CSCI 145 Problem 21: Audit a Scaling Memo

August 27, 2026

Submission Instructions

Please upload your work to Gradescope by **11:59pm on DATE**.

- Turn in a single PDF of *handwritten* solutions. Typed (including LaTeX) solutions will not be accepted.
- Problems are grouped into three-week **cycles**. Once per cycle, you will meet with the grader and present the solution to *one* problem from that cycle, selected at random when you arrive; bring your work and be ready to explain any of them.

Problem 21: Audit a Scaling Memo 2 points

A lab plans a transformer run using the empirical loss model

$$\mathcal{L}(N, D) = \mathcal{L}_\infty + aN^{-\alpha} + bD^{-\beta},$$

where N is the non-embedding parameter count and D is the number of training tokens. Its memo assumes $C = 6ND$ and concludes that the compute-optimal choices scale as

$$N \propto C^{\beta/(\alpha+\beta)}, \qquad D \propto C^{\alpha/(\alpha+\beta)}.$$

It then says the recommendation is unchanged if context length grows from 2,048 to 131,072 tokens.

The allocation is correct under the memo's compute model, including constants, when

$$N^* = \left(\frac{\alpha a}{\beta b}\right)^{1/(\alpha+\beta)} \left(\frac{C}{6}\right)^{\beta/(\alpha+\beta)}, \qquad D^* = \left(\frac{\beta b}{\alpha a}\right)^{1/(\alpha+\beta)} \left(\frac{C}{6}\right)^{\alpha/(\alpha+\beta)}.$$

Audit the long-context conclusion rather than re-deriving this optimization. Rebuild the training-compute constraint from a transformer block of width d, depth L, and sequence length n, accounting separately for parameter projections and attention mixing. Use $N = \Theta(Ld^2)$ to identify the regime in which $C = 6ND$ is a defensible approximation and the regime in which the omitted term can change the allocation.

End with a recommendation about what the lab must remeasure or re-optimize before accepting the long-context plan. A constants-free asymptotic audit is sufficient.

Results you may use. You may use the per-layer costs $O(nd^2)$ for projections and $O(n^2d)$ for attention mixing.