv.org/abs/2503.03862.

Zhengzhong Liu, Aurick Qiao, Willie Neiswanger, Hongyi Wang, Bowen Tan, Tianhua Tao, Junbo Li, Yuqi Wang, Suqi Sun, Omkar Pangarkar, Richard Fan, Yi Gu, Victor Miller, Yonghao Zhuang, Guowei He, Haonan Li, Fajri Koto, Liping Tang, Nikhil Ranjan, Zhiqiang Shen, Xuguang Ren, Roberto Iriondo, Mu Cun, Zhiting Hu, Mark Schulze, Preslav Nakov, Tim Baldwin, and Eric Xing. LLM360: Towards fully transparent open-source LLMs, 2023. URL https://arxiv.org/abs/2312.06550.

Yi Lu, Jing Nathan Yan, Songlin Yang, Justin T. Chiu, Siyu Ren, Fei Yuan, Wenting Zhao, Zhiyong Wu, and Alexander M. Rush. A controlled study on long context extension and generalization in LLMs, 2024a. URL https://arxiv.org/abs/2409.12181.

Yi Lu, Jing Nathan Yan, Songlin Yang, Justin T. Chiu, Siyu Ren, Fei Yuan, Wenting Zhao, Zhiyong Wu, and Alexander M. Rush. A controlled study on long context extension and generalization in llms, 2024b. URL https://arxiv.org/abs/2409.12181.

Xuezhe Ma, Xiaomeng Yang, Wenhan Xiong, Beidi Chen, Lili Yu, Hao Zhang, Jonathan May, Luke Zettlemoyer, Omer Levy, and Chunting Zhou. Megalodon: Efficient LLM pretraining and inference with unlimited context length. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/840abfadd04c967feaa2a49aba94a32d-Abstract-Conference.html.

Ian Magnusson, Akshita Bhagia, Valentin Hofmann, Luca Soldaini, Ananya Harsh Jha, Oyvind Tafjord, Dustin Schwenk, Evan Pete Walsh, Yanai Elazar, Kyle Lo, Dirk Groeneveld, Iz Beltagy, Hannaneh Hajishirzi, Noah A. Smith, Kyle Richardson, and Jesse Dodge. Paloma: A benchmark for evaluating language model fit. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/760b2d94398aa61468aa3bc11506d9ea-Abstract-Dataset
and_Benchmarks_Track.html.

Ian Magnusson, Nguyen Tai, Ben Bogin, David Heineman, Jena D. Hwang, Luca Soldaini, Akshita Bhagia, Jiacheng Liu, Dirk Groeneveld, Oyvind Tafjord, Noah A. Smith, Pang Wei Koh, and Jesse Dodge. DataDecide: How to predict best pretraining data with small experiments. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*. PMLR, 2025. URL https://proceedings.mlr.press/v267/magnusson25a.html.

Marin Community. Marin 32b retrospective. Technical report, Marin, 2025. URL https://marin.readthedocs.io/en/latest/reports/marin-32b-retro/.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models, 2016.

Jacob Daniel Morrison, Clara Na, Jared Fernandez, Tim Dettmers, Emma Strubell, and Jesse Dodge. Holistically evaluating the environmental impact of creating language models. *ArXiv*, abs/2503.05804, 2025. URL https://api.semanticscholar.org/CorpusID:276902612.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Allyson Ettinger, Michal Guerquin, David Heineman, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Jake Poznanski, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2 olmo 2 furious, 2025. URL https://arxiv.org/abs/2501.00656.

Olmo Team, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, Pradeep Dasigi, Robert Berry, Saumya Malik, Saurabh Shah, Scott Geng, Shane Arora, Shashank Gupta, Taira Anderson, Teng Xiao, Tyler Murray, Tyler Romero, Victoria Graf, Akari Asai, Akshita Bhagia, Alex Wettig, Alisa Liu, Aman Rangapur, Chloe Anastasiades, Costa Huang, Dustin Schwenk, Harsh Trivedi, Ian Magnusson, Jaron Lochner, Jiacheng Liu, Lj Miranda, Maarten Sap, Malia Morgan, Michael Schmitz, Michal Guerquin, Michael Wilson, Regan Huff, Ronan Le Bras, Rui Xin, Rulin Shao, Sam Skjonsberg, Shannon Zejiang Shen, Shuyue Stella Li, Tucker Wilde, Valentina Pyatkin, Will Merrill, Yapei Chang, Yuling Gu, Zhiyuan Zeng, Ashish Sabharwal, Luke Zettlemoyer, Pang Wei Koh, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. Olmo 3: Charting a path through the model flow to lead open-source ai. Technical report, Allen Institute for AI, 2025. Technical report.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The fineweb datasets: Decanting the web for the finest text data at scale, 2024. URL `https://arxiv.org/abs/2406.17557`.

Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, Huanqi Cao, Xin Cheng, Michael Chung, Matteo Grella, Kranthi Kiran GV, Xuzheng He, Haowen Hou, Jiaju Lin, Przemyslaw Kazienko, Jan Kocon, Jiaming Kong, Bartlomiej Koptyra, Hayden Lau, Krishna Sri Ipsit Mantri, Ferdinand Mom, Atsushi Saito, Guangyu Song, Xiangru Tang, Bolun Wang, Johan S. Wind, Stanislaw Wozniak, Ruichong Zhang, Zhenyuan Zhang, Qihang Zhao, Peng Zhou, Qinghua Zhou, Jian Zhu, and Rui-Jie Zhu. Rwkv: Reinventing rnns for the transformer era, 2023. URL `https://arxiv.org/abs/2305.13048`.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. Yarn: Efficient context window extension of large language models, 2026. URL `https://arxiv.org/abs/2309.00071`.

Ofir Press, Noah A. Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. In *The Tenth International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=R8sQPpGCv0`.

Jiarui Qin, Yunjia Xi, Junjie Huang, Renting Rui, Di Yin, Weiwen Liu, Yong Yu, Weinan Zhang, and Xing Sun. Aptbench: Benchmarking agentic potential of base llms during pre-training, 2025. URL `https://arxiv.org/abs/2510.24397`.

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 *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL `https://openreview.net/forum?id=1b7whO4SfY`.

Zihan Qiu, Zeyu Huang, Kaiyue Wen, Peng Jin, Bo Zheng, Yuxin Zhou, Haofeng Huang, Zekun Wang, Xiao Li, Huaqing Zhang, Yang Xu, Haoran Lian, Siqi Zhang, Rui Men, Jianwei Zhang, Ivan Titov, Dayiheng Liu, Jingren Zhou, and Junyang Lin. A unified view of attention and residual sinks: Outlier-driven rescaling is essential for transformer training, 2026. URL `https://arxiv.org/abs/2601.22966`.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL `http://jmlr.org/papers/v21/20-074.html`.

Machel Reid, Victor Zhong, Suchin Gururangan, and Luke Zettlemoyer. M2D2: A massively multi-domain language modeling dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 964–975, 2022. doi: 10.18653/v1/2022.emnlp-main.63.

Rylan Schaeffer, Hailey Schoelkopf, Brando Miranda, Gabriel Mukobi, Varun Madan, Adam Ibrahim, Herbie Bradley, Stella Biderman, and Sanmi Koyejo. Why has predicting downstream capabilities of frontier AI models with scale remained elusive? In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*. PMLR, 2025. URL `https://proceedings.mlr.press/v267/schaeffer25b.html`.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Evan Walsh, Luke Zettlemoyer, Noah Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. Dolma: an open corpus of three trillion tokens for language model pretraining research. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15725–15788, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.840. URL https://aclanthology.org/2024.acl-long.840/.

Yutao Sun, Li Dong, Barun Patra, Shuming Ma, Shaohan Huang, Alon Benhaim, Vishrav Chaudhary, Xia Song, and Furu Wei. A length-extrapolatable transformer. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14590–14604, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.816. URL https://aclanthology.org/2023.acl-long.816/.

Jie Wang, Tao Ji, Yuanbin Wu, Hang Yan, Tao Gui, Qi Zhang, Xuanjing Huang, and Xiaoling Wang. Length generalization of causal transformers without position encoding. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14024–14040, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.834. URL https://aclanthology.org/2024.findings-acl.834/.

Zengzhi Wang, Fan Zhou, Xuefeng Li, and Pengfei Liu. Octothinker: Mid-training incentivizes reinforcement learning scaling, 2025. URL https://arxiv.org/abs/2506.20512.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. Retrieval head mechanistically explains long-context factuality, 2024. URL https://arxiv.org/abs/2404.15574.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks, 2024. URL https://arxiv.org/abs/2309.17453.

Wenhan Xiong, Jingyu Liu, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou, Louis Martin, Rashi Rungta, Karthik Abinav Sankararaman, Barlas Oguz, Madian Khabsa, Han Fang, Yashar Mehdad, Sharan Narang, Kshitiz Malik, Angela Fan, Shruti Bhosale, Sergey Edunov, Mike Lewis, Sinong Wang, and Hao Ma. Effective long-context scaling of foundation models, 2023. URL https://arxiv.org/abs/2309.16039.

Wenhan Xiong, Jingyu Liu, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou, Louis Martin, Rashi Rungta, Karthik Abinav Sankararaman, Barlas Oguz, Madian Khabsa, Han Fang, Yashar Mehdad, Sharan Narang, Kshitiz Malik, Angela Fan, Shruti Bhosale, Sergey Edunov, Mike Lewis, Sinong Wang, and Hao Ma. Effective long-context scaling of foundation models. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4643–4663, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.260. URL https://aclanthology.org/2024.naacl-long.260/.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025a. URL https://arxiv.org/abs/2505.09388.

Bowen Yang, Bharat Venkitesh, Dwarak Talupuru, Hangyu Lin, David Cairuz, Phil Blunsom, and Acyr Locatelli. Rope to nope and back again: A new hybrid attention strategy, 2025b. URL https://arxiv.org/abs/2501.18795.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated delta networks: Improving mamba2 with delta rule, 2025c. URL https://arxiv.org/abs/2412.06464.

Tianzhu Ye, Li Dong, Yuqing Xia, Yutao Sun, Yi Zhu, Gao Huang, and Furu Wei. Differential transformer, 2025. URL https://arxiv.org/abs/2410.05258.

Howard Yen, Tianyu Gao, Minmin Hou, Ke Ding, Daniel Fleischer, Peter Izsak, Moshe Wasserblat, and Danqi Chen. HELMET: How to evaluate long-context models effectively and thoroughly. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=293V3bJbmE.

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*, pp. 4791–4800, 2019. doi: 10.18653/v1/P19-1472.

Liang Zhao, Tianwen Wei, Liang Zeng, Cheng Cheng, Liu Yang, Peng Cheng, Lijie Wang, Chenxia Li, Xuejie Wu, Bo Zhu, Yimeng Gan, Rui Hu, Shuicheng Yan, Han Fang, and Yahui Zhou. Longskywork: A training recipe for efficiently extending context length in large language models, 2024. URL `https://arxiv.org/abs/2406.00605`.

Itamar Zimerman and Lior Wolf. On the long range abilities of transformers, 2023. URL `https://arxiv.org/abs/2311.16620`.

A. ArchPool

Tables 1 and 2 describe all 26 models in ArchPool. The architectural columns of the table are identical, and replicated in the second table solely for ease of reading; Table 1 reports HELMET and LongPPL and Table 2 reports RULER scores for all runs. Both tables are sorted in order of HELMET score at 32K, with the best score bolded and the worst score italicized in each column.

The column labeled SWA indicates training with sliding window layers: three out of every four layers use a 4096 local window in lieu of full attention. The QKNorm column indicates the presence of layerwise (✓) or headwise (`hw`) QK norm. The initializations for each run is labeled with a letter code. The number of KV heads indicates the presence and degree of GQA used: 32 KV heads indicates full MHA. The feedforward dimension d_{ff} was adjusted for some runs with GQA to keep the parameter counts as close as possible across training runs.

Architecture									**HELMET**				**LongPPL**
SWA	QKNorm	Norm	fp8	Init	KV	Ctx	d_{ff}	Params	8K	16K	32K	64K	
✓	hw	post	✓	A	8	8K	13,312	7.4B	*42.9*	*34.0*	*29.9*	*27.0*	4.94
✓	✓	post	×	B	32	4K	11,008	7.3B	48.1	41.9	35.0	30.1	5.08
✓	✓	post	×	C	8	8K	13,312	7.4B	45.1	40.1	35.2	32.1	5.05
✓	✓	post	×	H	8	4K	13,312	7.4B	50.0	44.1	36.9	31.8	5.16
✓	✓	post	×	D	8	8K	13,312	7.4B	49.5	44.8	38.8	33.2	4.83
✓	✓	post	×	G	8	4K	14,336	7.8B	49.0	43.0	39.9	34.4	*5.21*
✓	✓	post	✓	D	8	8K	13,312	7.4B	56.4	50.2	42.7	36.1	4.74
✓	✓	post	✓	E	32	8K	11,008	7.3B	54.6	48.7	42.8	37.4	4.31
×	✓	post	×	B	32	4K	11,008	7.3B	53.8	50.9	42.9	32.8	3.83
✓	×	pre	×	F	8	8K	14,336	7.8B	53.7	48.2	43.6	39.2	4.49
✓	×	pre	×	G	8	8K	14,336	7.8B	54.1	49.1	44.8	36.1	4.66
✓	✓	pre	×	H	32	8K	11,008	7.3B	57.2	51.5	46.3	39.1	4.67
✓	✓	post	✓	H	32	8K	11,008	7.3B	55.9	52.8	46.7	39.6	4.23
×	hw	pre	×	K	8	8K	12,288	7.0B	56.9	54.1	46.9	40.5	3.47
✓	hw	post	×	H	32	8K	11,008	7.3B	55.0	48.9	46.9	43.4	4.62
×	✓	post	×	H	32	4K	11,008	7.3B	54.3	52.3	47.4	37.6	3.76
✓	✓	post	×	H	32	8K	11,008	7.3B	58.4	52.5	47.5	41.1	4.38
×	✓	post	×	G	8	8K	14,336	7.8B	58.2	53.0	48.5	39.6	3.85
×	✓	post	×	H	32	8K	11,008	7.3B	57.7	54.3	48.7	39.7	3.70
×	×	pre	×	G	4	8K	14,336	7.7B	55.5	53.8	49.1	43.5	3.63
×	×	pre	×	G	8	4K	14,336	7.8B	58.5	55.2	50.2	42.1	3.78
×	✓	pre	×	G	8	8K	14,336	7.8B	59.4	55.5	51.4	42.4	3.70
×	×	pre	×	G	8	8K	14,336	7.8B	57.8	54.5	52.4	48.4	3.57
✓	×	pre	×	H	32	8K	11,008	7.3B	59.6	57.4	53.7	47.8	4.28
×	×	pre	×	I	32	8K	12,288	7.8B	59.3	55.5	54.4	49.3	**3.41**
×	×	pre	×	J	16	8K	14,336	8.1B	**60.8**	**58.6**	**56.4**	**50.0**	3.50

Table 1. All runs sorted by HELMET 32K (worst to best): HELMET scores and LongPPL.

Architecture									**RULER**				
SWA	QKNorm	Norm	fp8	Init	KV	Ctx	d_{ff}	Params	4K	8K	16K	32K	64K
✓	hw	post	✓	A	8	8K	13,312	7.4B	80.5	68.3	55.4	45.7	38.8
✓	✓	post	×	B	32	4K	11,008	7.3B	81.1	67.0	54.7	45.7	35.2
✓	✓	post	×	C	8	8K	13,312	7.4B	76.3	*60.5*	*51.9*	45.0	39.6
✓	✓	post	×	H	8	4K	13,312	7.4B	78.2	65.3	55.6	*44.4*	*35.0*
✓	✓	post	×	D	8	8K	13,312	7.4B	79.4	64.2	57.0	46.8	38.2
✓	✓	post	×	G	8	4K	14,336	7.8B	*76.2*	63.6	54.9	48.8	40.2
✓	✓	post	✓	D	8	8K	13,312	7.4B	80.3	70.7	61.1	52.2	45.6
✓	✓	post	✓	E	32	8K	11,008	7.3B	81.0	71.0	62.8	54.6	45.8
×	✓	post	×	B	32	4K	11,008	7.3B	83.5	76.6	69.0	57.0	40.9
✓	×	pre	×	F	8	8K	14,336	7.8B	81.6	70.6	63.9	55.3	49.2
✓	×	pre	×	G	8	8K	14,336	7.8B	83.2	73.7	64.9	55.1	47.1
✓	✓	pre	×	H	32	8K	11,008	7.3B	84.6	74.3	64.1	55.8	49.5
✓	✓	post	✓	H	32	8K	11,008	7.3B	79.5	71.8	64.2	56.7	48.9
×	hw	pre	×	K	8	8K	12,288	7.0B	83.9	76.7	69.4	62.9	52.9
✓	hw	post	×	H	32	8K	11,008	7.3B	81.1	68.9	61.8	57.4	55.8
×	✓	post	×	H	32	4K	11,008	7.3B	–	–	–	–	–
✓	✓	post	×	H	32	8K	11,008	7.3B	81.2	70.4	61.2	54.1	48.4
×	✓	post	×	G	8	8K	14,336	7.8B	79.8	74.6	64.6	57.1	44.2
×	✓	post	×	H	32	8K	11,008	7.3B	84.9	75.0	66.8	59.3	50.8
×	×	pre	×	G	4	8K	14,336	7.7B	80.7	73.3	67.5	61.2	54.0
×	×	pre	×	G	8	4K	14,336	7.8B	81.1	67.0	54.7	45.7	35.2
×	✓	pre	×	G	8	8K	14,336	7.8B	82.5	75.6	68.3	62.8	50.7
×	×	pre	×	G	8	8K	14,336	7.8B	83.2	78.1	72.0	64.6	56.0
✓	×	pre	×	H	32	8K	11,008	7.3B	82.8	74.6	66.2	64.2	57.3
×	×	pre	×	I	32	8K	12,288	7.8B	85.2	**81.6**	75.5	67.5	58.9
×	×	pre	×	J	16	8K	14,336	8.1B	**86.0**	79.9	**76.0**	**67.7**	**64.9**

Table 2. All runs sorted by HELMET 32K (worst to best): RULER scores.

B. Additional features studied

Norm ordering and type of QK norm have small effects. We ablate the three combinations of QK norm and norm order that are stable in our training regime: preorder without QK norm, preorder with QK norm, and post-sublayer order with QK norm in Figure 7. We find that norm order alone has an inconsistent effect, with QK norm accounting for the vast majority of the performance difference between runs. Headwise QK norm causes an additional slight degradation. QK norm is most often added to improve training stability (e.g. Chameleon Team (2025); Marin Community (2025)); we also confirm that it improves stability in our setting in Appendix D.3.

Float8 pretraining does not appear to have a consistent effect. We consider training linear layers in float8 instead of bfloat16 precision. In two controlled comparisons, adding float8 training results in a slight degradation in one comparison and an improvement in the other. We find no evidence that float8 pretraining degrades long context performance—instead, it appears that float8's effect in ArchPool models may be purely noise, like initialization, because float8 optimization results in a slightly different optimization path. We note that we do not consider a setting where attention weights are quantized during pretraining; it is possible that this would result in degradation, as long context abilities inherently require the model to be able to attend precisely over a long context window.

Initialization causes more variation in short context than long context. We are not able to completely standardize initialization across runs because some models differ in parameterization. In ArchPool, we construct four pairs of runs that are identical except for initialization and measure the swing in performance for both short context and long context metrics. To standardize across metrics with differing scales, we compute the swing due to initialization as a percentage of the maximum variation between runs in ArchPool. Initialization has minimal impact on train loss and perplexity (shifting runs by less than 3% of the cross-run range on average) and causes the largest swings in the short context benchmarks and pre-context-extension HELMET scores, where runs with the same architecture but different initialization may vary by up to 58% of the total range of variance. This effect is reduced after long context extension; on the long context benchmarks, intialization can account for smaller but still substantial swings of up to 17% (and on average 7.7%) of the observed range of values. In the results, we discuss only differences across architectural variations that result in long context score shifts that are substantially larger than the mean difference attributable to initialization. Note that the sole dimension where we cannot construct an initalization-controlled trial is in measuring the impact of varying GQA degree; thus, the effect size for the difference there may be slightly larger or smaller than the one we observe.

C. Additional Training Details

For additional details on training and the pretraining data used, we refer the reader to the Olmo 3 technical report (Olmo Team et al., 2025); we follow the configuration for Olmo 3 7B Base pretraining, although we use substantially less than 1024 concurrent GPUs for these much shorter runs. The project github repositiory provides configurations for each pretraining and context extension run, and is the best reference point for specific training details for each run.

One important detail that we highlight here: like all recent Olmo models, ArchPool models are trained with a skip-step optimizer, which skips steps that have an abnormally high gradient norm to increase run stability. This triggers very rarely, but when it does occur, it causes a small amount of variance in pretraining data across ArchPool (i.e., for a small percentage of data, not every model performs a gradient update). For more details, see the documentation.

D. Further evaluation details

This appendix briefly describes each short-context evaluation metric and validation perplexity set and its correlation with long context downstream.

D.1. Loss

In Figure 8, we graph the (slight) correlation between loss and long context ability. Both pretraining and long context loss fail to correlate with long context capabilities.

D.2. Perplexity on held-out text

We measure perplexity across a diverse set of held-out validation splits, drawn mostly from Paloma (Magnusson et al., 2024). **C4** (Raffel et al., 2020) is a filtered and deduplicated web corpus derived from Common Crawl; we use the English validation split. **Dolma** (Soldaini et al., 2024) is Ai2's open pretraining corpus; we evaluate on six domain-specific splits: books, Common Crawl, peS2o (scientific papers), Reddit, Stack Exchange, and Wikipedia. **ICE** (Greenbaum & Nelson, 1996) (International Corpus of English) provides text spanning multiple varieties of dialectal English. **M2D2** (Reid et al., 2022) is a massively multi-domain dataset; we use the S2ORC split covering scientific text. **The Pile** (Gao et al., 2020) is an 800GB diverse text corpus from EleutherAI, evaluated on its held-out validation set. **WikiText-103** (Merity et al., 2016) is a standard language modeling benchmark of Wikipedia articles.

Figure 9 shows the per-split correlation with the downstream

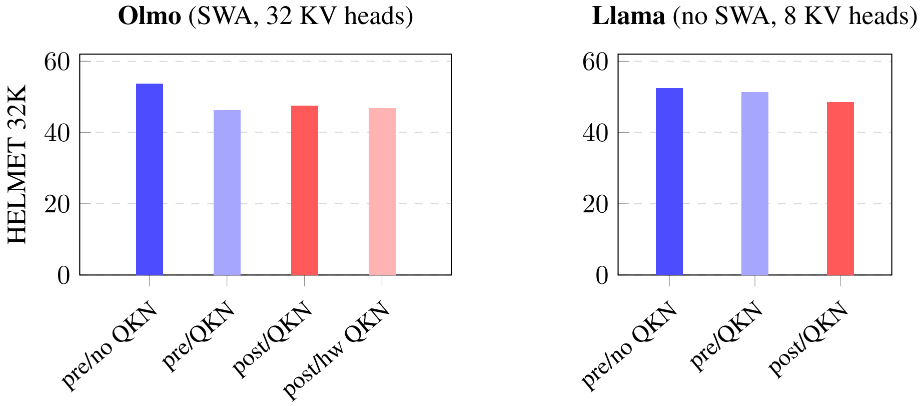

Figure 7. QK norm is harmful for long context (as first observed by Yang et al. (2025b)); we further note that the headwise variant is an additional slight detriment to downstream performance, and the less harmful norm order with QK norm is model-family-dependent.

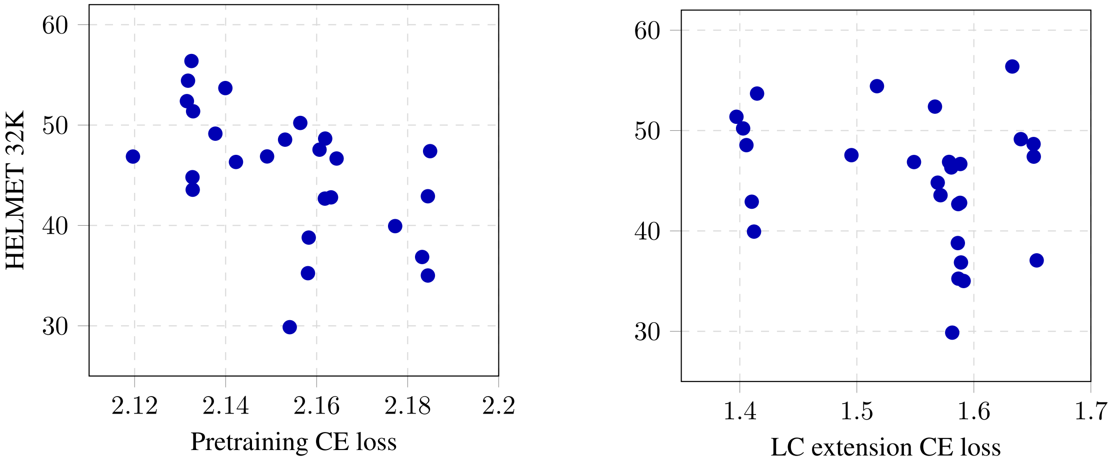

Figure 8. Training loss does not correlate well with long context ability, in either the pretraining or long context extension runs.

long context metrics.

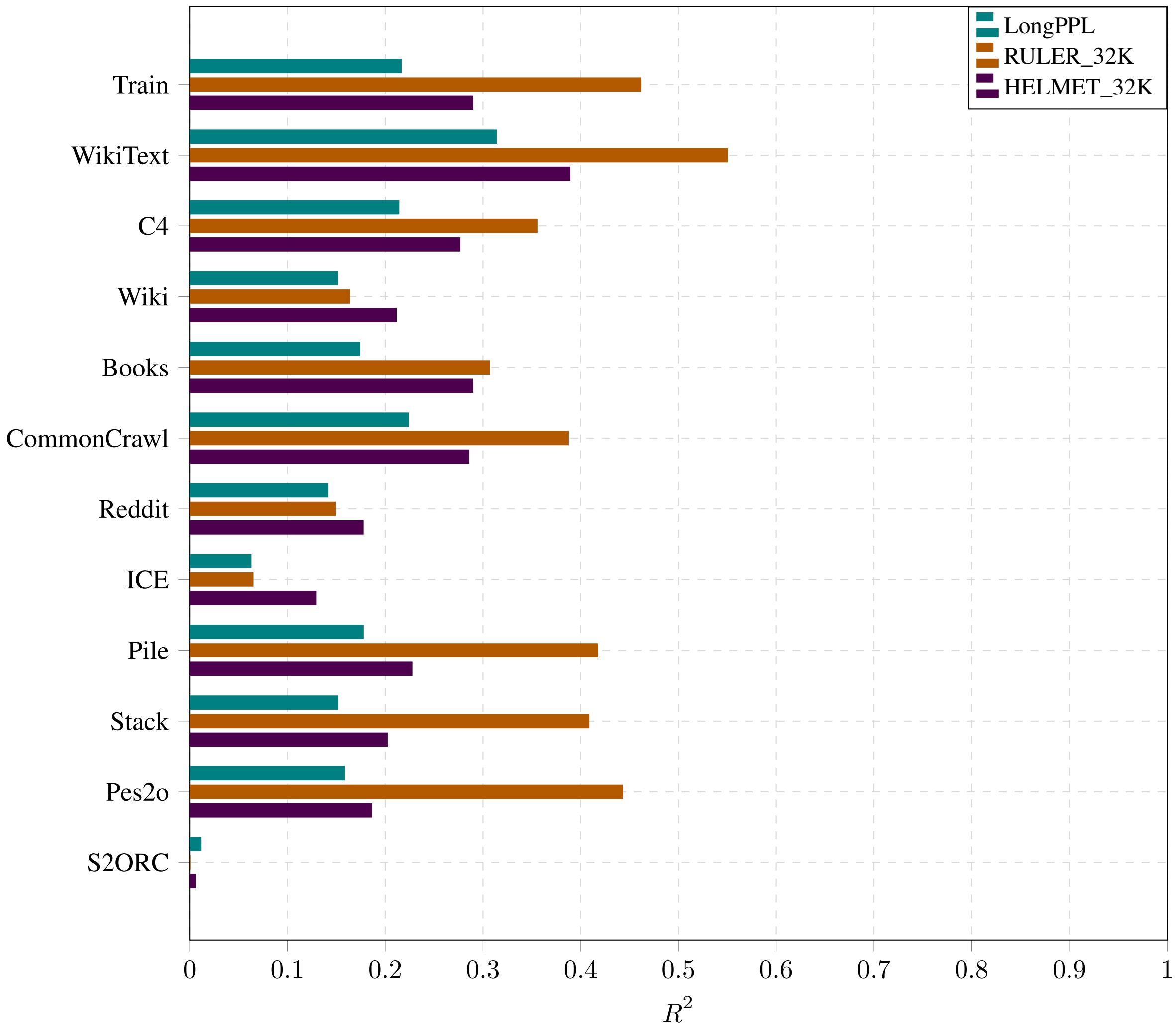

Figure 9. R^2 between each standard validation perplexity metric and downstream long-context metrics (HELMET 32K, RULER 32K, LongPPL).

D.3. BPB on downstream benchmarks

We evaluate short-context capabilities across several standard benchmarks. **ARC** (Easy and Challenge splits) (Clark et al., 2018) is grade-school science question answering, with the Challenge split requiring more complex reasoning. **HellaSwag** (Zellers et al., 2019) measures commonsense natural language inference via sentence completion. **MMLU** (Hendrycks et al., 2021) measures graduate-level knowledge across 57 subjects; we use the general split into humanities, social sciences, STEM, and other domains . **HumanEval** (Chen et al., 2021) requires Python code generation conditioned on docstrings. **MBPP** (Austin et al., 2021) similarly evaluates code generation on entry-level Python programming problems. **MINERVA** (Lewkowycz et al., 2022) is a set of 500 math problems requiring multi-step numerical and symbolic computation. Finally, **Basic Skills** (Olmo Team et al., 2025) is a datatset covering six fundamental competencies: arithmetic, coding, common knowledge, logical reasoning, pattern recognition, and string operations. All downstream metrics are reported as bits-per-byte (BPB) on the gold answer string.

Figure 10 shows the correlation between each individual short context benchmark BPB and the downstream long context scores. No benchmark correlates strongly, and we do not observe any consistent trend in which types of benchmarks correlate most. Note one of the three coding-related benchmarks correlates strongly, while the others do not.

D.4. Training stability

Another important feature in pretraining is the stability of the training process. We measure a score for stability by computing the percentage of gradient norms that are more than 6 standard deviations from the mean during the pretraining run. We observe that runs with QK norm have, on average, a lower spike score, which aligns with prior observations that QK norm is beneficial for stability.

We then calculate correlation between this score and long context performance downstream and visualize this in Figure 11. In ArchPool, there is a slight negative correlation ($R^2 = 0.22$) between pretraining stability and long context performance, mostly because QK norm both improves stability and damages long context performance.

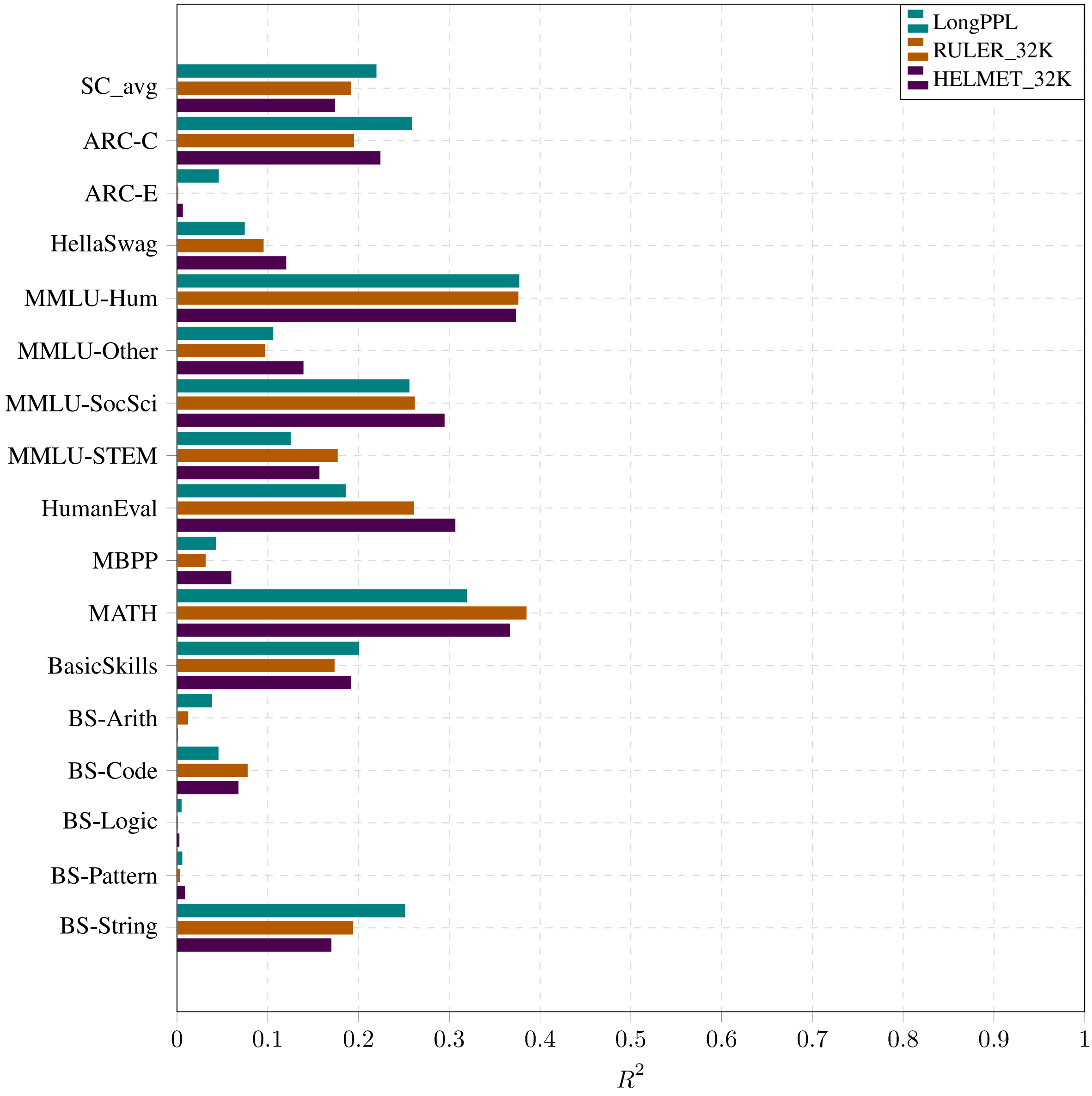

Figure 10. R^2 between each short-context benchmark metric and downstream long-context metrics (HELMET 32K, RULER 32K, LongPPL).

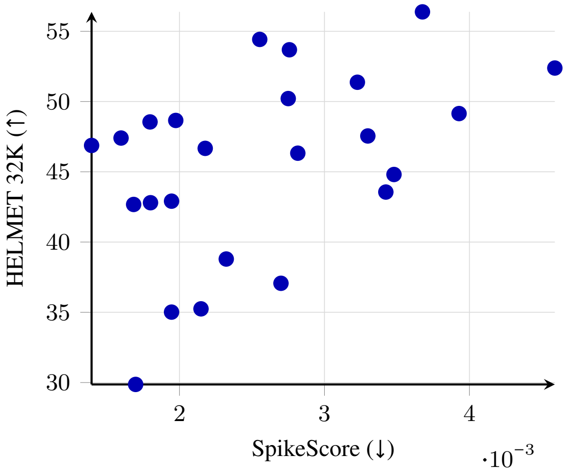

Figure 11. Spike score vs. HELMET 32K across ArchPool. Stability weakly correlates with worse LC performance downstream, mostly due to QK-norm's influence on both factors.