



Fractional Rotation, Full Potential?

Investigating Performance and Convergence of Partial RoPE

Mohammad Aflah Khan, Krishna P. Gummadi, Manish Gupta, Abhilasha Ravichander
afkhan@mpi-sws.org

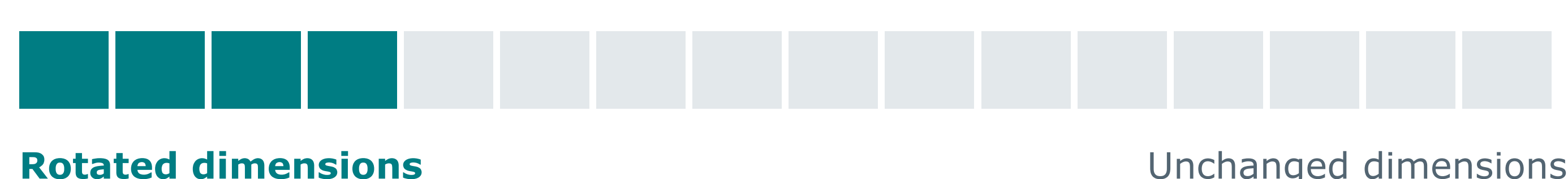
$\approx 10\%$
of query/key dimensions

**Comparable convergence
to full RoPE**

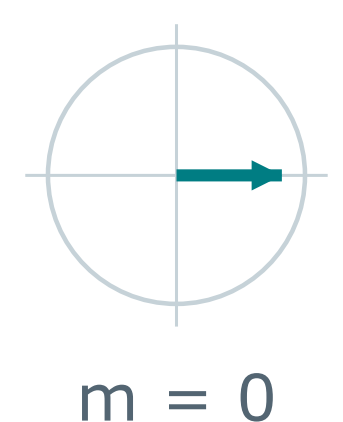
Upto 10x
smaller RoPE cache

What is Partial RoPE?

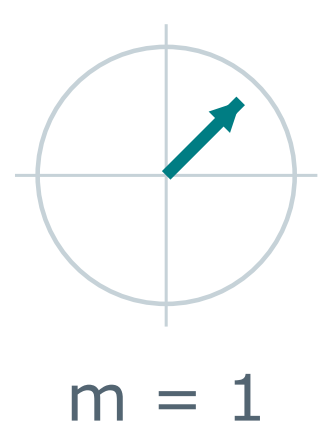
Rotate a fraction of the dimensions in each query/key head. Leave the rest unchanged.



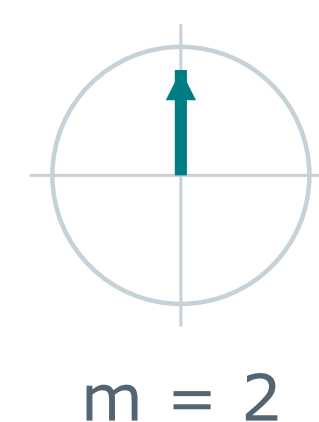
Each selected pair rotates with token position



$m = 0$



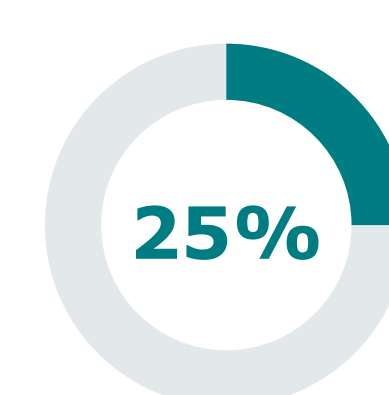
$m = 1$



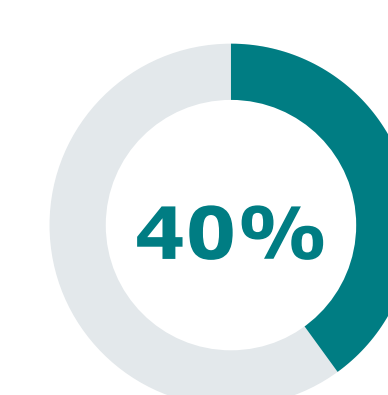
$m = 2$

Existing models vary widely

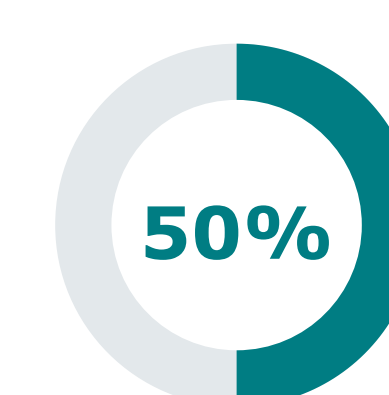
Fraction of query/key dimensions within each head



GPT-J, GPT-NeoX
Pythia,
Qwen3-Next



Phi-2



Nemotron-4
340B, GLM-4.5



Llama,
most Qwen
models

What fraction is enough for stable, effective pretraining?

1B and 8B

Llama-style / Pythia
style architectures

100B tokens

FineWeb / FineWeb-Edu

1024–8192

Training sequence length

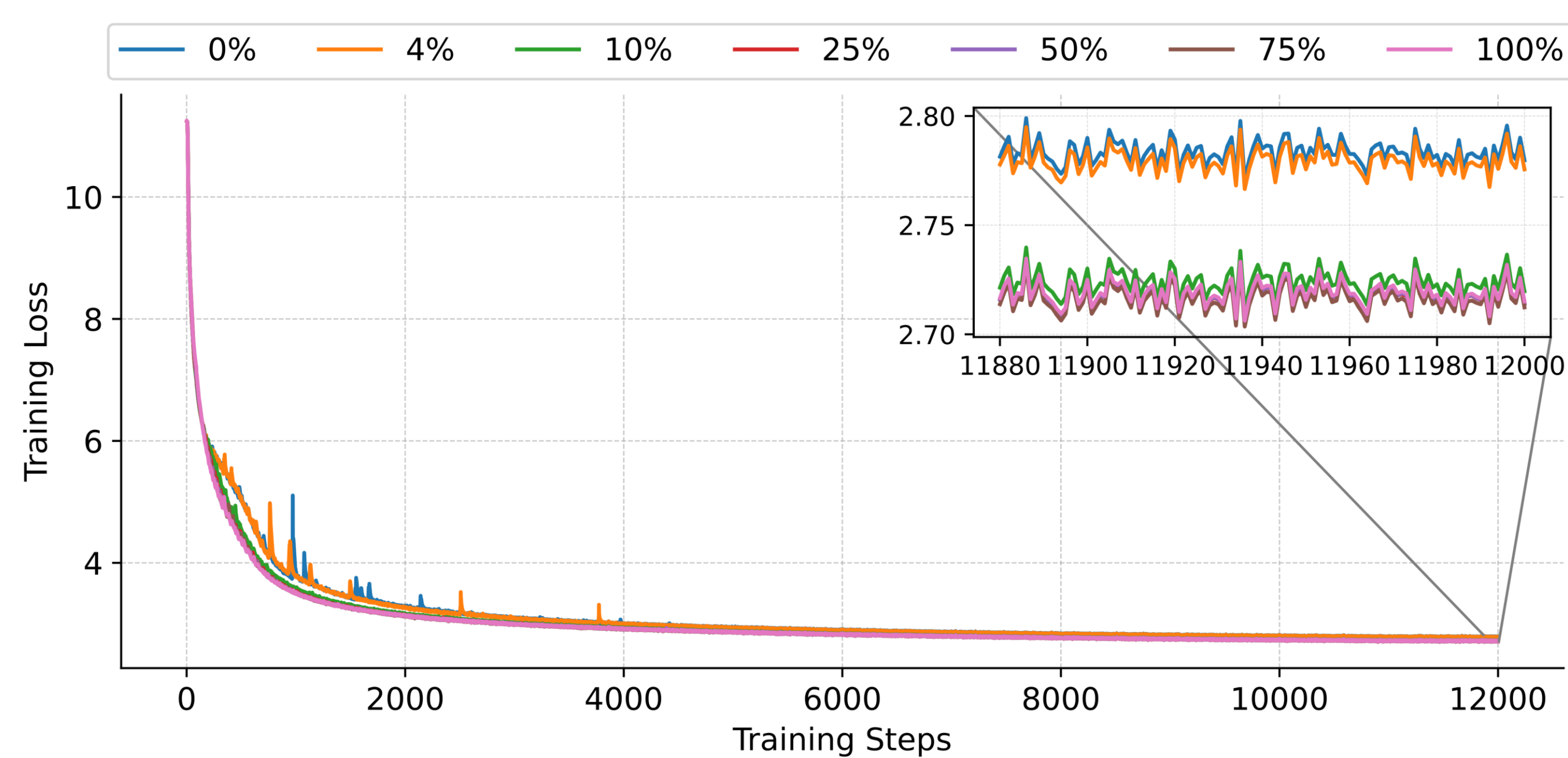
0–100%

Fraction of dimensions rotated

About 10% reaches a similar lower-loss band

We pretrain **55 models from scratch** across varying configurations -

1B model, FineWeb, 2048 tokens

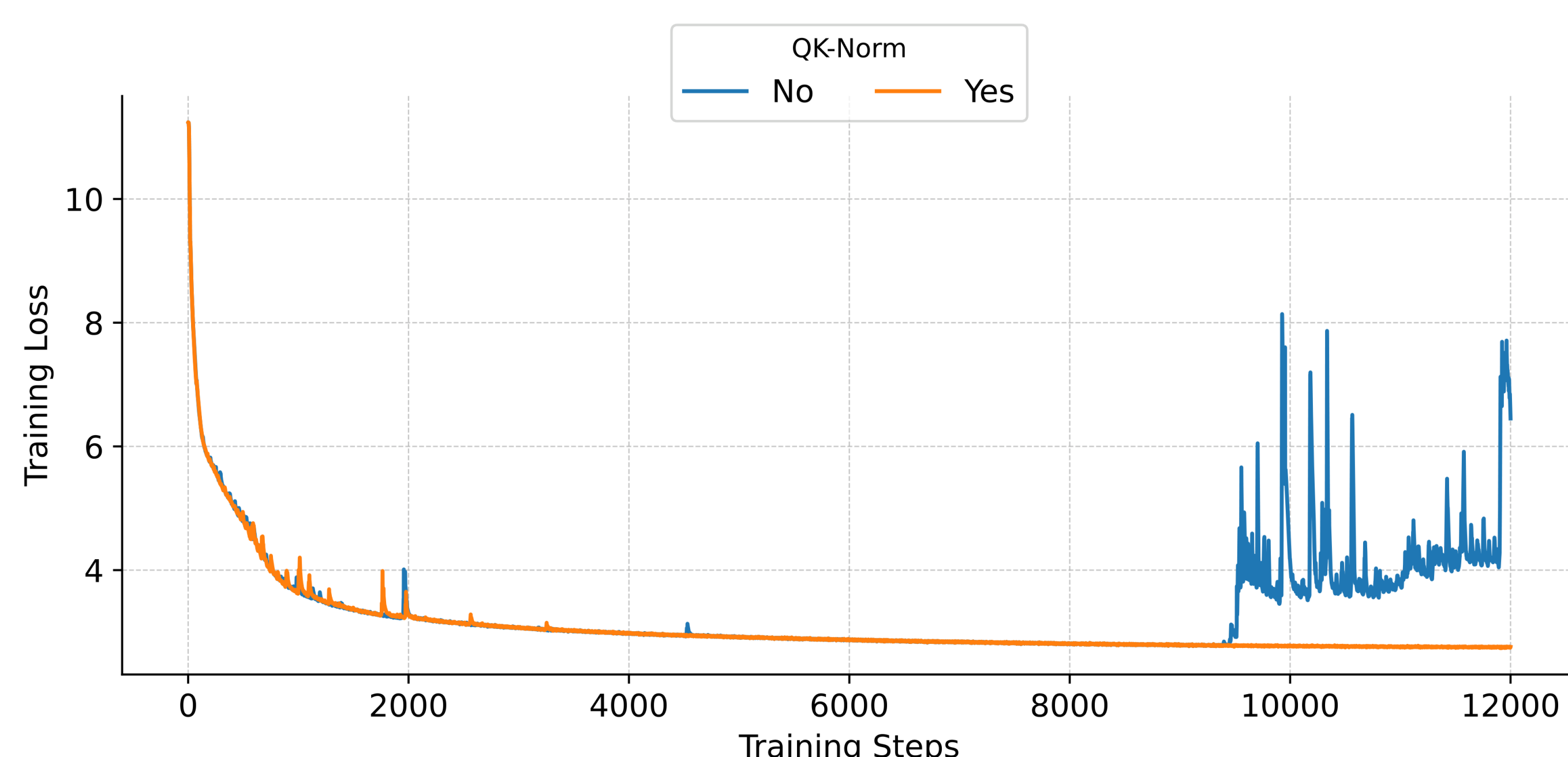


NoPE and 4% finish at a higher loss.
10–100% cluster in the lower band.

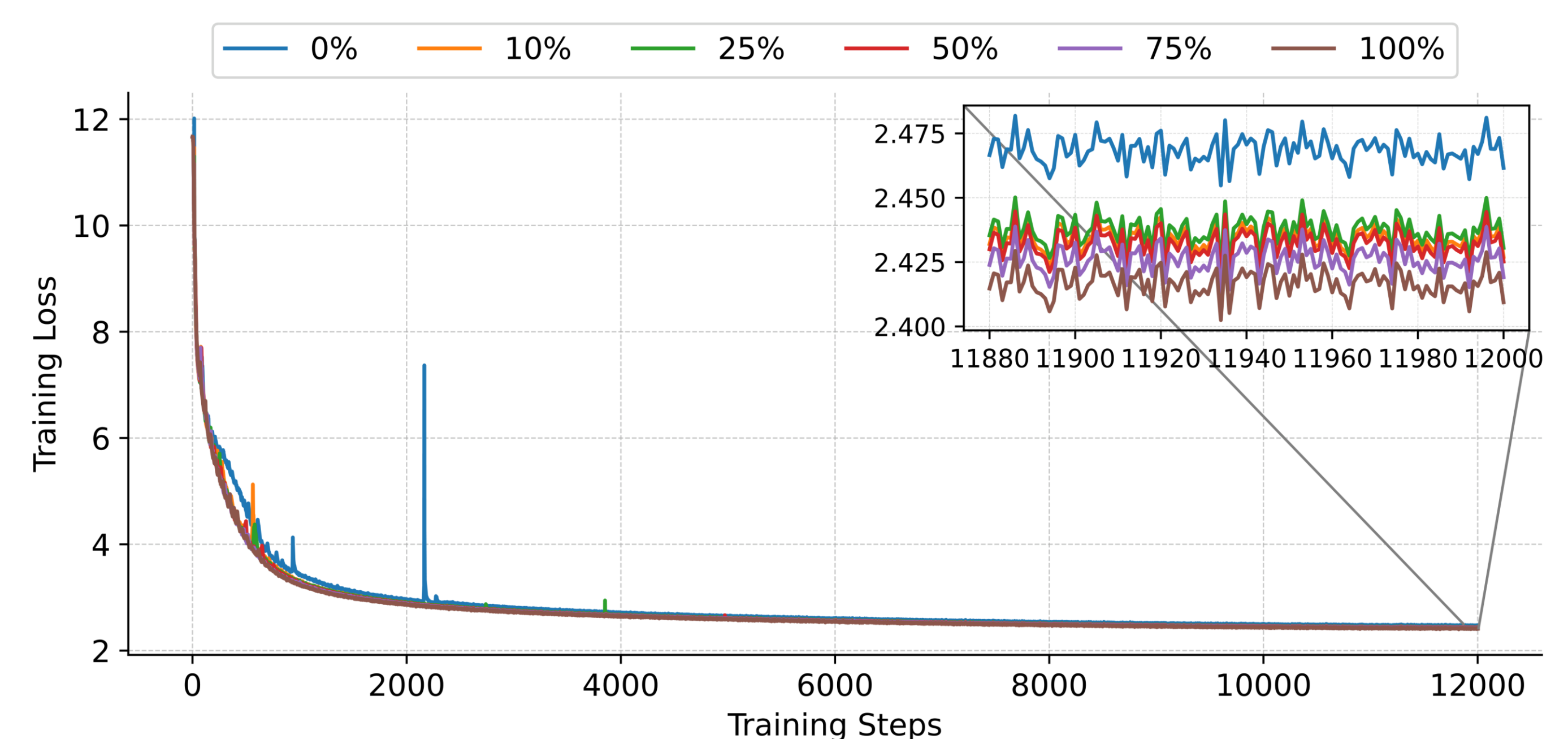
NoPE can destabilize training

Parallel 1B, FineWeb-Edu, 2048 tokens.

QK-Norm removes spikes, but loss remains higher.



8B model, FineWeb, 2048 tokens



The pattern persists at 8B parameters.

Small differences remain among RoPE runs.

Perplexity evals. further support

LAMBADA (OpenAI), lower is better

1B model, FineWeb, 2048 tokens

RoPE fraction	Perplexity
0% (NoPE)	14.90 ± 0.46
10%	12.86 ± 0.39
100%	12.78 ± 0.39

Most MCQ benchmarks show similar accuracy.

Inline with prior work which finds perplexity is a more sensitive metric

