Quiz 6

601.471/671 NLP: Self-Supervised Models

Johns Hopkins University Fall 2026

Name: _____ **Student version**

Time: 20 minutes. Total: 100 points.

Give short answers and show calculations. No calculator is needed; log is the natural logarithm.

Q1. FFNs and Block Layouts *30 points*

(a) **15 points. FFN budgets.** Compare bias-free FFNs on a row vector $x \in \mathbb{R}^D$:

$$\text{ordinary: } f(xW_1)W_2, \qquad \text{gated: } \big(f(xU) \odot (xV)\big)W_o.$$

Here $W_1 : [D, F]$, $W_2 : [F, D]$, $U, V : [D, G]$, and $W_o : [G, D]$. All matrices are separately trainable; f is parameter-free; $\odot$ multiplies elementwise.

Review: Parameter counting from Quiz 1, Q3(b). Gated FFNs and matching their budgets are new applications.

(i) **3 points.** Give the ordinary FFN's matrix-parameter count in terms of D, F.

Your answer:

(ii) **3 points.** Give the gated FFN's matrix-parameter count in terms of D, G.

Your answer:

(iii) **6 points.** With $D = 6, F = 24$, find the gated width G that matches the ordinary FFN's parameter count.

Your answer:

(iv) **3 points.** For those widths, find the common matrix-parameter count.

Your answer:

(b) **15 points. Serial or parallel.** A is attention, M an FFN, and N, N_1, N_2 normalization:

$$\text{serial: } h = x + A(N_1(x)), \quad y = h + M(N_2(h)); \qquad \text{parallel: } y = x + A(N(x)) + M(N(x)).$$

(i) **3 points.** Which layout's FFN receives a state containing this block's attention update?

Your answer:

(ii) **3 points.** Write the input to the serial FFN.

Your answer:

(iii) **3 points.** Write the input to the parallel FFN.

Your answer:

(iv) **6 points.** Name one computation that the parallel branches can share or combine.

Your answer:

Q2. Pre-training Data and Optimization *30 points*

(a) **10 points. Held-out data.** A pipeline randomly splits pages into training and validation, then removes exact duplicate documents *within each split* separately.

(i) **4 points.** Why can identical document text still appear in both splits?

Your answer:

(ii) **6 points.** Give one pipeline change that prevents this exact overlap while keeping held-out validation.

Your answer:

(b) **10 points. Fixed compute.** These synthetic runs have comparable architecture family, tokenizer, training mixture, validation set, and suitable optimizer settings at each size. Measured training compute is equal. Lower validation loss is better.

Run	Parameters	Processed tokens	Validation loss
A	100M	8B	3.4
B	200M	4B	3.1
C	400M	2B	3.3

(i) **4 points.** Which measured run is best?

Your answer:

(ii) **6 points.** Why would extending only C to 4B tokens, at additional compute cost, change the fixed-compute comparison?

Your answer:

(c) **5 points. Optimizer memory.** AdamW stores two moment tensors, each with one entry per trainable scalar parameter. A model has 100 million parameters and each moment entry occupies 4 bytes. Count only these tensors; $1\,\mathrm{GB} = 10^9$ bytes.

(i) **5 points.** Find the combined storage of both moment tensors in decimal GB.

Review: Tensor-memory counting from Quiz 5, Q2(a)(iv)–(v), applied here to optimizer state.

Your answer:

(d) **5 points. Learning-rate schedules.** A warmup–stable–decay schedule first increases the learning rate, then holds it steady, then reduces it.

(i) **2 points.** Give one reason to use a smaller learning rate at the very start.

Your answer:

(ii) **3 points.** Why might an extremely small learning rate for the *entire* fixed-budget run be a poor choice?

Your answer:

Q3. Instruction Tuning *20 points*

(a) **20 points. Response-only loss.** An instruction/context consists of tokens P_1, P_2; its demonstrated response is R_1, R_2, EOS. EOS marks the end of the response and is included in its loss. The table shows already aligned next-token predictions; each probability is for the listed target given the *gold* prefix.

Prediction after	P_1	P_2	R_1	R_2
Target token	P_2	R_1	R_2	EOS
Probability of target	$1/4$	$1/2$	$1/4$	$1/2$

(i) **8 points.** Which target tokens in the table should count toward the instruction-tuning loss? List them.

Your answer:

(ii) **12 points.** Using the target tokens you selected in (i), compute the mean negative log-likelihood in terms of $\log 2$.

Review: Mean negative log-likelihood from Quiz 2, Q2(a), and averaging over selected targets from Quiz 5, Q3(ii). The response-only instruction-tuning setup is new.

Your answer:

Q4. Learning a Reward Model *20 points*

(a) **5 points. Tokens as actions.** Consider the token-level RL description of response generation in the lecture.

(i) **2 points.** What is the state just before choosing the next token?

Your answer:

(ii) **2 points.** What is the action?

Your answer:

(iii) **1 point.** When does the response-level reward arrive in this setup?

Your answer:

(b) **15 points. Preference probability and loss.** For a prompt p, humans prefer response w over l. The reward model uses

$$P(w \succ l \mid p) = \sigma(\Delta), \qquad \Delta = R(w;p) - R(l;p), \qquad \sigma(z) = \frac{1}{1+e^{-z}}.$$

The example's loss is $-\log \sigma(\Delta)$. Let $R(w;p) = \log 3$ and $R(l;p) = 0$. Use exact expressions.

(i) **6 points.** Find the preference probability.

Your answer:

(ii) **6 points.** Find this example's loss.

Review: Negative-log-probability calculation from Quiz 2, Q2(a), applied here to a preference probability.

Your answer:

(iii) **3 points.** If the two scores are swapped, does the loss increase or decrease?

Your answer: