CS 839: Foundation Models

Homework 1

Instructions. Read the two problems below. Type up your answers in LaTeX. Submit your answers in two weeks (i.e., Oct. 2, 2025, end of day). You will need a computer for this assignment, but a laptop without a GPU should still be sufficient. Work **individually**.

Deliverables. Submit:

- A PDF write-up answering all questions clearly.
- All supporting code as separate files (zip or repo link).

Due: Thur Oct 2, 11:59pm CT.

LLM policy. You may use LLMs for writing help, brainstorming, or code scaffolding. Attach an **LLM usage log** with: (i) prompts, (ii) model name/version, and (iii) how you verified outputs. Hallucinated or unverified content may result in loss of credit.

1. GPT-2 parameter counts *[20 pts]*
Read Michael Wornow, "Counting Parameters in Transformers" (2024) carefully, and compute the number of parameters in GPT-2:

1. Write symbolic formulas for embeddings, attention, MLP, and LayerNorms. Give a total in terms of V, E, H, L, P, where: V = vocabulary size, E = model/embedding dimension, H = number of attention heads, L = number of transformer layers, and P = maximum positional indices (context length). *[8 pts]*
2. Plug in the hyperparameters for **GPT-2**. Report totals and a breakdown by component as concrete parameter counts (numbers). Verify with code and explain any mismatch. *[6 pts]*
3. Estimate totals for **GPT-2 Medium**, **GPT-2 Large**, and **GPT-2 XL** using documented (E, H, L). Briefly explain which terms dominate scaling. *[6 pts]*

2. Modern model study *[80 pts]*
Choose *one*: **gpt-oss-20b**, **Qwen3-8B**, or **Gemma3-4B**. For the chosen model:

1. Report the exact configuration you use (cite the source). *[4 pts]*
2. Derive **component-wise** formulas (as in Problem 1) consistent with the chosen architecture, and compute both *per-layer* and *total* parameter counts using your conventions. *[20 pts]*
3. Report the final totals as concrete parameter counts (numbers). If MoE, also report *active* parameters per token. *[12 pts]*
4. Provide code that demonstrates a programmatic check. *[12 pts]*
5. Identify the key components that differ from GPT-2. For each:
   - Summarize the design motivation in your own words. *[12 pts]*
   - Provide evidence of effectiveness from the paper/tech report (ablation, benchmark, or efficiency claim). *[12 pts]*
6. Give a short overall summary (bullets or a short paragraph) of the main design trends you observe. *[8 pts]*

Suggested resources.

- GPT-2 counting blog: Counting Parameters in Transformers (Wornow, 2024).
- GPT-2 docs: huggingface.co/docs/transformers/en/model_doc/gpt2.
- GPT-OSS 20B model card: arXiv:2508.10925.
- Qwen3 tech report: arXiv:2505.09388.
- Gemma 3 tech report: arXiv:2503.19786.