Understanding DeepSeek: Architecture, Reasoning Capabilities and Training Processes (Hanna's Notes)

‼️ Update May 20th:

I am currently still deep diving into the literature and gathering relevant information. It is too early to draw any conclusions. Below is my literature list so far, organized by areas of interest. I have defined goals for each outcome I aim to achieve and developed a few sub-questions to guide my research more specifically. Key information, including figures and tables, is generally marked in the PDFs listed in the literature table below.

Anticipated Outcomes:

1. Explain how reinforcement learning and chain-of-thought replace standard fine-tuning.
2. Show how the reward systems improve step-by-step reasoning.

Research Question:

How does DeepSeek's pure reinforcement learning strategy, integrated with chain-of-thought prompting, improve reasoning accuracy compared to conventional supervised fine-tuning?

Github

https://github.com/deepseek-ai/DeepSeek-R1

https://github.com/deepseek-ai/DeepSeek-V3

https://github.com/deepseek-ai/DeepSeek-Coder-V2

https://github.com/deepseek-ai/DeepSeek-VL

https://github.com/deepseek-ai/DeepSeek-V2

https://github.com/deepseek-ai/DeepSeek-Coder

https://github.com/deepseek-ai/DeepSeek-Math

📒 **Outcome 1: Explain how reinforcement learning and chain-of-thought replace standard fine-tuning.**

Goal 1: Evaluate Reasoning Accuracy

Goal 2: Analyze the Role of Chain-of-Thought Prompting

Goal 3: Identify Generalization Capabilities

Goal 4: Measure Training Efficiency

Literature

Aa Title	Goal
DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning	
DeepSeek-Prover-V2: Advancing Formal Mathematical Reasoning via Reinforcement Learning for Subgoal Decomposition	
Med-R1: Reinforcement Learning for Generalizable Medical Reasoning in Vision-Language Models	
(How) Do reasoning models reason?	
DeepSeek-Coder: When the Large Language Model MeetsProgramming - The Rise of Code Intelligence	
DeepSeek LLMScaling Open-Source Language Models with Longtermism	

Github

https://github.com/deepseek-ai/DeepSeek-R1

https://github.com/deepseek-ai/DeepSeek-Math

https://github.com/deepseek-ai/DeepSeek-V3

Sub-questions

- How Do Chain-of-Thought and Reinforcement Learning Replace Fine-Tuning?
- How Do Reward Systems Improve Reasoning?

Main Findings (so far)

- Reinforcement Learning (RL) surpasses traditional supervised fine-tuning
  - DeepSeek-R1 and DeepSeek-V2 use RL (e.g. R1-Zero, GRPO) to fine-tune on intermediate reasoning traces → outperforming models trained with only supervised objectives

  > Traditional LLMs use supervised fine-tuning (SFT) on human-written or synthetic Q&A pairs → DeepSeek-R1 and DeepSeek-V2 introduce post-training RL using correctness-based rewards to refine outputs → These models are not trained to mimic reasoning but to generate correct answers via derivational traces → As shown in benchmarks this leads to higher accuracy than SFT-trained counterparts

- Chain-of-Thought (CoT) prompting significantly boosts reasoning accuracy
  - CoT leads to better performance on complex tasks (math, planning) by encouraging structured intermediate outputs (step-by-step reasoning)

  > Instead of outputting direct answers CoT prompts like "Let's think step-by-step" guide the model to generate intermediate steps → DeepSeek-Math and R1 models internally or explicitly produce multi-step traces even during generation → Studies show performance jumps significantly on math and logic tasks with CoT formatting

- Generalization is enhanced by derivational trace learning
  - DeepSeek-R1 generalizes to abstract planning domains → indicating reasoning capability rather than memorization

DeepSeek-R1 is tested on unseen concealed planning domains → Success is attributed to training on intermediate reasoning steps and not just solutions → This demonstrates the model's ability to generalize beyond memorized patterns → a hallmark of reasoning ability

---

📒 **Outcome 2: Show how the reward systems improve step-by-step reasoning.**

Goal 5: Assess Process and Output Quality

Goal 6: Establish the Impact of Reward Modeling

Literature

Aa Title	Goal
DeepSeek-Coder: When the Large Language Model MeetsProgramming - The Rise of Code Intelligence	
DeepSeek-V2: A Strong, Economical, and EfficientMixture-of-Experts Language Model	
DeepSeek-Prover-V1.5: Harnessing Proof Assistant Feedback for Reinforcement Learning and Monte-Carlo Tree Search	

Github

https://github.com/deepseek-ai/DeepSeek-Math

https://github.com/deepseek-ai/DeepSeek-V2

https://github.com/deepseek-ai/DeepSeek-V3

Sub-questions

- What is the purpose of step-by-step Reasoning?
- What are reward systems?
- What types of reward systems does DeepSeek use?
- How do Reward systems change model behavior?
- What effects do reward systems have in DeepSeek models?
- Why do reward systems improve step-by-step reasoning?

Main Findings (so far)

- Reward modeling improves both solution accuracy and reasoning trace quality
  - RL trained with reward signals from correctness (compilers, math solvers) leads to better intermediate steps and final answers in code and math benchmarks

> DeepSeek-Math and DeepSeek-V2 Chat (RL) generate multiple completions and score them using reward models → Correct or more useful completions receive higher rewards → Over time models learn to prioritize intermediate steps that are not just fluent but verifiably correct

- Process-aware rewards (not just final answers) lead to more reliable outputs
  - Models like DeepSeek-Math benefit from program-aided reward signals, scoring higher on logic-intensive benchmarks such as GSM8K and MATH

> Rather than scoring only the final answer, reward models evaluate the entire reasoning process → For example: in program-aided math reasoning each code step is executed and verified → This approach reduces hallucinations and aligns model behavior with logical correctness

- RL alignment increases open-ended reasoning quality (chat, explanation)
  - DeepSeek-V2 Chat (RL) improves on open-ended generation benchmarks (MT-Bench, AlpacaEval 2.0) by using multi-reward feedback from helpfulness, safety and rule compliance models
  - DeepSeek-V2 Chat (RL) is trained with feedback from:
    - Helpfulness reward model
    - Safety reward model
    - Rule-compliance reward model
  - This multi-objective RL leads to:
    - More factual well-structured answers in chat
    - Higher scores on MT-Bench and AlpacaEval 2.0
  - Users benefit from outputs that are not only correct but also aligned with human expectations