Pontificia Universidad Católica del Perú

Faculty of Science and Engineering

Intro to LLMs and Agents

Homework 2 - Solutions

Author: Saymon Nicho

Date: March 20, 2025

1 Questions and Answers

1.1 Question 1

Which of the following metrics is the correct one to evaluate LLaMA 3 intrinsically in the next-token-prediction task?

- F1 Score
- BLEU
- **Perplexity**
- Accuracy

Perplexity is the standard metric for language modeling tasks (i.e., next token prediction). It measures how "surprised" the model is by the real sequence.

1.2 Question 2

What is the difference between Cross-Attention and Causal Attention?

Cross-Attention is used when the decoder must attend to a different sequence (e.g., the encoder's outputs). Causal Attention (masked self-attention in autoregressive models) only allows each token to attend to previous tokens, preventing access to future tokens.

1.3 Question 3

Which of the following Preference Optimization algorithms requires a dedicated (or separate) reward model?

- Direct Preference Optimization
- **Proximal Preference Optimization (PPO)**

Inspired by Proximal Policy Optimization, PPO requires training a separate reward model to guide the main policy.

1.4 Question 4

Which of the following generation techniques randomly samples one word at each step? Mark all that apply.

- Greedy Decoding
- **Nucleus Sampling**
- **Top-K Decoding**
- Beam Search

Both use sampling (top-p or restricting to top-k most probable tokens). Greedy Decoding and Beam Search do not introduce randomness.

1.5 Question 5

Which prompting technique is most suitable to generate a step-by-step solution for an algebraic problem?

- Zero-shot prompting
- **Chain of Thought**
- Tree of Thought
- [START] text [SEP] text [EXTRACT]

Chain of Thought explicitly elicits intermediate reasoning steps, which is ideal for mathematical or logical problem solving.