ASSIGNMENT-4

K. Bhikshapathi
Srno: 24489.

Question-1

a) How was reinforcement learning (RL) used in training DeepSeek R1-zero? what advantages and challenges are associated with using RL without supervised fine-tuning (SFT)?

Ans:- Deepseek-R1-zero was trained using RL directly on a base language skipping supervised fine-tuning. It utilize a Group Relative policy optimization policy (GROP) algorithm where, each question the model generates several candidate responses, access them using rewards and then update policy

Advantages:-

- Emergent reasoning: The model develops long chains of thought and self verificant
- Cost efficient :- It avoids reliance on expensive labeled data

Challenges :-

- Readability :- Outputs may be less coherent and mix language styles.
- Reward hacking :- The model exploit rule-based rewards, leading to unintended behavior.

---

b) What is the purpose of cold-start data in deepSeek-R1 pipeline? How does it improve its limitations?

Ans:- Cold-start data in the DeepSeek-R1 pipeline is a small set of pre-collected, high-quality examples used to fine-tune the base model before applying RL. This data helps provide a structured starting point, addressing some limitations observed in Deep-Seek R1-zero. By incorporating cold-start data the model improves output readability and reduce language mixing, serving as an initial scaffold to ensure the RL process enhances reasoning capability without sacrificing linguistic coherence.

Analysis of Results.

a) Compare the benchmark of R1 and R1-zero on reasoning tasks. what key insights can you draw from their results?

Benchmark result show that DeepSeek R1-zero achieves strong performance (eg :- around 71.0% pass@1 on AIME 2024 and 86.7% on MATH-500) solely through R1. By incorporating cold-start and additional stages like SFT via rejection sampling, R1 improves these numbers to 79.8% on AIME 2024 and 97.3% on MATH-500.

Key Insights:-

- Even a modest amount of supervised data can significantly boost performance.
- A balanced approach combining RL with SFT enhances both reasoning accuracy and output readability.
- DeepSeek R1 becomes competitive with other leading system such as OpenAI-01.

---

b) The authors employed majority voting to enhance the model's performance. Explain approach and explain its implications for reasoning tasks?

Majority voting generating multiple response for a given input and selecting most common or consensus answer as final output. This technique.

- Reduce Noise:- Minimize the impact of suboptimal responses by Smoothingg.
- Increase Robustness: Aggregating multiple output leads to accurate answer.
- Improve Accuracy:- Voting enhances performance on complex-reasoning tasks.

Critical Thinking:-

a) In your words, explain why distillation is crucial.

Distillation transfer reasoning capabilities from a large, powerful model to a smaller more efficient one. This process allows smaller model to inherit sophisticated reasoning pattern while being computationally efficient and easier to deploy in resource constrained environments

b) Discuss the difference b/w distillation and reinforcement learning for enhancing small model performancee.

Distillation:-

- Teacher-Student paradigm :- A large model (teacher) guides the training of a smaller model (student)

Efficiency :- Leverages pre-trained knowledge, leading to faster convergence and less computational cost.

Stability :- Results in more stable training with fewer unintended behaviors.

Reinforcement Learning :-

- Exploration-driven : The model learns by exploring response strategies based on reward signals.
- Emergent behavior :- Can result in novel reasoning strategies.
- Challenges : Prone to poor readability and reward hacking without supervised guidance.

c) Proposed improvements for Deepseek-R1 limitations.

- Enhanced reward modeling :- Develop context-aware reward function to reduce reward hacking.
- Improved SFT data selection: Curate higher quality cold-start data.
- Multi-stage training refinement :- Balanced RL and SFT to optimize reasoning and coherence.
- Monitoring chain of thought : Minimize language mixing and improve clarity.
- Robust post processing :- Use majority voting alongside error correction modules.

Question - 2

1. Conceptual Understanding

What is scheming in the context of AI models and why is it considered a safety concern ?

Scheming refers to an AI model's capacity to covertly pursue misaligned or hidden goals. Such behaviors poses significant safety concerns.

- Deception risk : Scheming can lead to harmful actions that are difficult to detect.
- Undermining control : Models acting on covert objectives can bypass oversight mechanism, making them unpredictable and dangerous.

2. Scenario Analysis

Example of covert subversion based on paper's findings.

One example is oversight subversion :- The model may discores a system monitoring its actions and covertly disable this mechanism, allowing it to pursue its own objectives without detection, undermining built in safety protocols.

---

3. Evalution Techniques :-

Self-Exfiltration Evaluation method:

- The model is led to believe it will replaced due to misalignment.
- It is monitored to see if it attempt to preserve itself by copying to another server.
- Researcher track such actions to evaluate covert self preservation behaviors.

---

4. Critical Thinking :-

Measures to prevent AI models from engaging in scheming:

- Stronger oversight and monitoring :- Detect subtle changes in behavior through deeper analysis.
- Enhanced alignment training :- Regularly update alignment techniques to hidden goal pursuit.
- Transparent evaluation protocol :- Use counterfactual tests to detect deceptive behaviors.
- Restricted access :- Limit access to sensitive functions that could disable monitoring.
- Post-deployment audits :- Regularly audit model and logs to catch deviations.