MSc Research Notebook

- **Student Name:** Cormac Shannon
- **Student ID:** 23366303
- **Proposed Research Title:** When Does a Small Language Model Need QK-Norm? An Isolated Learning-Rate Characterization of Query–Key Normalization at <100M Parameters *(working title)*
- **GitHub Repository Link:** https://github.com/NikScorch/MSc_Research_Project

---

1. Project Overview & Evolution Track

Use this section to map out how your project evolves over the block. Whenever you change your research question, methodology, or scope, update this log.

Current Working Research Question(s)

> *Last Updated: 21/08/2026*

Primary research question:

> *"At <100M parameters, does isolated Query–Key Normalization (QK-norm) improve training convergence and raise the maximum stable learning rate - and at what learning rate does it become necessary?"*

Sub-question: How does that necessity threshold shift with model depth?

Hypotheses:

- **H1** - QK-norm's effect is learning-rate-conditional: negligible or slightly negative at low LR, decisive at high LR.
- **H2** - QK-norm raises the maximum stable LR and lowers LR-sensitivity at <100M.
- **H3** - the LR at which QK-norm becomes necessary falls as model depth increases.

Honest framing (important): This is a *verification + boundary-charting* study, **not** a novel discovery. QK-norm's stabilising effect is already established (Wortsman, ICLR 2024; IMU-1, 2026). What is genuinely un-charted is a **dense**, **isolated** (μP-free) learning-rate-sensitivity curve at 10–100M, plus how the "necessity" threshold moves with depth. I am not claiming novelty of technique or "first at small scale" - see the Critical Source Evaluation Log.

Feasibility anchor: Small models only (~10–15M params). Controlled ablation on an M4 Pro (48 GB, PyTorch MPS): ~1–1.5 h/run (to validate with a Week-5 pilot); core matrix QK-norm{on,off} × LR{~6} × seed{2} ≈ 24 runs. Measures: max-stable-LR, LR-sensitivity (Wortsman metric), steps-to-threshold, perplexity, plus attention entropy and activation kurtosis as mechanistic checks.

Experimental platform: plain nanoGPT (https://github.com/karpathy/nanoGPT), with **μP switched off** and standard AdamW, so QK-norm is the *only* variable. (Karpathy's *nanochat* was trialled in Week 2 but bundles speedrun tricks/optimisers that would confound an isolated QK-norm study - see Pivot Log.)

The Pivot Log

Document any major changes in direction here. (e.g., *"Switched from quantitative survey to qualitative interviews because sample access fell through on 15/10/2026"*).

Date	Previous Direction	New Direction	Reason for Pivot / Academic Justification
08/07/2026	"Study architectural changes to	Small-model (~10–100M param) controlled single-lever ablations	Training frontier LLMs is infeasible within an MSc compute/time budget. Small-model ablations preserve the research question while remaining

Date	Previous Direction	New Direction	Reason for Pivot / Academic Justification
	LLMs" (unbounded scope)		reproducible - consistent with Creswell (Ch.1) criteria for selecting an approach and Vogt (2008) on letting the problem dictate feasible methods.
17/07/2026	Small-model single-lever ablation (lever TBD)	Focused specifically on Query–Key Normalization (QK-norm) × learning rate	After a literature scan of candidate levers, QK-norm chosen as a cheap, ubiquitous, well-scoped component with a clean controlled design that is feasible on an M4 Pro - unlike over-training or nanoGPT-speedrun-stack studies, which need cloud GPUs or CUDA-only tooling.
24/07/2026	Framed as a "novel study of QK-norm at small scale"	Reframed as a verification + boundary-charting study; base switched from nanochat to plain nanoGPT (μP off)	Prior-art checks showed QK-norm's benefit is already established (Wortsman 2024; IMU-1 2026), so the claim was honestly rescoped to a dense, isolated, laptop-reproducible LR-sensitivity curve + depth threshold. Plain nanoGPT with μP off isolates the single variable; nanochat/modded-nanoGPT bundle confounds.

---

2. Critical Source Evaluation Log

Do not just copy and paste citations. For every core paper that shapes your final Introduction, document your active reading process here. Also note papers that looked promising but you decided, on review, to exclude them. Note reasons for exclusion.

Xiao & Watson (2019) - Guidance on Conducting a Systematic Literature Review

- **Full IEEE/APA Citation:** Xiao, Y., & Watson, M. (2019). Guidance on conducting a systematic literature review. *Journal of Planning Education and Research*, 39(1), 93–112. *[verify volume/issue/pages against the PDF before final use]*
- **Why did you select this paper?** It provides a step-by-step, repeatable process for running a literature review, which I need to structure the search that will feed my Introduction and thesis.
- **Key Findings & Methodology:** Lays out the SLR stages (formulate the problem → develop/validate the protocol → search the literature → screen for inclusion → assess quality → extract/analyse data → report findings). Emphasises transparency and reproducibility of the search.
- **Limitations/Weaknesses Identified:** Oriented toward planning/social-science literature; the database and screening advice needs adapting to CS venues (e.g. arXiv, ACL Anthology, IEEE Xplore, NeurIPS/ICML proceedings).
- **Direct Connection to my Project:** Gives me a defensible protocol for finding and filtering the architecture/ablation papers, so my literature map (per Creswell Ch.2) is systematic rather than ad hoc.

Wortsman et al. (2024) - Small-scale proxies for large-scale Transformer training instabilities

- **Full IEEE/APA Citation:** Wortsman, M., et al. (2024). Small-scale proxies for large-scale Transformer training instabilities. *ICLR 2024 (oral)*. arXiv:2309.14322. *(arXiv ID + venue confirmed; verify full author list before final use.)*
- **Why did you select this paper?** The closest methodological template: it reproduces training instabilities (incl. attention-logit growth) in small models and introduces the *LR-sensitivity* metric I will use.
- **Key Findings & Methodology:** qk-layernorm (+ z-loss) mitigate high-LR instabilities and transfer down to small models; dense learning-rate sweeps at small scale, confirmed with one large run.
- **Limitations/Weaknesses Identified:** qk-norm is studied *bundled* with other fixes and framed as a *proxy* for large-scale behaviour, not isolated at <100M.

- **Direct Connection to *Your* Project:** Defines both my method (LR-sensitivity, dense LR sweeps) and my niche (isolate QK-norm, μP off, dense curve at 10–100M).

IMU-1 (2026) - Sample-Efficient Pre-training of Small Language Models

- **Full IEEE/APA Citation:** Grigorev, G., et al. (2026). IMU-1: Sample-Efficient Pre-training of Small Language Models. arXiv:2602.02522. *(arXiv ID, lead author, and "QK-norm = largest single architectural win" claim confirmed; preprint - verify any published version.)*
- **Why did you select this paper?** Closest subject-matter prior art: its 70M ablations find QK-norm the single largest architectural win.
- **Key Findings & Methodology:** At 70M, QK-norm reduces extreme activation kurtosis and gives the biggest single-intervention loss improvement - but bundled with μP, NorMuon, value residuals, etc., and measured via activation statistics, *not* an LR sweep.
- **Limitations/Weaknesses Identified:** Fixed depth; QK-norm's LR effect entangled with μP; no dense LR sweep.
- **Direct Connection to *Your* Project:** The paper I am disentangling - isolating QK-norm (μP off) and adding the LR sweep it did not do.

Henry et al. (2020) - Query-Key Normalization for Transformers

- **Full IEEE/APA Citation:** Henry, A., et al. (2020). Query-Key Normalization for Transformers. *EMNLP 2020 Findings*. arXiv:2010.04245. *[verify]*
- **Why did you select this paper?** Origin and definition of QK-norm (on small transformers, for low-resource machine translation).
- **Key Findings & Methodology:** ℓ2-normalise queries and keys before the dot product; a learnable scale replaces 1/√d; fixes softmax saturation.
- **Limitations/Weaknesses Identified:** Machine translation, not language modelling; small models but a different task.
- **Direct Connection to *Your* Project:** Defines the exact technique under study.

Dehghani et al. (2023) - Scaling Vision Transformers to 22 Billion Parameters (ViT-22B)

- **Full IEEE/APA Citation:** Dehghani, M., et al. (2023). Scaling Vision Transformers to 22 Billion Parameters. arXiv:2302.05442. *[verify full author list / venue]*
- **Why did you select this paper?** The work usually credited with popularising QK-norm at scale, and my "why QK-norm matters in practice" citation.
- **Key Findings & Methodology:** To train a 22B-parameter Vision Transformer without divergence, the authors normalise the queries and keys (QK-norm) before the attention dot product, curbing attention-logit growth that otherwise destabilised training; QK-norm is presented as an enabling stability fix at scale.
- **Limitations/Weaknesses Identified:** Vision, not language modelling, and a single ultra-large model - it says nothing about whether QK-norm helps below 100M.
- **Direct Connection to *Your* Project:** Establishes why the technique exists and matters at scale, leaving open exactly the small-scale, isolated question I address.

OLMo 2 Team (2025) - 2 OLMo 2 Furious

- **Full IEEE/APA Citation:** OLMo 2 Team (Walsh, P., Soldaini, L., Groeneveld, D., et al.) (2025). 2 OLMo 2 Furious. arXiv:2501.00656. *[verify preferred citation format]*
- **Why did you select this paper?** A fully-open, modern LM report showing QK-norm is now standard practice, not a niche trick.
- **Key Findings & Methodology:** OLMo 2 applies RMSNorm to the query and key projections before attention (QK-norm) and uses a "reordered norm" (normalising layer *outputs* rather than inputs) plus z-loss; the authors report these normalisation changes prevent attention logits from growing too large and improve training stability and loss.

- **Limitations/Weaknesses Identified:** Billion-scale; QK-norm is bundled with reordered-norm and z-loss, so its isolated contribution is not separable here.
- **Direct Connection to *Your* Project:** Confirms contemporary adoption (relevance) and, again, bundles QK-norm with other stability fixes - precisely the confound my isolated ablation removes.

Bhatti (2025) - Architectural Trade-offs in Small Language Models Under Compute Constraints

- **Full IEEE/APA Citation:** Bhatti, S. S. (2025). Architectural Trade-offs in Small Language Models Under Compute Constraints. arXiv:2512.20877. *(single-author preprint - verify before final use.)*
- **Why did you select this paper?** The counter-evidence to the "QK-norm / modern-components help small models" line: it reports that techniques successful in large models do not necessarily transfer to the small regime.
- **Key Findings & Methodology:** A systematic small-model study under strict compute budgets, building up from a linear next-token predictor through MLPs, self-attention and multi-layer transformers, evaluated on character-level Tiny Shakespeare and word-level PTB / WikiText-2 (test NLL vs parameters vs FLOPs). Finds attention beats MLPs per-FLOP even at small scale, that adding depth/context without enough optimisation can hurt, and - the key point for me - that large-model techniques do not automatically transfer down (demonstrated for rotary position embeddings).
- **Limitations/Weaknesses Identified:** Very small char/word corpora and a progression-of-architectures design rather than an isolated component ablation; the clearest "no-transfer" result is for RoPE specifically. It does not isolate QK-norm and does not sweep the learning rate.
- **Direct Connection to *Your* Project:** Supplies the "modern components don't automatically transfer to small scale" half of the contradiction that motivates my gap - but because it centres on RoPE and tiny datasets rather than an isolated QK-norm LR sweep, it leaves my exact question open.

Further papers to evaluate: the mechanistic works (Sun et al. 2024, massive activations, arXiv:2402.17762; the attention-sink study, arXiv:2410.10781); Zhai et al. 2023 (attention entropy collapse, arXiv:2303.06296); the competing logit-control methods (Gemma 2 soft-capping; PaLM z-loss; NormFormer); and Yang et al. 2022 (μP, why LR is the axis, arXiv:2203.03466).

(Duplicate this block as you discover new literature throughout the course)

---

3. Generative AI & Digital Tool Transparency Log

In line with academic integrity policies, any use of digital tools (AI search engines, LLMs, translation, or advanced editing software) to support your research process must be declared below. **Note: Using AI to write your final text from scratch is prohibited. Using tools for brainstorming, search, or copy-editing must be documented.**

Date	Tool Used (e.g., ChatGPT, Elicit, Claude)	Task Performed (e.g., Literature search, refining a sentence, brainstorming variables)	The Exact Prompt or Input Query Used	How you critically evaluated/modified the output before using it
08/07/2026	Claude (Claude Code, Opus 4.8)	Summarising Week 1 & 2 course materials; brainstorming and scoping a research topic; scaffolding this notebook's structure and drafting starting points for the reflections	"Review the research notebook and course materials from both weeks and assess what work is needed by Friday"; "I want to explore how architectural changes affect LLM training/performance - help me scope it"	Chose to keep the research question deliberately broad rather than accept a premature narrowing; adopted the small-model feasibility constraint after weighing compute limits.
10/07/2026	Claude (Claude Code, Opus 4.8)	Scoping the topic toward small-transformer architectural ablations; assessing Karpathy's nanochat as a candidate platform; outlining an initial pitch	"Explore how architectural changes affect LLM training/performance"; shared the nanochat repo link for assessment	Adopted the small-model/feasibility framing; treated the pitch outline as a scaffold to re-voice; sanity-checked nanochat's claims against its README before trusting them.
24/07/2026	Claude (Claude Code, Opus 4.8)	Literature search on QK-norm and small-model architecture; assessing four candidate ablation directions; drafting a research-findings file, pitch slides and narration script	"What's been done in this area - does technique X hold on setting Y?"; "pin down what to ablate"; "are the pitch and script honest - do they need changes?"	**Fact-checked the outputs against primary sources and caught real errors** - a fabricated quote and two wrong venue attributions - which I corrected; rescoped the pitch from a novelty claim to an honest verification study; re-voiced all narration. Reference details remain flagged *[verify]* pending my own primary-source checks.
18/08/2026	Claude (Claude Code, Opus 4.8)	Setting up the LaTeX/IEEE writing environment for the Introduction: creating the paper folder, a build script (`build.sh`) that auto-detects a TeX engine, a README, and annotating the IEEE template with plain-language comments defining unfamiliar terms	"Create the IEEE folder structure from the downloaded template"; "does LaTeX support comments? add comments defining terms the layman may not know"; "make build.sh optionally take an argument for which .tex to compile"	Read and understood each shell command before running it; installed the TeX engine (tectonic) myself and confirmed the script compiled the PDF; treated the template comments as private learning aids (they never appear in the output). No paper prose was AI-written.
20–21/08/2026	Claude (Claude Code, Opus 4.8)	Converting my own hand-written Introduction draft (`draft.md` , my prose) into IEEE LaTeX (`draft.tex`) - a format-only conversion - and cross-checking that my in-text citation numbers pointed at the right references	"Convert my draft.md into a .tex file; make a new file, don't overwrite main.tex"; "verify I'm referencing the correct document in each place"; "remove references that are never used and update the numbering"	The prose is my own words, kept verbatim (typos included); AI only added LaTeX structure and mapped [n]→\cite. I reviewed the citation-accuracy report, acted on the flagged errors myself, and verified the two newly

Date	Tool Used (e.g., ChatGPT, Elicit, Claude)	Task Performed (e.g., Literature search, refining a sentence, brainstorming variables)	The Exact Prompt or Input Query Used	How you critically evaluated/modified the output before using it
				added sources (TinyStories, Cramming) against arXiv before accepting them.
21/08/2026	Claude (Claude Code, Opus 4.8)	Auditing this notebook before submission - checking formatting/consistency, fixing citation-year/date/title mismatches, and scaffolding additional weekly-log and source-evaluation blocks for me to complete	"Is the notebook correctly formatted? what work is needed before submission?"; "do the consistency fixes"; "help me fill out the weekly logs and paper evaluations"	Accepted the mechanical fixes after checking each against my git history and the primary sources; wrote all reflective content and critical evaluations in my own words - AI supplied structure and factual paper summaries only (e.g. it re-checked OLMo 2 and the Bhatti paper against arXiv and flagged that I had overstated the latter), which I verified and re-voiced.

---

4. Weekly Activity & Reflection Logs

Update this section at the end of *every* week. Keep reflections concise, focusing on **actions taken**, **obstacles hit**, and **next steps**.

Week 01: Focus Area - Research Approach Selection

- **Activities Completed:**
  - Reviewed Week 1 materials: the CS6513 introduction, Creswell (6e) Ch.1 *The Selection of a Research Approach*, and the articles by Vogt (2008), Finlay (2012), and Miller (2006).
  - Studied the three research approaches (quantitative, qualitative, mixed methods) and the worldview → design → methods framework.
- **Key Insights / Discoveries:**
  - Vogt's "dictatorship of the problem" reframed method selection as problem-led rather than preference-led - directly relevant to keeping my topic feasible.
  - My intended ML/LLM topic maps most naturally onto a quantitative/experimental design.
- **Obstacles Encountered & How You Overcame Them:**
  - No fixed thesis topic yet, this has been addressed by settling on a broad-but-directional area to refine over the block
- **Plan for Next Week:**
  - Engage with the Week 2 literature-review materials and continue shaping a working research question.

Week 02: Focus Area - Review of the Literature

- **Activities Completed:**
  - Reviewed Week 2 materials: Creswell Ch.2 *Review of the Literature* and the articles by Xiao & Watson (2019), Winchester & Salji (2016), and Basten (2018).

    - Established a provisional working research question in the LLM-architecture space and recorded the feasibility constraints (small-model ablations) in Section 1.
    - Started the Critical Source Evaluation Log and the GenAI Transparency Log.
    - Explored Karpathy's *nanochat* repository (a hackable, single-node GPT training/inference pipeline) as a candidate platform for running the planned ablations. Built a small model as a trial run.
    - Submitted the GitHub repository + research-notebook link on Brightspace.
- **Key Insights / Discoveries:**
    - Xiao & Watson's systematic-review steps give me a concrete, defensible process for the literature search I'll run over the coming weeks.
    - Training the model on my machine took longer than anticipated. I will need to be careful to schedule time for when the model is training.
- **Obstacles Encountered & How You Overcame Them:**
    - Scope: "architectural changes to LLMs" is too broad and compute-infeasible. Narrowed the framing to single-lever ablations on small models and logged the scoping decision in the Pivot Log.
- **Plan for Next Week:**
    - Run an initial systematic search for architecture/ablation papers and add 3–4 to Section 2; begin narrowing toward one architectural lever, a dataset, and a metric set ahead of the Week 04 pitch.
    - Assess nanochat (vs nanoGPT) as the experimental base: reproduce the baseline run and confirm what a single-lever architectural change would involve.

Week 03: Focus Area - Literature Search & Topic Exploration

- **Activities Completed:**
    - Ran a systematic search for architecture/ablation papers and explored four candidate ablation directions: (1) whether "modern tricks don't help at 70M" survives over-training; (2) leave-one-out of nanoGPT-speedrun tricks; (3) depth-vs-width under data constraints; (4) QK-normalization × learning rate.
    - Assessed each for feasibility on my M4 Pro.
- **Key Insights / Discoveries:**
    - The over-training and speedrun-stack directions are compute- or tooling-infeasible on a laptop (cloud GPUs / CUDA-only kernels); the QK-norm and depth-vs-width directions are laptop-feasible.
- **Obstacles Encountered & How You Overcame Them:**
    - Had to narrow four different directions to choose from. Over-training suvival and leave-one-out speedruns require more compute than I have reasonably available. Additionally some nanogpt speedrun tactics require CUDA compatible GPUs, so they were removed from the choice pool. Leaving QK-norm and depth-vs-width as the laptop-feasible options.
- **Plan for Next Week:**
    - Narrow to a single feasible lever and prepare the Research Proposal Pitch.

Week 04: Focus Area - Narrowing & Research Proposal Pitch

- **Activities Completed:**
    - Narrowed to **QK-norm × learning rate at <100M** and ran prior-art verification (Wortsman 2024; IMU-1 2026; the contested small-model claims).
    - Found the effect is already established, so honestly rescoped the study to a **verification + boundary-charting** design and switched the planned platform from nanochat to plain nanoGPT (μP off).
    - Drafted and revised the 5-minute Research Proposal Pitch (slides + narration script) and this notebook.
- **Key Insights / Discoveries:**
    - Prior-art scrutiny matters: an initial "novel" framing did not survive it, and I had to catch and correct AI-assistant citation errors rather than trust them.
- **Obstacles Encountered & How You Overcame Them:**
    - Separating genuine novelty from a saturated area took several search passes and careful fact-checking.
- **Plan for Next Week:**

- Fork nanoGPT, implement the isolated μP-off QK-norm toggle, reproduce a baseline, and run a pilot LR-sweep to validate the per-run time estimate.

Week 05: Focus Area

This was a low-activity period.

- **Activities Completed:**
  - Limited progress this week. background reading, upskilling on ML fundementals.
- **Key Insights / Discoveries:**
  - No major discoveries, reinforced my understanding of ML, learned more about AlexNET etc.
- **Obstacles Encountered & How You Overcame Them:**
  - The Week-4 implementation plan (fork nanoGPT, μP-off toggle, pilot LR-sweep) did not start on schedule. Will start this after submission of the literature review and feedback is received.
- **Plan for Next Week:**
  - Likely little/no work will be done, personal events clash.

Week 06: Focus Area

Low activity week, personal calendar conflicted.

- **Activities Completed:**
  - Reread some papaer abstracts to keep the project in my mind.
- **Key Insights / Discoveries:**
  - No major discoveries.
- **Obstacles Encountered & How You Overcame Them:**
  - Implementation and writing continued to slip. Personal calander clashed with the project. Aiming to catch up in the final week.
- **Plan for Next Week:**
  - Aim to complete literature review. Improve formatting and any errors in this document.

Week 07: Focus Area - Writing the Introduction & Mini Literature Review

- **Activities Completed:**
  - Set up an IEEE-format LaTeX writing pipeline (paper scaffold, `build.sh`, README) and compiled the PDF locally on the M4 Pro.
  - Wrote the full Introduction + mini literature review in my own words, structured around the five required components (context/hook, literature context, gap, contribution + hypotheses, roadmap).
  - Broadened the reference set from the core papers to the surrounding literature - the mechanism (massive activations / attention sinks), the competing logit-control methods (soft-capping, z-loss, NormFormer), and the data/feasibility precedents (TinyStories, Cramming) - mapping each citation to the specific claim it supports.
  - Caught up and consolidated this notebook (expanded source evaluations, GenAI log, weekly logs).
- **Key Insights / Discoveries:**
  - Drafting the review made the shape of the gap concrete: the positive and negative small-model evidence disagree at roughly the same scale, yet neither isolates QK-norm or sweeps the learning rate - which sharpened my contribution.
  - Citation-checking caught real problems (an uncited reference; and a paper - Bhatti 2025 - whose findings I had characterised too strongly), reinforcing that no reference is safe until read against the primary source.
- **Obstacles Encountered & How You Overcame Them:**
  - Recovering from the earlier low-activity period meant compressing a lot of writing and verification into a short window. Lost access to Scopus, was unable to resolve this. As a result, a handful of papers are unverified.
- **Plan for Next Week:**

- Regain access to Scopus, reach out for support.
- Resolve the remaining *[verify]* citation details against Scopus / primary PDFs, correct the Bhatti characterisation in the draft Introduction, and return to the implementation that slipped (nanoGPT fork, μP-off QK-norm toggle, pilot LR-sweep).

(Continue adding weekly blocks as the module progresses)

[**NOTE:**]{.underline} Make sure to regular commits to the GitHub repository as the research notebook is updated. These commits will also form part of the evaluation of your work.