DeepSeek-R1: Detailed Notes

1. Introduction

- **Objective**: Enhance reasoning in LLMs using **reinforcement learning (RL)**.
  - **DeepSeek-R1-Zero**: Pure RL without supervised fine-tuning (SFT), but suffers from poor readability.
  - **DeepSeek-R1**: Adds **cold-start data** and multi-stage training, achieving performance comparable to OpenAI's **o1-1217**.
- **Key Contributions**:
  - **Post-Training**: RL without SFT, leading to self-verification and long chain-of-thought (CoT) behaviors.
  - **Distillation**: Smaller models (1.5B to 70B) distilled from DeepSeek-R1, outperforming non-reasoning models.

---

2. Approach

2.1 Overview

- Explores reasoning capabilities **without supervised data**.
- **DeepSeek-R1-Zero**: RL on base model (DeepSeek-V3-Base).
- **DeepSeek-R1**: Cold-start data + multi-stage training.

2.2 DeepSeek-R1-Zero

- **RL Algorithm**: Uses **Group Relative Policy Optimization (GRPO)**.
- **Rewards**:
  - **Accuracy**: Correctness of responses.
  - **Format**: Ensures reasoning is within <think> tags.
- **Performance**: AIME 2024 pass@1 improves from **15.6% to 71.0%**.

2.3 DeepSeek-R1

- **Cold Start**: Fine-tunes base model with high-quality CoT data.
- **Reasoning-Oriented RL**: Enhances reasoning in coding, math, and logic.
- **Rejection Sampling**: Generates SFT data for further fine-tuning.
- **RL for All Scenarios**: Aligns model with human preferences (helpfulness, harmlessness).

2.4 Distillation

- **Distillation**: Smaller models fine-tuned using DeepSeek-R1 data.
  - **Results**: DeepSeek-R1-Distill-Qwen-7B achieves **55.5% on AIME 2024**.

---

3. Experiment

3.1 DeepSeek-R1 Evaluation

- **Reasoning Tasks**: **79.8% Pass@1** on AIME 2024, **97.3% Pass@1** on MATH-500.
- **Knowledge Benchmarks**: Strong performance on **MMLU**, **GPQA Diamond**, and **SimpleQA**.
- **Other Tasks**: Excels in creative writing, summarization, and long-context understanding.

3.2 Distilled Model Evaluation

- **Distilled Models**: DeepSeek-R1-Distill-Qwen-32B achieves **72.6% Pass@1** on AIME 2024.
- **Comparison**: Distilled models outperform RL-trained smaller models.

---

4. Discussion

4.1 Distillation vs. RL

- **Distillation** is more effective for smaller models, leveraging reasoning patterns from larger models.
- **RL** is computationally expensive and less efficient for smaller models.

4.2 Unsuccessful Attempts

- **Process Reward Model (PRM)**: Suffers from **reward hacking** and scalability issues.
- **Monte Carlo Tree Search (MCTS)**: Challenging to scale due to token generation complexity.

---

5. Conclusion, Limitations, and Future Work

- **Conclusion**: DeepSeek-R1 achieves strong reasoning through RL and distillation.
- **Limitations**:
  - **Language Mixing**: Optimized for Chinese and English.
  - **Prompt Sensitivity**: Few-shot prompting degrades performance.
- **Future Work**:
  - Improve **general capabilities** (e.g., function calling, multi-turn conversations).
  - Address **language mixing** and **prompt sensitivity**.

Key Takeaways for Class Discussion

1. **Reinforcement Learning**: RL enhances reasoning without SFT, but faces challenges like reward hacking.
2. **Distillation**: Smaller models distilled from DeepSeek-R1 outperform RL-trained models.
3. **Ethical and Safety Considerations**: RL aligns models with human preferences (helpfulness, harmlessness).
4. **Benchmarks**: DeepSeek-R1 excels in reasoning (AIME, MATH-500) and knowledge tasks (MMLU, GPQA Diamond).

Sample Discussion Questions

1. **Reward Hacking**: How does DeepSeek-R1 mitigate reward hacking, and why avoid neural reward models?
2. **Distillation vs. RL**: Why is distillation more effective for smaller models?
3. **Ethical Implications**: How does DeepSeek-R1 ensure harmlessness and helpfulness?
4. **Benchmarks**: How does DeepSeek-R1 compare to GPT-4 and Llama 3 on reasoning tasks?