HW2: Reproducing ChatGPT

1. Create an account on HPRC (https://hprc.tamu.edu/apply/)
    a. Apply for Basic Allocation on Grace (20,000 Service Units)
    b. 20,000 Service Units (SUs) ~= 250 A100 (40G) GPU hours
2. Train a variant of GPT-2
    a. Limit training time/resource to max **24 hours** w/ **one A100 40G GPU**
    b. Follow instructions at https://github.com/parasol-aser/hw-reproduce-chatgpt

Your Goal: train the best GPT model ***from scratch*** within the resource budget - Top 10 submissions with the highest HellaSwag accuracy will each earn 1 bonus point - Top 3 will earn 4, 2, 1 additional bonus points respectively

Your Strategies:

- Tune hyper-parameters guided by the scaling laws
- Try different architectures, e.g.:
    - Group Query Attention
    - Replace LayerNorm by RMSNorm
    - Replace absolute positional encoding by RoPE
    - Replace GeLU activation function by SwiGLU
    - Drop Positional Encoding
    - Change KQV (e.g., merge K and Q)
    - Elimination or Modification of FFN Layers
    - Mixture of Experts (MoE)
    - …

Submission (5pt):

- Your final model checkpoint and original logs stored on Grace (2pt)
    - Need to share a folder with our grader
- Your training code (only diff is required if based on karpathy/llm.c) (1pt)
- A report that describes your solution and results (including remaining challenges and failures if any) (2pt)

- Limit your report to three pages with 10pt font size