Pseudocode for Grouped-Query Attention (GQA) Uptraining

Siyu Yang, Based on the paper by Joshua Ainslie et al.

Algorithm 1 Uptraining for Grouped-Query Attention

Require: M_{MHA}: Multi-Head Attention model
Require: D_{train}: Training dataset
Require: G: Number of groups in GQA
Require: $\alpha = 0.05$: Proportion of original training compute for uptraining

$M_{\text{GQA}} \leftarrow \text{ConvertToGQA}(M_{\text{MHA}}, G)$
function ConvertToGQA(M_{MHA}, G)
    Initialize M_{GQA} with M_{MHA}'s architecture
    for each attention layer in M_{MHA} **do**
        for $g = 1$ to G **do**
            $K^g_{\text{group}} \leftarrow \text{mean}(\{K_h | h \in \text{Group}_g\})$
            $V^g_{\text{group}} \leftarrow \text{mean}(\{V_h | h \in \text{Group}_g\})$
        end for
        Assign grouped $K_{\text{group}}, V_{\text{group}}$ to M_{GQA}
    end for
    return M_{GQA}
end function
$steps_{\text{total}} \leftarrow$ Number of steps in M_{MHA}'s original pre-training
$steps_{\text{uptrain}} \leftarrow \lceil steps_{\text{total}} \times \alpha \rceil$
for $step = 1$ to $steps_{\text{uptrain}}$ **do**
    $batch \leftarrow$ Sample from D_{train}
    Update M_{GQA} on $batch$ ▷ Using grouped K and V
end for
return M_{GQA}