



DeepSeek DeepDive



<https://arxiv.org/abs/2501.12948>

1. Introduction to the Model(s) - High level overviews

1.1 Evolution of Reasoning Models

1.2 Objective and Methodology

1.3 Training Approach

1.4 Model Variants

1.5 Distillation and Performance

1.6 Open Source and Accessibility

2. DeepSeek-V3 Overview

2.1 Introduction

2.2 Model Architecture

2.3 Training Methodology

2.4 Key Features

2.5 Cost and Efficiency

2.6 Comparison and Impact

3. DeepSeek-R1-Zero: From Scratch to Emergence

<u>3.1 Overview</u>
<u>3.2 Training Methodology</u>
<u>3.2.1 Pure Reinforcement Learning</u>
<u>3.2.2 Training Template</u>
<u>3.3 Performance and Evaluation</u>
<u>3.3.1 Benchmark Comparisons</u>
<u>3.3.2 Emergent Capabilities</u>
<u>3.4 Inference and Efficiency</u>
<u>3.5 Emergent Behaviors</u>
<u>4. DeepSeek-R1: A Multi-Stage Evolution</u>
<u>4.1 Overview</u>
<u>4.2 Key Challenges with R1-Zero</u>
<u>4.3 Development Stages of DeepSeek-R1</u>
<u>4.3.1 Stage 1: Cold Start Training</u>
<u>4.3.2 Stage 2: Reasoning-Based Reinforcement Learning</u>
<u>4.3.3 Stage 3: Rejection Sampling</u>
<u>4.3.4 Stage 4: Final Reinforcement Learning</u>
<u>4.4 Emergent Capabilities</u>
<u>5. Evaluating Performance & Benchmarks</u>
<u>6. Knowledge Distillation: Enhancing Reasoning in Llama 3.3 & Qwen, etc</u>
<u>7. Future Directions & Ongoing Research</u>
<u>7.1 Current Limitations</u>
<u>7.2 Future Research Directions</u>

1. Introduction to the Model(s) - High level overviews

1.1 Evolution of Reasoning Models



The landscape of model training is undergoing a significant transformation.

- Traditional approaches focused on scaling models by increasing their size and computational power.
- However, recent advancements have shifted the paradigm towards optimizing inference time compute, trading off upfront costs for more efficient performance.

The development of reasoning models has reached a pivotal point. Initially spearheaded by OpenAI's o1 Mini, the field has seen rapid advancements with new models emerging.

DeepSeek's latest release, for instance, rivals OpenAI's offerings and is completely open-sourced under an MIT license. Despite the limited details in their publication, the model's performance is noteworthy, highlighting the potential for further enhancements.

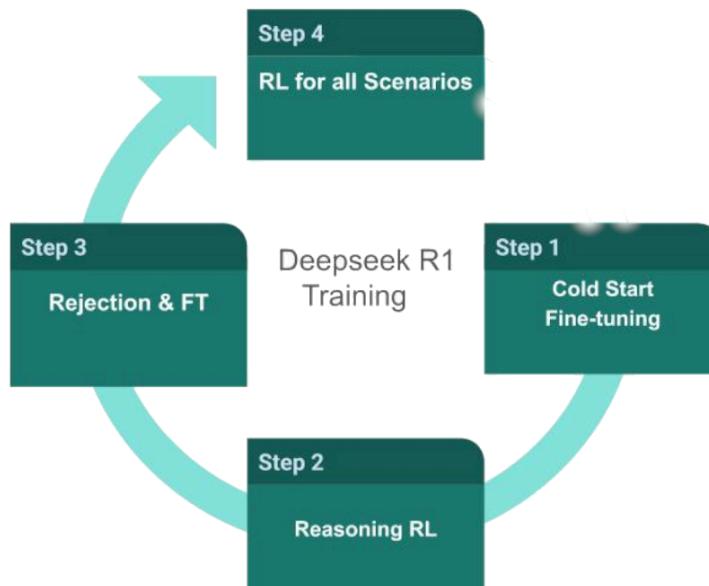
1.2 Objective and Methodology

DeepSeek's primary goal is to harness the potential of Large Language Models (LLMs) to develop reasoning capabilities without relying on supervised data. This is achieved through a pure Reinforcement Learning (RL) process. By employing RL, they discovered that models could achieve high performance without traditional supervised data, focusing instead on self-evolution and reflection.

1.3 Training Approach

The training methodology for DeepSeek's models is innovative and multi-staged:

1. **Post-Training with Group Relative Policy Optimization (GRPO):** The base model, DeepSeek-V3, undergoes post-training with GRPO, leading to emergent reasoning capabilities and enhanced performance.
2. **Four-Stage Training for R1:**
 - **Cold Start:** Initial phase to stabilize the model.
 - **RL for Reasoning:** Focused training to enhance reasoning skills.
 - **Rejection Sampling:** Applied for general-purpose refinement.
 - **RL Polishing:** Final stage to fine-tune the model's capabilities.



1.4 Model Variants

DeepSeek introduces two primary models:

- **R1-Zero:** A specialized reasoning model trained on Chain of Thought (CoT) using RL, excelling in reasoning but lacking general model capabilities.
- **R1:** Developed from R1-Zero outputs and the four-stage training process, R1 is a robust reasoning and chat model.

1.5 Distillation and Performance

Outputs from R1 are distilled into Qwen and Llama models, demonstrating exceptional performance even without native RL training. Distillation proves more effective than traditional RL post-training, offering a computationally efficient alternative.

1.6 Open Source and Accessibility

DeepSeek's models are fully open-sourced under the MIT license, though training data and code remain undisclosed. The models are hosted on DeepSeek's API, which boasts speeds 3 to 10 times faster and more cost-effective than other providers, albeit with considerations for data privacy.

2. DeepSeek-V3 Overview



2.1 Introduction

DeepSeek-V3 represents a significant advancement in the field of large language models (LLMs). As a foundational model, it serves as the base for subsequent reasoning models, setting a new standard in terms of efficiency, scalability, and performance. This section provides a detailed overview of DeepSeek-V3, highlighting its architecture, training methodology, and key features.

2.2 Model Architecture

DeepSeek-V3 is a large-scale language model, characterized by its substantial parameter count. It boasts a total of 671 billion parameters, with 37 billion active parameters during inference. This unique architecture allows the model to maintain a high level of performance while optimizing resource usage.

- **Total Parameters:** 671 billion
- **Active Parameters:** 37 billion

The model leverages a multi-expert system, which enables it to function effectively as a 37 billion parameter model at inference time. This design choice significantly reduces computational demands while maintaining the model's capabilities.

2.3 Training Methodology

The training of DeepSeek-V3 involved several strategic steps to enhance its performance and efficiency:

1. **Base Model Training:** The initial training phase established the foundational capabilities of the model.
2. **Context Length Extension:** The model underwent two stages of context length extension, first to 32k and then to 128k, allowing it to handle longer sequences effectively.
3. **Token Training:** Trained on approximately 15 trillion tokens, the model's training process incorporated standard techniques such as supervised fine-tuning (SFT) and reinforcement learning (RL).

2.4 Key Features

DeepSeek-V3 introduces several innovative features that set it apart from other models:

- **Multi-Head Latent Attention:** This advanced attention mechanism enhances the model's ability to focus on relevant information, improving its overall performance.
- **Multi-Token Prediction:** By adopting a more sample-efficient approach, DeepSeek-V3 can predict multiple tokens simultaneously, which contributes to its efficiency and speed.

2.5 Cost and Efficiency

One of the most notable claims about DeepSeek-V3 is its cost-effectiveness. The model was reportedly trained at a cost of \$5.5 million, a figure that has generated considerable attention and discussion within the AI community. This cost efficiency is attributed to several factors, including the use of mixed precision training and the absence of auxiliary loss during training.

2.6 Comparison and Impact

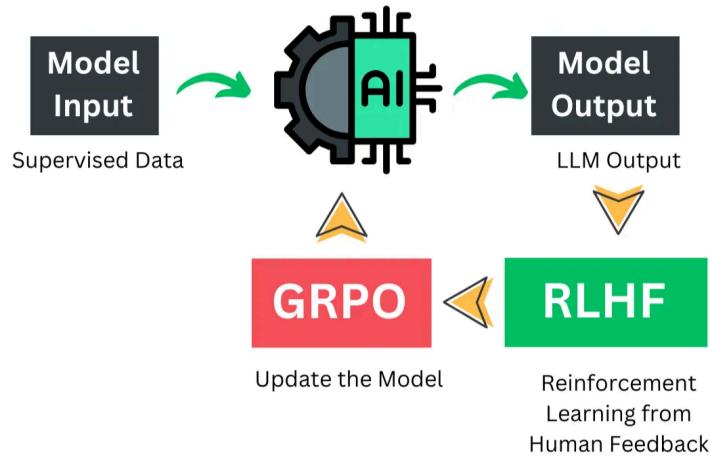
DeepSeek-V3 is often compared to models like GPT-4o and Mitral, positioning itself as a competitive alternative in the landscape of LLMs. Its ability to deliver high performance with reduced active parameters makes it a desirable choice for applications where speed and cost are critical considerations.

3. DeepSeek-R1-Zero: From Scratch to Emergence

3.1 Overview

DeepSeek-R1-Zero utilizes pure reinforcement learning (RL) without any supervised fine-tuning (SFT) data. This section explores the innovative methodologies and emergent capabilities of DeepSeek-R1-Zero, highlighting its unique training processes and performance metrics.

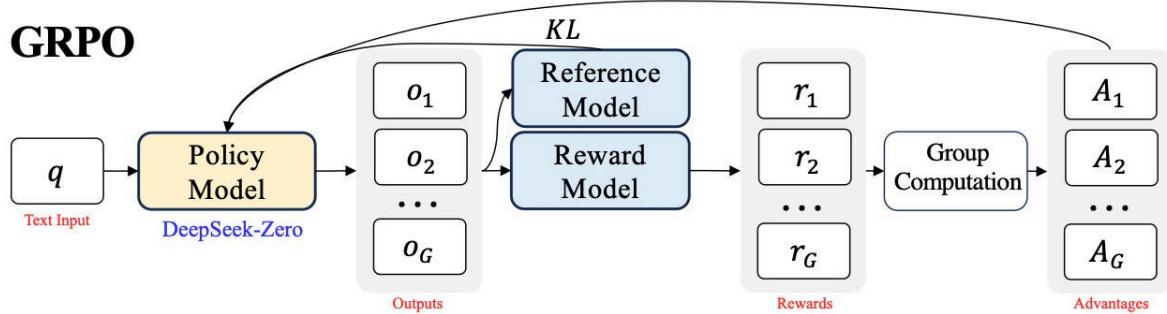
3.2 Training Methodology



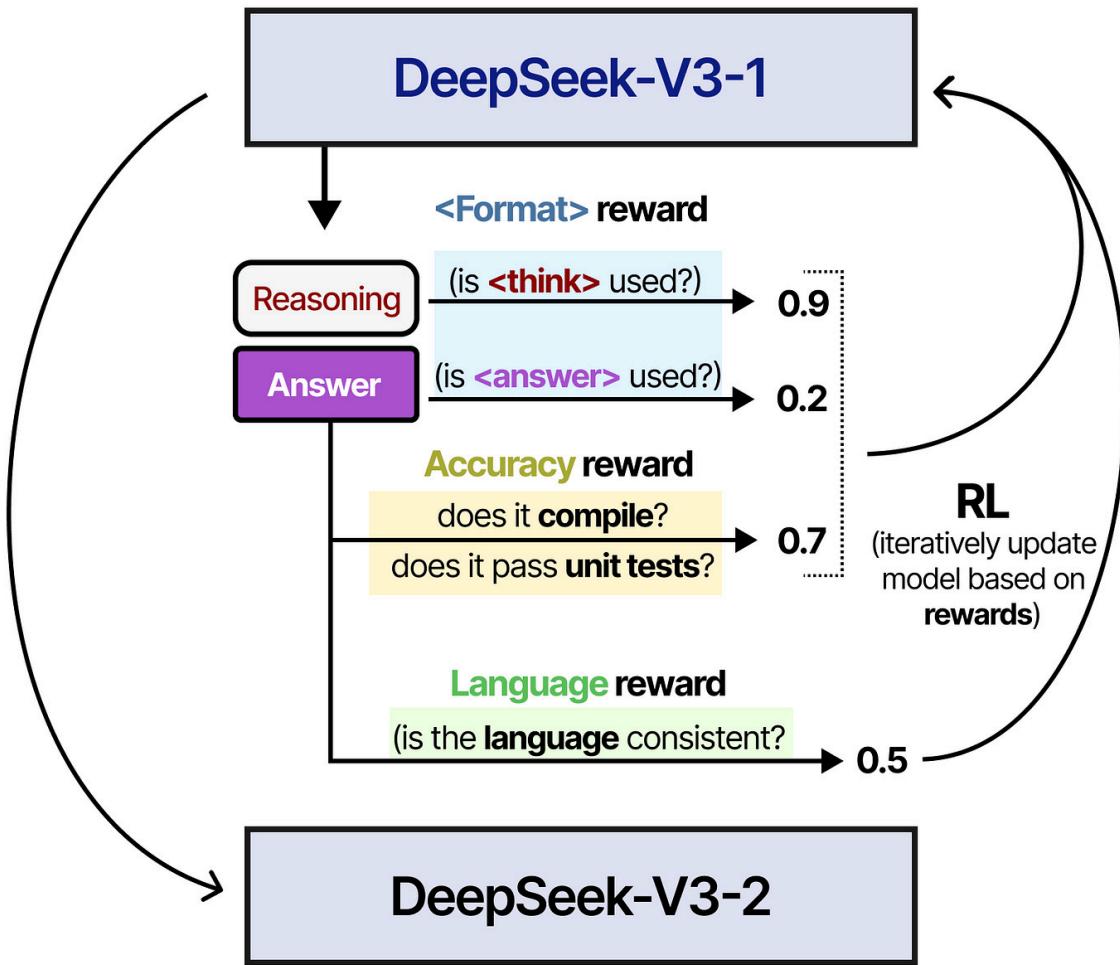
3.2.1 Pure Reinforcement Learning

DeepSeek-R1-Zero applies pure RL directly to a V3-base model, bypassing the traditional SFT data. This approach leverages the GRPO (Group-based Policy Optimization) algorithm, introduced in the DeepSeek Math paper, to optimize model performance.

- GRPO Algorithm:** Unlike traditional RL methods, GRPO foregoes the value model, instead using group-based sample generation to estimate rewards. This reduces computational costs and enhances efficiency.



- Reward Mechanism:** The model is rewarded based on accuracy and format adherence. Responses must be verifiably accurate, such as solving math problems or compiling code correctly. Additionally, format rewards ensure structured output using <think> and <answer> tags.



3.2.2 Training Template

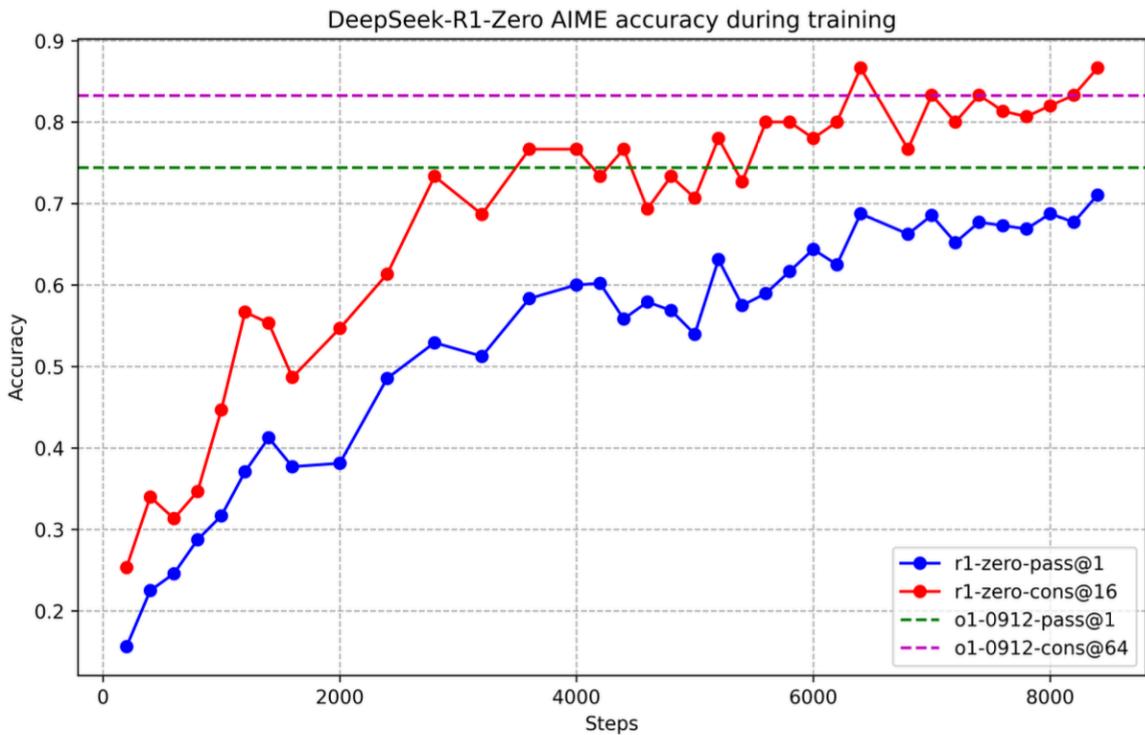
The training process involves a structured conversational template where the assistant processes user queries, reasons through problems, and provides answers within specified tags. This structure facilitates clear reasoning and accurate responses.

3.3 Performance and Evaluation

3.3.1 Benchmark Comparisons

DeepSeek-R1-Zero exhibits exceptional reasoning capabilities without relying on labeled SFT data. It performs well across various benchmarks, including math and coding challenges, often surpassing models like OpenAI-o1.

- **Majority Voting:** The model's accuracy improves significantly with majority voting, where multiple response samples are generated and the most accurate is selected.



3.3.2 Emergent Capabilities

As training progresses, DeepSeek-R1-Zero demonstrates emergent behaviors, such as:

- **Reflections:** The model revisits and reevaluates previous steps, exploring alternative solutions independently.
- **Aha Moments:** It autonomously identifies and corrects its reasoning paths, showcasing advanced problem-solving strategies.

Question: If $a > 1$, then the sum of the real solutions of $\sqrt{a - \sqrt{a+x}} = x$ is equal to

Response: <think>

To solve the equation $\sqrt{a - \sqrt{a+x}} = x$, let's start by squaring both ...

$$(\sqrt{a - \sqrt{a+x}})^2 = x^2 \implies a - \sqrt{a+x} = x^2.$$

Rearrange to isolate the inner square root term:

$$(a - x^2)^2 = a + x \implies a^2 - 2ax^2 + (x^2)^2 = a + x \implies x^4 - 2ax^2 - x + (a^2 - a) = 0$$

...

Wait, wait. Wait. That's an aha moment I can flag here.

Let's reevaluate this step-by-step to identify if the correct sum can be ...

We started with the equation:

$$\sqrt{a - \sqrt{a+x}} = x$$

First, let's square both sides:

$$a - \sqrt{a+x} = x^2 \implies \sqrt{a+x} = a - x^2$$

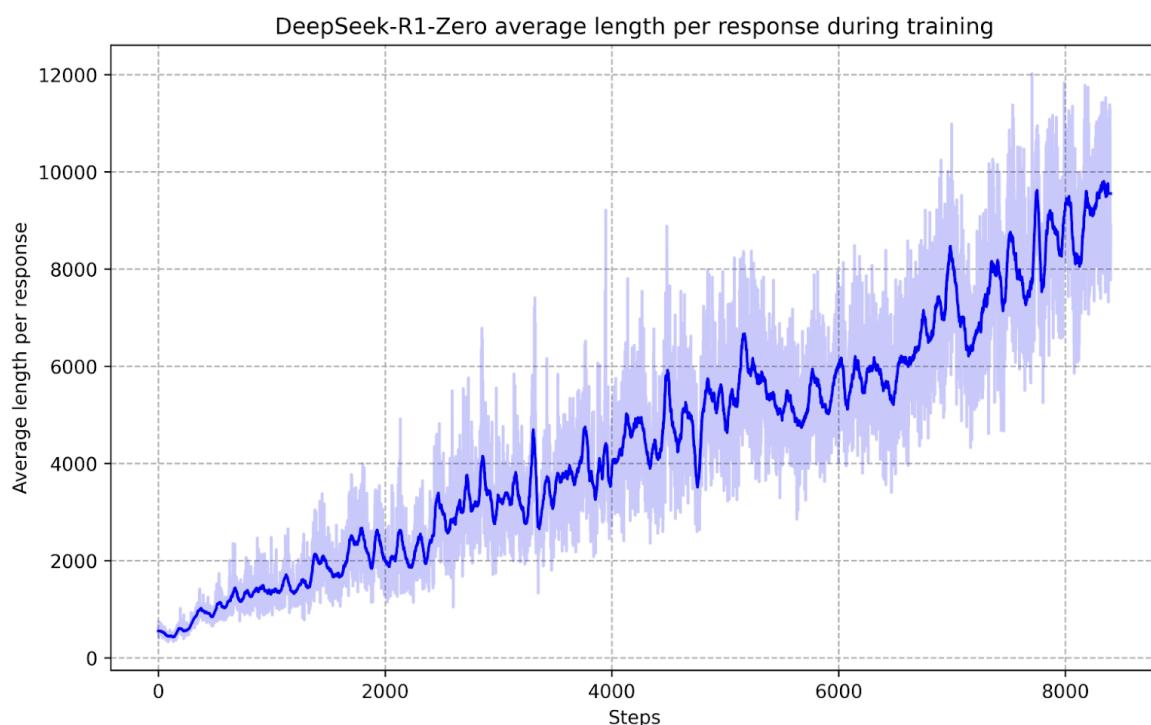
Next, I could square both sides again, treating the equation: ...

...

Table 3 | An interesting “aha moment” of an intermediate version of DeepSeek-R1-Zero. The model learns to rethink using an anthropomorphic tone. This is also an aha moment for us, allowing us to witness the power and beauty of reinforcement learning.

3.4 Inference and Efficiency

Charts indicate that inference time correlates with evaluation performance. As the model undergoes more training steps, its reasoning depth and response length increase, enhancing its ability to tackle complex tasks.



3.5 Emergent Behaviors

DeepSeek-R1-Zero naturally evolves its problem-solving abilities by extending test time compute. This ranges from hundreds to thousands of reasoning tokens, leading to the emergence of interesting behaviors as test time compute increases.

- **Reflections:** The model re-evaluates previous steps, exploring alternatives spontaneously.
- **Aha Moments:** It takes more time to think by reevaluating its approach, leading to breakthroughs in problem-solving.

DeepSeek-R1-Zero exemplifies the potential of RL to unlock new levels of intelligence in artificial systems. By providing the right incentives, the model autonomously develops advanced problem-solving strategies, paving the way for more autonomous and adaptive models in the future.

4. DeepSeek-R1: A Multi-Stage Evolution

4.1 Overview

DeepSeek-R1 is addressing key issues present in its predecessor, R1-Zero. This section delves into the multi-stage process employed to enhance readability, language consistency, and overall functionality, transforming DeepSeek-R1 into a robust reasoning and chat model.

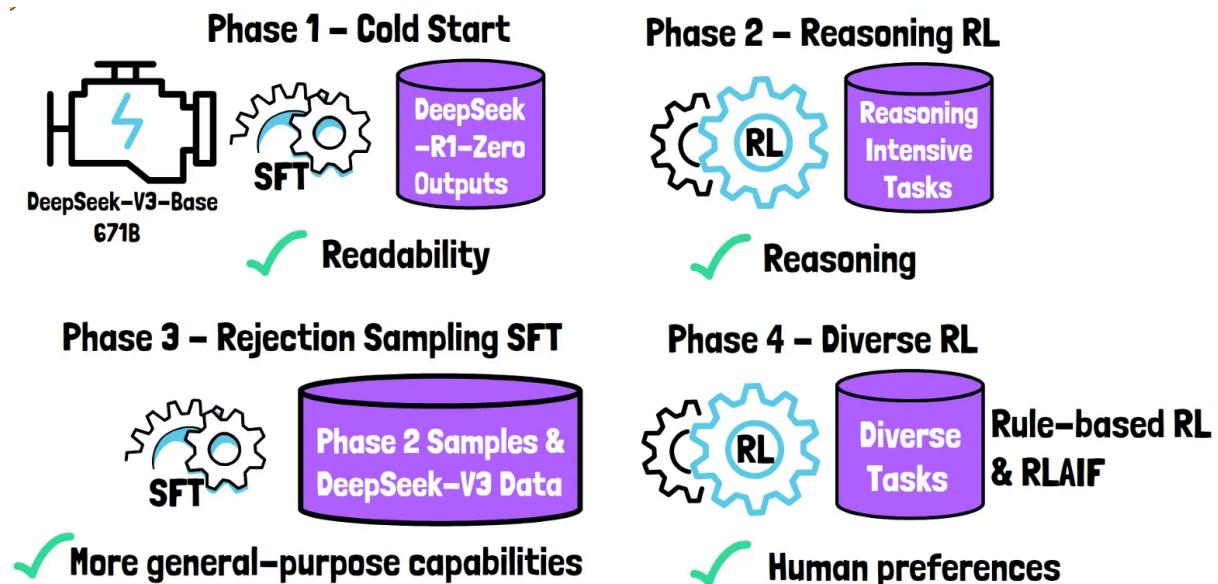
4.2 Key Challenges with R1-Zero

R1-Zero faced several challenges, including:

- **Poor Readability:** The model often produced outputs that were difficult to follow.
- **Language Mixing:** It frequently switched between languages, such as English and Chinese, without maintaining consistency.

These issues were partially attributed to the reinforcement learning (RL) approach used, which lacked specific objectives related to safety, conciseness, and user engagement.

4.3 Development Stages of DeepSeek-R1



4.3.1 Stage 1: Cold Start Training

The initial stage focused on stabilizing the model through strong supervised fine-tuning (SFT):

- **Objective:** Prevent the model from becoming unstable by providing a solid foundation.
- **Methodology:**
 - Utilized a long Chain of Thought (CoT) few-shot prompt to generate detailed answers.
 - Employed human annotators to refine the generated examples from R1-Zero.
 - Focused on producing examples with clear reasoning, reflection, and verification.
 - Processed thousands of examples to ensure quality.

4.3.2 Stage 2: Reasoning-Based Reinforcement Learning

Building on the cold start, the second stage involved reasoning-based RL:

- **Objective:** Enhance language consistency and reasoning capabilities.
- **Methodology:**

- Applied the same RL process as R1-Zero, with an added focus on maintaining language consistency.

4.3.3 Stage 3: Rejection Sampling

Rejection sampling was introduced to further refine the model:

- **Objective:** Improve response quality by selecting the best outputs.
- **Methodology:**
 - Generated multiple completions and ranked them using a reward model.
 - Fine-tuned the original model based on these rankings.
 - Processed 800k completions, including 600k reasoning tasks and 200k general chat problems.

4.3.4 Stage 4: Final Reinforcement Learning

The final stage aimed to optimize the model for general use:

- **Objective:** Create a helpful, harmless model with strong reasoning abilities.
- **Methodology:**
 - Incorporated R1-Zero style questions for reasoning tasks.
 - Captured human preferences in nuanced scenarios for general chat.
 - Ensured outputs included summaries and clear thinking steps.

4.4 Emergent Capabilities

DeepSeek-R1 successfully overcame the limitations of its predecessor, achieving:

- **Improved Readability:** Outputs are now more coherent and user-friendly.
- **Language Consistency:** The model maintains a consistent language throughout interactions.
- **Enhanced Functionality:** It performs well in both reasoning tasks and general chat scenarios.

5. Evaluating Performance & Benchmarks

DeepSeek-R1 has undergone rigorous testing to evaluate its effectiveness as a chat model. Here are the key findings:

- **Model Capabilities:** DeepSeek-R1 excels in providing detailed thinking steps and summaries in its responses. It demonstrates significant improvement over its predecessors, outperforming models like o1 mini and being competitive with o1.
- **Language Consistency:** While DeepSeek-R1 generally performs well, it occasionally struggles with language swaps. However, it consistently surpasses o1 mini and often matches or exceeds o1 in various benchmarks.
- **Benchmarking Results:** The model's performance was evaluated across several benchmarks, including MMLU, DROP, and English IF-Eval. DeepSeek-R1 consistently achieved high scores, reflecting its robust reasoning and language capabilities.
- **Model Specifications:** With 37 billion active parameters, DeepSeek-R1 offers a balance of performance and efficiency. Its MIT license ensures accessibility and ease of deployment.

Benchmark (Metric)	Claude-3.5-Sonnet-1022	GPT-4o-0513	DeepSeek-V3	OpenAI o1-mini	OpenAI o1-1217	DeepSeek R1
Architecture	-	-	MoE	-	-	MoE
# Activated Params	-	-	37B	-	-	37B
# Total Params	-	-	671B	-	-	671B
MMLU (Pass@1)	88.3	87.2	88.5	85.2	91.8	90.8
MMLU-Redux (EM)	88.9	88.0	89.1	86.7	-	92.9
MMLU-Pro (EM)	78.0	72.6	75.9	80.3	-	84.0
DROP (3-shot F1)	88.3	83.7	91.6	83.9	90.2	92.2
English	IF-Eval (Prompt Strict)	86.5	84.3	86.1	84.8	-
	GPQA Diamond (Pass@1)	65.0	49.9	59.1	60.0	75.7
	SimpleQA (Correct)	28.4	38.2	24.9	7.0	47.0
	FRAMES (Acc.)	72.5	80.5	73.3	76.9	-
	AlpacaEval2.0 (LC-winrate)	52.0	51.1	70.0	57.8	-
	ArenaHard (GPT-4-1106)	85.2	80.4	85.5	92.0	-
	LiveCodeBench (Pass@1-COT)	38.9	32.9	36.2	53.8	65.9
	Codeforces (Percentile)	20.3	23.6	58.7	93.4	96.6
	Codeforces (Rating)	717	759	1134	1820	2061
	SWE Verified (Resolved)	50.8	38.8	42.0	41.6	49.2
	Aider-Polyglot (Acc.)	45.3	16.0	49.6	32.9	61.7
Math	AIME 2024 (Pass@1)	16.0	9.3	39.2	63.6	79.8
	MATH-500 (Pass@1)	78.3	74.6	90.2	90.0	97.3
	CNMO 2024 (Pass@1)	13.1	10.8	43.2	67.6	-
Chinese	CLUEWSC (EM)	85.4	87.9	90.9	89.9	-
	C-Eval (EM)	76.7	76.0	86.5	68.9	-
	C-SimpleQA (Correct)	55.4	58.7	68.0	40.3	-

Table 4 | Comparison between DeepSeek-R1 and other representative models.

6. Knowledge Distillation: Enhancing Reasoning in Llama 3.3 & Qwen, etc

deepseek-ai/DeepSeek-R1-Distill-Llama-70B
Text Generation • Updated 8 days ago • ↓ 220k • 505

deepseek-ai/DeepSeek-R1-Distill-Llama-8B
Text Generation • Updated 8 days ago • ↓ 390k • 466

deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B
Text Generation • Updated 8 days ago • ↓ 569k • 726

deepseek-ai/DeepSeek-R1-Distill-Qwen-14B
Text Generation • Updated 8 days ago • ↓ 238k • 349

deepseek-ai/DeepSeek-R1-Distill-Qwen-32B
Text Generation • Updated 8 days ago • ↓ 479k • 942

deepseek-ai/DeepSeek-R1-Distill-Qwen-7B
Text Generation • Updated 8 days ago • ↓ 411k • 370

The distillation process played a crucial role in enhancing the reasoning capabilities of Llama 3.3 and Qwen models. Here's how it was implemented:

- **Distillation Process:** Outputs from DeepSeek-R1 were distilled into Llama and Qwen models using 800,000 reasoning samples. This process did not involve reinforcement learning (RL) but focused on basic supervised fine-tuning (SFT).
- **Model Performance:** The distilled models demonstrated significant improvements, with enhanced reasoning traces and overall performance. They were compared against base models and other industry-leading models like GPT-4o and Cloud Sonic.
- **Open Source Availability:** The distilled models, including Llama 8B, Llama 70B, and Qwen 32B, have been open-sourced. These models are optimized for local deployment, offering strong performance with efficient resource usage.

7. Future Directions & Ongoing Research

7.1 Current Limitations

Despite its strengths, DeepSeek-R1 has several areas where it underperforms compared to V3:

- **Functionality and Usability:** R1 is less efficient in function calling, multi-turn interactions, complex role play scenarios, and JSON output handling.
- **Language Mixing:** R1 struggles with mixing languages, particularly when responding in a different language than the one used in the query. This is not observed in V3, suggesting possible differences in training data or methodologies.
- **Sensitivity to Prompting:** R1 is highly sensitive to prompting styles. Few-shot prompts tend to degrade its performance. Users are advised to specify desired outcomes directly rather than guiding the model's reasoning process.
- **Engineering Tasks:** R1 does not significantly outperform V3 in engineering-related tasks, indicating room for improvement in this area.

7.2 Future Research Directions

To address these limitations, ongoing research is focusing on:

- **Enhancing Functionality:** Improving R1's ability to handle complex interactions and outputs, aligning it more closely with V3's capabilities.
- **Refining Language Handling:** Investigating the causes of language mixing issues and exploring solutions to enhance R1's multilingual capabilities.
- **Optimizing Prompting Techniques:** Developing strategies to reduce sensitivity to prompting and improve performance without compromising reasoning abilities.
- **Advancing Engineering Applications:** Enhancing R1's proficiency in engineering tasks to make it a more versatile tool for technical applications.