How DeepSeek Works

🧠 Architecture: Mixture-of-Experts (MoE) Model

DeepSeek utilizes a **Mixture-of-Experts (MoE)** architecture, which means:

- **Sparse Activation**: Instead of activating all neurons for every task, only a subset (experts) is engaged, making computations more efficient.
- **Specialized Experts**: Each expert specializes in specific tasks, enhancing performance without increasing overall model size.

This approach allows DeepSeek to achieve high performance with reduced computational resources.

🛠️ Training Methodology: Reinforcement Learning (RL)

DeepSeek employs **Reinforcement Learning (RL)**, a technique where the model learns by trial and error:

1. **Initial Fine-Tuning**: The model undergoes supervised fine-tuning on curated datasets to establish a foundational understanding.
2. **Reinforcement Learning Phases**: The model interacts with tasks, receives feedback (rewards or penalties), and adjusts its behavior to maximize positive outcomes.
3. **Rejection Sampling**: Incorrect or suboptimal responses are filtered out, ensuring the model learns from high-quality data.

This methodology enhances DeepSeek's reasoning and problem-solving capabilities.

🔄 Distillation Techniques: Creating Efficient Models

DeepSeek uses **distillation**, a process where a large "teacher" model trains a smaller "student" model:

- **Knowledge Transfer**: The student model learns to replicate the teacher's behavior, capturing essential features with fewer parameters.
- **Efficiency**: This results in models that are faster and require less computational power, making them suitable for various applications.

Distillation enables DeepSeek to offer powerful AI solutions that are both cost-effective and efficient.

⚙️ Key Components of DeepSeek's Functionality

- **Multi-Head Latent Attention (MLA)**: An advanced attention mechanism that compresses information, allowing the model to focus on relevant data efficiently.
- **Extended Context Length**: DeepSeek can process longer sequences of text, up to 128K tokens, enabling it to understand and generate more coherent and contextually accurate responses.

These components contribute to DeepSeek's ability to handle complex tasks effectively.

-notes part of the Udemy course (Instructor: Bibhatsu Kuiri)