Last Name: _____ **First Name:** _____

SUNet ID: _____@stanford.edu

Final Exam Solutions — December 10, 2025

Closed-book exam. No notes, no laptop, no phone, no internet, no AI assistance, no anything. Just your brain and a pen!

Duration: 1 hour 30 minutes **Total number of points: 100**

Instructions

Questions are grouped into logical sections to ease your thought process. There are two types of questions:

- **Multiple-choice**: One correct choice per question. Please (circle) the correct answer.
- **Free-form**: Short and concise answers.

In all cases, there is no penalty for wrong answers.

I. LLM tuning (25 points)

1. (1 point) What is a key limitation of standard Supervised Fine-Tuning (SFT) that Preference Tuning (e.g., RLHF) is specifically designed to address?
   - A. SFT inherently prevents the model from understanding structural dependencies required for complex instruction following or formatting outputs.
   - B. SFT maximizes likelihood on good examples but lacks an explicit mechanism to penalize (provide negative signals for) undesirable/harmful outputs.
   - C. SFT computationally demands an exponentially higher budget compared to Reinforcement Learning due to the need for labeled data.
   - D. SFT causes a catastrophic forgetting phenomenon that strictly prevents the model from generalizing to any prompts not seen during training.

2. (1 point) The Reward Model (RM) in a standard RLHF pipeline takes as input:
   - A. Only the prompt.
   - B. Only the response.
   - C. A prompt and a response pair.
   - D. Two competing model responses in isolation, without conditioning on the initial prompt.

3. (2 points) The Bradley-Terry formulation is used in reward modeling to:

   A. Auto-regressively generate high-quality synthetic text data to augment the SFT training set.

   B. Precise calculation of the Kullback-Leibler divergence between the reference model distribution and the current policy.

   C. Dynamically optimize the learning rate schedule of the Proximal Policy Optimization (PPO) algorithm.

   D. Model the probability that response A is better than response B based on their reward difference.

4. (2 points) In the PPO objective function, the term $-\beta \text{KL}(\pi_\theta(y|x)||\pi_{\text{ref}}(y|x))$ is included to:

   A. Prevent the model from deviating too far from the base model.

   B. Maximize the raw reward signal indiscriminately, prioritizing high scores regardless of the semantic coherence or quality of the text.

   C. Force the model to strictly memorize the training data to ensure zero hallucinations during inference.

   D. Significantly increase the entropy and diversity of generated responses to prevent mode collapse.

5. (2 points) The key theoretical insight of Direct Preference Optimization (DPO) is that:

   A. It is possible to train a reward model directly using reinforcement learning without any preference data.

   B. PPO is mathematically simpler and more stable than standard supervised learning approaches for all tasks.

   C. The optimal policy can be solved analytically, allowing implicit reward optimization via a supervised loss.

   D. Reward models become entirely redundant if the dataset size for Supervised Fine-Tuning exceeds a certain threshold.

6. (2 points) Best-of-N (BoN) is an alternative to RL training that works by:

   A. Generating a single response and accepting it only if the log-probability exceeds a pre-defined confidence threshold.

   B. Recruiting a group of N human annotators to manually write and curate the best possible response for every query.

   C. Averaging the synaptic weights of N distinct fine-tuned models to produce a robust ensemble prediction.

   D. Generating N responses and selecting the one with the highest reward model score.

7. (1 point) What is a common symptom of "reward hacking" in RLHF?

   A. The model generates high-reward outputs that are actually of low quality.

   B. The model enters a degenerative state where it completely ceases to generate any tokens after the prompt.

   C. The training loss function instantaneously drops to absolute zero within the first few optimization steps.

   D. The model completely loses its pre-trained ability to parse external function calls or utilize tools.

8. (1 point) In PPO, the value function estimates:

   A. The immediate probability distribution over the vocabulary for the very next token in the sequence.

   B. The expected future reward from the current token sequence.

   C. The statistical difference (divergence) between the current policy's weights and the old policy's weights.

   D. The binary classification accuracy of the reward model on the validation set.

9. (3+4 points) **RLHF Pipeline.** (i) List the two main training stages involved in RLHF *after* the SFT phase. (ii) Briefly explain the specific role/output of the model trained in the first of those two stages.

   (i) 1. Reward modeling. 2. Reinforcement Learning (PPO).
   (ii) The reward model (trained in the first stage) takes a prompt and response pair and outputs a scalar score representing the quality or human preference of that response.

10. (3+3 points) **DPO vs. PPO.** (i) Explain the primary architectural advantage of DPO over PPO during the training process (think about the number of models loaded). (ii) Name one possible reason why one might still use PPO over DPO.

    (i) DPO is more memory efficient because it eliminates the need to load a separate reward model and value function; it only requires the policy model and reference model.
    (ii) PPO allows for non-differentiable or sparse reward signals (e.g., from a code compiler or math verifier) which cannot be easily captured in the static pairwise preference datasets required by DPO.

II. LLM reasoning (25 points)

1. (1 point) In the context of LLMs, "reasoning" is best defined as:
   A. The capacity to accurately retrieve and regurgitate specific factual strings seen during pretraining.
   B. The computational throughput speed at which tokens are generated per second on a GPU.
   C. The ability to solve complex problems, often via multi-step logic.
   D. The ability to follow formatting instructions.

2. (1 point) What is the core idea behind Chain of Thought (CoT)?
   A. Forcing the model to output the final answer immediately to minimize token usage and latency.
   B. Fine-tuning the model exclusively on high-fidelity encyclopedia articles to improve factual density.
   C. Prompting the model to explain its thinking process before producing the final answer.
   D. Utilizing an external vector database retrieval system to fetch the answer directly from a corpus.

3. (2 points) In DeepSeek R1-Zero, what phenomenon was observed regarding the length of the model's output during RL training?
   A. It naturally increased as the model generated more reasoning steps.
   B. It consistently decreased over time as the model learned to be more concise and efficient.
   C. It remained statistically invariant regardless of the complexity of the reasoning task.
   D. It fluctuated randomly without any correlation to the reward signal or task difficulty.

4. (2 points) Group Relative Policy Optimization (GRPO) differs from PPO primarily because:
   A. It utilizes a significantly more complex neural value function that requires separate pre-training.
   B. It replaces the reward model entirely with a GAN-style discriminator network.
   C. It necessitates real-time human feedback for every single optimization step during training.
   D. It eliminates the need for a value function and estimates advantages using the average reward of a group of outputs.

5. (2 points) Which of the following is considered a "verifiable reward" for training reasoning models?
   A. A subjective human rating regarding the "politeness" or "helpfulness" of the generated tone.
   B. Whether code passes unit tests or a math answer matches the ground truth.
   C. A standard perplexity score derived from the pre-training corpus distribution.
   D. The total number of tokens in the generated response, rewarding longer outputs.

6. (1 point) What is the "aha moment" observed in reasoning models like DeepSeek R1-Zero?
   A. The model suddenly overfitting and memorizing the exact solutions to the entire evaluation dataset.
   B. The exact iteration where the training loss function mathematically converges to zero.
   C. When the model strictly refuses to answer any question it deems potentially unsafe or controversial.
   D. The model learning to self-evaluate and backtrack/correct its own mistakes during the thinking process.

7. (2 points) Why is a "cold start" (small-scale SFT with high-quality data) often used before RL training for reasoning?
   A. To teach the base model fundamental English grammar rules that were missing from pretraining.
   B. To forcefully restrict the model from utilizing excessive GPU compute resources during the RL phase.
   C. To stabilize the output format and make reasoning traces readable for humans.
   D. To artificially expand the model's context window size beyond its architectural limits.

8. (1 point) In the context of reasoning models, "distillation" refers to:
   A. Compressing the model's floating-point weights into 4-bit integers using quantization techniques.
   B. Using reasoning traces generated by a strong model (like R1) to SFT a smaller/weaker model.
   C. Systematically stripping out all reasoning tokens during the inference phase to speed up generation.
   D. Training a separate reward model based solely on human preference labels.

9. (4+2 points) **GRPO Mechanics.** (i) Explain how the advantage is calculated for an output o_i in a group of G outputs in GRPO. (ii) Name one computational benefit of this approach compared to standard PPO.

> (i) The advantage is calculated by standardizing the rewards within the group. For a specific output o_i with reward r_i, the advantage is $A_i = \frac{r_i - \text{mean}(r)}{\text{std}(r)}$, where mean and std are computed over the group of G outputs.
>
> (ii) It removes the need for a separate value function, reducing memory usage and training complexity.

10. (3+4 points) **Test-time scaling.** (i) Define test-time scaling. (ii) List two methods discussed in class to control or increase the "thinking" budget during inference.

> (i) Test-time scaling refers to the method of improving model performance by increasing the amount of compute used during inference, rather than just increasing model parameters or training data.
>
> (ii) 1. Best-of-N (sampling multiple solutions and verifying). 2. Context awareness, budget forcing.

III. Agentic LLMs (25 points)

1. (1 point) Retrieval-Augmented Generation (RAG) primarily solves which limitation of frozen LLMs?

   A. The inherent latency issues and slow inference speeds associated with large parameters.

   B. The model's inability to perform complex floating-point arithmetic calculations.

   C. Knowledge cutoff and lack of private domain knowledge.

   D. The lack of built-in safety guardrails against generating toxic or harmful content.

2. (1 point) In RAG, "contextual retrieval" refers to:

   A. Retrieving the entire database content and stuffing it into the context window.

   B. Ignoring the hard limits of the model's context window during the generation phase.

   C. Adding context (e.g., document summaries) to chunks to improve retrieval accuracy.

   D. Restricting the retrieval mechanism to look only at the very last sentence of the query.

3. (1 point) What is the "ReAct" framework in the context of agents?

   A. An interleaving loop of "Reasoning" (plan) and "Acting" (act, observe).

   B. A popular front-end JavaScript library used for building interactive user interfaces.

   C. A fine-tuning method that retrains models based on user emotional reactions.

   D. A specific type of sparse attention mask designed to optimize transformer efficiency.

4. (2 points) In a tool-use scenario, if an LLM outputs a function call (e.g., `find_teddy_bear(loc='Stanford')`), who typically executes this function call?

   A. The LLM itself simulates the execution internally and hallucinates the return value.

   B. The external runtime/system (backend), which returns the result to the LLM.

   C. The human user manually executes the code and types the result back into the chat context.

   D. The separate reward model intercepts the call and computes the result.

5. (2 points) The Model Context Protocol (MCP) aims to standardize:

   A. The internal architecture and layer normalization of standard Transformer blocks.

   B. The mathematical formulation of the cross-entropy loss function used in pre-training.

   C. The specific evaluation metrics used for scoring performance on standard benchmarks.

   D. The connection between LLM clients/hosts and data sources/tools (servers).

6. (1 point) A key difference between a standard chatbot and an "agent" is:

   A. Agents are strictly required to process multimodal inputs like images and audio.

   B. Chatbots are technically incapable of accessing the internet under any circumstances.

   C. Agents typically have autonomy to pursue goals and execute multi-step actions on behalf of the user.

   D. Agents rely on Recurrent Neural Networks (RNNs) instead of Transformer architectures.

7. (2 points) When an agent hallucinates a tool that does not exist, a common remedy is to:

   A. Update the system prompt with clear tool definitions or check for logic errors in the tool router.

   B. Increase the sampling temperature to 1.0 to encourage more creative output generation.

   C. Completely disable all tool functionality and revert to a pure text-generation mode.

   D. Switch to a significantly smaller parameter model to reduce the complexity of the output.

8. (2 points) What is the A2A (Agent2Agent) protocol designed to facilitate?

   A. The standard communication interface between the human user and the AI agent.

   B. Communication and hand-offs between different specialized agents.

   C. A technique for transfer learning between two distinct base model architectures.

   D. The process of aligning autonomous agents to strict human moral values.

9. (3+3 points) **RAG vs. long context.** (i) Why might one use RAG even if the model has a 1M token context window? (ii) Name one challenge specific to RAG systems.

   (i) RAG is often cheaper (fewer input tokens) and faster (lower latency). Additionally, models with long context can still suffer from the "needle in a haystack" problem where they fail to retrieve information from the middle of a massive context.
   (ii) Retrieval accuracy: retrieving irrelevant chunks or missing the correct ones.

10. (3+4 points) **Tool calling workflow.** (i) Describe the three high-level steps in a tool execution loop (from the LLM's perspective). (ii) What happens if the tool execution returns an "error"?

    (i) 1. The LLM generates a tool call. 2. The backend executes the tool. 3. The tool output is fed back to the LLM and returns a response.
    (ii) The error message is returned to the LLM as an observation. A capable agent will analyze the error and attempt to correct its parameters or try a different approach.

IV. LLM evaluation (25 points)

1. (1 point) Goodhart's Law in the context of LLM benchmarks states that:
   A. All standardized benchmarks are inherently useless for measuring model performance.
   B. Larger parameter models will universally outperform smaller models on every task.
   C. Human evaluation is always perfectly consistent and free from subjective bias.
   D. When a benchmark becomes a target, it ceases to be a good measure.

2. (1 point) "Data contamination" in evaluation refers to:
   A. The presence of toxic or harmful content within the supervised fine-tuning dataset.
   B. Formatting the SFT data in a JSON structure that the model cannot parse.
   C. Injecting random Gaussian noise into the prompt embeddings during inference.
   D. Evaluation/test data leaking into the model's pretraining corpus.

3. (1 point) "LLM-as-a-judge" typically involves:
   A. Using a strong model (e.g., Gemini 3 Pro) to rate the outputs of other models based on a rubric.
   B. Using a very small, distilled model to evaluate the capabilities of a much larger model.
   C. Asking the model to output a confidence score regarding its own generation.
   D. Utilizing a legacy BERT-based model specifically fine-tuned for binary sentiment analysis.

4. (2 points) Which of the following is a "pairwise" evaluation method?
   A. Assigning an absolute scalar quality score to a single response.
   B. Presenting two model responses A and B and asking "Which is better?".
   C. Calculating the n-gram overlap BLEU score of a response against a reference.
   D. Running a static analysis tool to verify if the generated code compiles successfully.

5. (2 points) "Position bias" in LLM-as-a-judge refers to:

   A. The model demonstrating a preference for answers that align with specific political ideologies.

   B. The model focusing exclusively on instructions located at the very end of the prompt.

   C. The model favoring an option at a specific position, regardless of content quality.

   D. The model systematically assigning higher scores to answers that are significantly longer.

6. (2 points) The "pass@k" metric is defined as:

   A. The probability that at least one of the k generated solutions is correct.

   B. The probability that every single one of the k generated solutions is correct.

   C. The percentage of test cases strictly passed by the very first generation attempt.

   D. The average quality score across the top k models on the leaderboard.

7. (2 points) Which benchmark is primarily designed to test *multitask knowledge* across subjects like STEM, humanities, and social sciences?

   A. MMLU

   B. GSM8K

   C. HumanEval

   D. HarmBench

8. (1 point) Why are n-gram metrics like BLEU and ROUGE often considered insufficient for evaluating modern open-ended chat LLMs?

   A. They are computationally too expensive and slow to calculate for large datasets.

   B. They strictly require high-end GPU acceleration to be computed effectively.

   C. They are linguistically limited and can fundamentally only handle English text.

   D. They focus on exact word overlap rather than semantic meaning and reasoning correctness.

9. (3+4 points) **LLM-as-a-Judge biases.** (i) Define "verbosity bias". (ii) Propose one concrete method to mitigate a bias of your choice in LLM-based evaluation.

   (i) Verbosity bias is the tendency of LLM judges to prefer longer answers, even if they are less accurate or more repetitive than shorter ones.

   (ii) To mitigate position bias, use "position swapping": evaluate the pair (A, B) and then (B, A) and check for consistency.

10. (3+3 points) **Benchmarks.** (i) What does the "SWE-bench" benchmark evaluate? (ii) Explain the difference between static benchmarks (like MMLU) and dynamic leaderboards (like Chatbot Arena).

> (i) SWE-bench evaluates a model's ability to solve real-world software engineering problems, specifically by resolving GitHub issues in popular Python repositories.
>
> (ii) Static benchmarks use a fixed set of questions/answers, which are prone to contamination. Dynamic leaderboards rely on live human voting, making them more reflective of general user preference.

*
* *

We hope you enjoyed spending this quarter with us. Wishing you to spend an amazing holiday season, and the best of success ahead!