WHEN DOES A SMALL MODEL NEED QK-NORM?

Presentation subtitle

THE PROBLEM & TOPIC

1. Modern LLMs stack many components: *RoPE, RMSNorm, GQA, SwiGLU,* ***QK-norm***

2. **What's already known:** QK-norm improves training stability, *especially at high learning rates*

3. **What's still fuzzy, my focus:** the *qualitative* rule is known; a **clean, dense measurement is not**. Small-scale evidence is **patchy and inconsistently measured**

SIGNIFICANCE

"SO WHAT?"

- **The field is going small**
- **A usable rule of thumb:**
  "below this LR you can skip it; above it you can't."
- **Reproducibility:** a **laptop-runnable, open** artifact adds trustworthy, re-runnable evidence to a literature dominated by expensive, hard-to-reproduce large-scale results.

CONCEPTUAL FRAMEWORK QUESTIONS & OBJECTIVES

Primary RQ: At <100M parameters, does *isolated* QK-norm's stability benefit reproduce under a **dense learning-rate sweep**, and **at what learning rate does it become necessary?**
Sub-RQ: How does that **necessity threshold shift with model depth?**

Hypotheses:

1. **H1:** QK-norm's effect is **LR-conditional**: negligible/slightly negative at low LR, decisive at high LR.
2. **H2:** QK-norm **raises the max stable LR** and **lowers LR-sensitivity** at <100M.
3. **H3:** the LR at which QK-norm becomes necessary **falls as depth increases** (deeper models need it sooner).

METHODOLOGICAL APPROACH

Quantitative, experimental: a controlled ablation

Independent variable	QK-norm **(on/off)**
Moderators	**Learning rate** (dense, log-spaced, safe to divergent), **depth** (2–3 values)
Controls	Identical width, data, tokens, optimizer, fixed seeds
Model	Plain nanoGPT decoder, ~10-15M params
Measures	**max stable LR, LR-sensitivity**, steps-to-threshold, perplexity, attention entropy

FEASIBILITY & TIMELINE

- **Runs on my hardware:** M4 Pro (48 GB)
- **Stretch** (the depth boundary): add 2 more depths near the threshold
- **Not** GPT-2 scale. An ablation needs *relative* effects, so tiny models are correct choice here, a larger model would be more costly financially and timewise.

ETHICS & DATA ACCESS

- **No human subjects, no personal or sensitive data:** no ethical-approval barriers.
- **Public, openly-licensed datasets:** (TinyStories, FineWeb-Edu), licences to be confirmed and respected, **no data-access constraints.**
- **Low environmental footprint:** a few GPU-hours total, energy-efficiency is itself a study motivation.
- **Reproducibility & integrity:** code, configs, seeds, and logs released openly. AI-tool use declared in the research notebook per module policy.

SUMMARY

- **Not "does QK-norm help" (known), but *when a small model needs it*, charted cleanly.**
- An honest **verification-plus-boundary-charting** study: firms up patchy small-scale evidence and hands practitioners a usable rule.