Assignment 2: Scaling Laws, Normalization, and the Geometry of Transformers

Course COMP5801H/4900A: Generative AI and LLMs

Release Date: January 30, 2026 — **Due Date:** February 12, 2026

Objective

To critically examine the internal mechanics and scaling properties of Transformer-based architectures through a rigorous evaluation of their stability, efficiency, and representational limits.

Questions

There are six questions provided below, each containing one assertion. Respond to **any three** questions from the six provided.

Collaboration

All **graduate and undergraduate students** (i.e., belonging to COMP5801H or COMP4900A) will complete and submit the assignment **individually**.

Format

For each of the questions, you must take a stand: **Agree** or **Disagree** with the provided assertion. You must support your position with a combination of:

- **Mathematical Justification:** Formal analysis of update equations, gradients, and/or gate mechanisms.
- **Experimental Justification:** Plots generated from scripts carrying out the experiments, along with detailed analysis/interpretation of the results.

For each position you take, you are required to provide at least **two strong** independent evidences. One of these evidences could involve a mathematical justification. So, if you are providing two evidences for a question, it could be either one mathematical justification and one experimental justification or two different experimental justifications. *Experimental investigations conducted across two different datasets comprising significantly distinct feature sets will be evaluated as two separate and independent pieces of empirical justification.* You should justify how the two datasets provide fundamentally "different" insights into the architecture. For every justification, it is not sufficient to only show the result (in terms of an equation or a plot). You should clearly explain

the result and elucidate how that result serves as a **strong evidence** to the position that you have taken.

For the experimental investigations, you are strongly encouraged to design simple, targeted experiments (roughly plan for experiments that can be executed on a standard laptop CPU); the focus should be on providing clear analysis and insights from the results rather than complex training runs requiring multiple GPUs. Any experimentation that relies on stochasticity or randomization should provide results involving at-least three different random seeds, along with the average and standard deviation of performance measures.

For every question, make sure to reference **at least two** academic papers to strengthen your arguments.

Submission Instructions

All materials must be submitted via **Brightspace** no later than **February 12, 11:59 PM ET**. Your submission must include:

1. **Formal Report:** A PDF document containing your stance on each assertion, mathematical derivations (as appropriate), experimental plots, associated discussions of the results (clarifying how they support your assertion), and academic citations. See "Grading Criteria" below for more details. Make sure to include the **question number** from this document for the selected three questions, so that the assertion you are responding to is clear to the grader.

2. **Implementation Files:** All source code used for the experiments. The experiments must be reproducible using the provided code. Include a `README` file providing clear instructions on how to run the scripts and generate the results. Include comments in your code as appropriate. The code can also be provided in Jupyter notebook files. The code should be runnable as submitted (they **cannot** be submitted in word or PDF documents).

Question 1: The Myth of "Infinite" Context in Transformers

Assertion: "While the Transformer's Self-Attention mechanism is mathematically defined for sequences of arbitrary length, the fixed frequency of Sinusoidal Encodings creates a 'structural myopia' that renders the architecture functionally inferior to an RNN when processing sequences significantly longer than those encountered during the model's training phase."

—

Question 2: Layer Normalization as a "Double-Edged Sword"

Assertion: "Layer Normalization (LayerNorm) is essential for stabilizing Transformer training, but its tendency to reduce feature variance across the hidden dimension (d_{model}) acts as a lossy compression that hurts performance on fine-grained regression tasks compared to Batch Normalization."

—

Question 3: Softmax Bottleneck in Large Vocabularies

Assertion: "While using Generative AI in production, the primary computational bottleneck is not the complexity of the Self-Attention layers, but rather the Output Projection layer. As the

model's vocabulary size (V) (meaning the total number of unique words/tokens it can predict) grows, the cost of the final linear transformation eventually surpasses the cost of processing the actual sequence, making large-scale vocabularies the true 'Hard Ceiling' for real-time inference on standard hardware."

—

Question 4: Weight Tying and Representation Collapse

Weight Tying: In the context of Transformer architectures, Weight Tying (often referred to as Shared Embeddings) is a parameter-sharing technique where the model uses the exact same set of weights for two different layers: the Input Embedding layer and the Output Projection layer.

Assertion: "Tying the weights of the Input Embedding and the Output Projection ($W_{embed} = W_{out}^T$) is a parameter-saving trick that fundamentally limits the semantic richness of the latent space, forcing the model to represent 'Input Meaning' and 'Output Probability' in the same geometric manifold."

—

Q5: The Redundancy of Multi-Head Attention

Assertion: "Multi-Head Attention (MHA) is computationally inefficient for small-to-medium scale tasks; a single-head attention mechanism with a larger dimension d_{head} can achieve identical performance with significantly lower memory overhead on a CPU."

—

Question 6: Residual Connections as "Identity" Shortcuts

Assertion: "Residual connections ($x + \text{SubLayer}(x)$) do not actually allow for 'deeper' representation learning; they simply transform a deep network into an ensemble of shallow networks, meaning that a 12-layer Transformer is functionally no 'smarter' than a 3-layer Transformer."

—

Grading Criteria

- **Strength of the evidences (45%):** Is the position clearly stated? Are there at-least two strong evidences for the position taken? Are the evidences appropriate for the stated position? Are there any technical errors in the arguments provided?
- **Clarity of report (45%):** Experimental investigations must be clearly explained (i.e., the dataset used, hyperparameters selected, the experimental procedure followed, and rational for the choices), plots need to be clear (clear labelling of axes and legends), and all terms/variables of an equation should be clarified for theoretical justifications. If a derivation is performed, each step of the derivation should be clarified. All equations should be clearly explained in words. The results of an experimental investigation must be clearly explained. For both mathematical and empirical results, reasons for how these results support the stated position should be clearly explained. There are no page limits to the report, just make sure that all justifications are clear, without ambiguities or confusions.

- **Implementation (0%):** Are empirical results reproducible from the submitted implementation files and provided instructions? The implementation files do not carry any points, but if the results are not reproducible, the empirical evidences in your report will **not be awarded** any points. To ensure the reproducibility of your empirical findings, any experimentation involving randomization must clearly specify the random seeds and environment configurations used.
- **Literature citations (10%):** Do the provided literature references strengthen the arguments presented?

Reminder: Course Policy on Generative AI Tool Usage

You are free to use any Generative AI Tool for your assignments or projects in the course, but you bear responsibility for all your submissions. Since Generative AI tools may plagiarize from other sources on the web, you should be extremely careful while using them. They are also prone to producing incorrect information with exaggerated confidence. All prior work should be rigorously cited, and any plagiarism will be considered an academic offence.