When Does a Small Language Model Need QK-Norm? An Isolated Learning-Rate Characterization of Query–Key Normalization at Sub-100M Parameters

(working title – rough draft)

Cormac Shannon
MSc Immersive Software Engineering
University of Limerick
Limerick, Ireland
23366303

Abstract—Modern language-model architectures bundle many components validated only at large scale; whether they help small models is under-studied and, for some components, actively disputed. This paper motivates a controlled, isolated study of one such component – Query–Key Normalization (QK-norm) – at sub-100M-parameter scale, characterizing when (at which learning rates and depths) it becomes necessary.

Index Terms—small language models, transformer architecture, training stability, query–key normalization, learning rate, ablation study

I. Introduction

- Broad domain: transformer language models now dominate NLP/ML; but the field is increasingly moving toward small / efficient models (on-device, edge, cost, energy, and reproducibility).
- Modern LLM architectures are an accretion of components – rotary position embeddings, RMSNorm, grouped-query attention, SwiGLU, QK-norm – almost all validated at the billion-parameter scale [2], [5].
- Hook / tension: it is not established that these components carry down to small models. Design rules for the sub-100M regime are thin, and for several components the evidence is contradictory (see below).
- Why now: small-model pretraining is a live research area (sample-efficiency work [6]; compute-constrained ablations [7]); getting the component choices right matters most exactly where the compute budget is smallest.
- What QK-norm is + why it exists.
  - Origin: Henry et al. [1] – normalize the query and key vectors before the attention dot product (for low-resource NMT).
  - Mechanism/problem it solves: without it, query/key norms grow during training → large attention logits → softmax saturates (near one-hot) → unstable/diverging loss, forcing smaller learning rates.
  - Deeper mechanism (why logits grow): "massive activations" – a few tokens acquire extreme-magnitude activations that become dominant key vectors and form "attention sinks" [10], [11]; QK-norm bounds Q/K magnitude, limiting this.
  - Popularized at scale for stability: ViT-22B [2].
  - Complementary mechanistic account: "attention entropy collapse" – low attention entropy co-occurs with instability [3]. (QK-norm bounds logits → keeps entropy up.)
- QK-norm is now standard in frontier models.
  - OLMo 2 applies RMSNorm to Q and K and reports improved training loss + stability [5]. Also adopted by Gemma 2/3, Qwen3, Chameleon.
- QK-norm is one of several logit-control methods (matters for the gap).
  - Alternatives/complements: attention logit soft-capping (Gemma 2, $z \leftarrow c\tanh(z/c)$) [12]; z-loss, which keeps the softmax normalizer near zero [13]; NormFormer's extra normalization [14].
  - These are almost always studied bundled together, so QK-norm's isolated effect is rarely attributed.
- Evidence at / near small scale (the closest work).
  - Wortsman et al. [4] – instabilities reproduce in small models at high LR; qk-layernorm (+ z-loss) mitigate them and "transfer down"; introduces

the LR-sensitivity metric. KEY methodological template for us.
  - IMU-1 (Grigorev et al.) [6] – at 70M, QK-norm is the single largest architectural win (measured via reduced activation kurtosis) – but bundled with μP, NorMuon, value residuals, etc.
- The countervailing line (the contradiction).
  - Compute-constrained small-model ablations report that modern components (RoPE, RMSNorm, GQA) give "no measurable benefit at ∼70M" [7]. Directly disputes the above.
- Why learning rate is the right axis.
  - μP / μTransfer [8]: optimal LR transfers across width; LR is the hyperparameter these stability effects hinge on. Motivates studying QK-norm as a function of LR, and switching μP OFF so LR effects are not masked.
  - Recent work keeps reaffirming the LR link – e.g. QK-Normed MLA reports QK-norm "permits significant increases in learning rate" [15].
- The positive small-scale evidence is confounded: Wortsman bundles qk-norm with z-loss and frames small models as a proxy for large-scale behaviour; IMU-1 bundles it with μP + NorMuon + value residuals and measures it via activation statistics, not an LR sweep, at a single (fixed) depth.
- The negative evidence [7] tests a single compute-optimal budget, does not isolate QK-norm, and does not sweep LR.
- So the two literatures disagree at the same scale while neither does the clean experiment.
- GAP: no dense, isolated (μP-off) learning-rate-sensitivity characterization of QK-norm at 10–100M, and no map of how the LR at which QK-norm becomes necessary moves with model depth.
- Secondary gap: reproducibility – most evidence is expensive / CUDA-only; a laptop-reproducible study is itself a contribution.
- Objective: cleanly isolate QK-norm and chart when a small model needs it.
- Research question: at <100M params, does isolated QK-norm improve convergence and raise the maximum stable LR – and at what LR does it become necessary? Sub-question: how does that threshold shift with depth?
- Hypotheses (state as such):
  - H1 – effect is LR-conditional: negligible/slightly negative at low LR, decisive at high LR.
  - H2 – QK-norm raises max stable LR and lowers LR-sensitivity.
  - H3 – the "necessity" LR falls as depth increases.
- High-level method: controlled ablation on plain nanoGPT [9], μP off, AdamW; single independent variable QK-norm{on,off} × dense LR sweep × depth; ∼10–15M params; measures = LR-sensitivity [4], max-stable-LR, steps-to-threshold, perplexity, plus attention entropy / activation kurtosis as mechanistic checks.
- Data + precedent: small models learn coherent language from tiny corpora (TinyStories) [16]; single-GPU near-BERT training is an established accessible-research precedent (Cramming) [17].
- Feasibility/repro: runs on a consumer laptop (Apple M4 Pro); code, seeds, logs released.
- "The remainder of this paper is structured as follows: Section II reviews related work on QK-norm and small-model training stability; Section III details the experimental design; Section IV reports results; and Section V concludes."

References

[1] A. Henry, P. R. Dahl, et al., "Query-key normalization for transformers," in Findings of EMNLP, 2020. arXiv:2010.04245.
[2] M. Dehghani et al., "Scaling vision transformers to 22 billion parameters," 2023. arXiv:2302.05442.
[3] S. Zhai et al., "Stabilizing transformer training by preventing attention entropy collapse," in ICML, 2023. arXiv:2303.06296.
[4] M. Wortsman et al., "Small-scale proxies for large-scale transformer training instabilities," in ICLR (oral), 2024. arXiv:2309.14322.
[5] OLMo 2 Team, "2 OLMo 2 furious," 2025. arXiv:2501.00656.
[6] G. Grigorev et al., "IMU-1: Sample-efficient pre-training of small language models," 2026. arXiv:2602.02522.
[7] "Architectural trade-offs in small language models under compute constraints," 2025. arXiv:2512.20877.
[8] G. Yang et al., "Tensor programs V: Tuning large neural networks via zero-shot hyperparameter transfer," 2022. arXiv:2203.03466.
[9] A. Karpathy, "nanoGPT," GitHub repository, 2022–. https://github.com/karpathy/nanoGPT.
[10] M. Sun et al., "Massive activations in large language models," in ICML, 2024. arXiv:2402.17762.
[11] "When attention sink emerges in language models: an empirical view," 2024. arXiv:2410.10781.
[12] Gemma Team, "Gemma 2," technical report, 2024.
[13] A. Chowdhery et al., "PaLM: Scaling language modeling with pathways," 2022.
[14] S. Shleifer et al., "NormFormer: Improved transformer pretraining with extra normalization," 2021. arXiv:2110.09456.
[15] "QK-Normed MLA: QK normalization without full key caching," 2026. arXiv:2606.16310.
[16] R. Eldan and Y. Li, "TinyStories: How small can language models be and still speak coherent English?," 2023. arXiv:2305.07759.
[17] J. Geiping and T. Goldstein, "Cramming: Training a language model on a single GPU in one day," in ICML, 2023. arXiv:2212.14034.