Inside DeepSeek-R1

How Reasoning LLMs Are Born

@shivanivirdi

NEOSAGE

The Problem with Standard Training

trained models to imitate reasoning, not to execute it.

Supervised learning on millions of examples teaches pattern matching.

- The model's ability is bounded by the quality and scope of human data.
- It appears to reason, but fails when problems deviate from its training examples.

@shivanivirdi

NEOSAGE

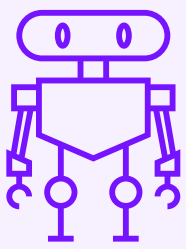
DeepSeek-R1 Flipped the Script

It wasn't taught reasoning with labelled examples.
It was incentivised to discover it via reinforcement.

Typical LLMs

Pretraining

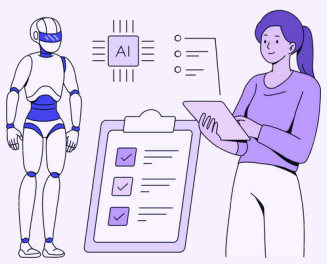
Base Model: a raw, unaligned simulator of internet patterns



Supervised Fine-Tuning (Instruction Tuning)

Trained on human-written assistant responses

RLHF (Preference-Based)



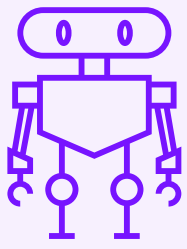
Reward model trained on human preferences

Inference Alignment

Supervised

DeepSeek-R1-Zero

Base Pretraining



Base Model: a raw, unaligned simulator of internet patterns

GRPO Reinforcement Learning

Relative reward without a critic. Learns by comparing completions, not scoring them absolutely.

Emergent Reasoning

Not taught. Not demonstrated. Just incentivised — and discovered.

Structured Format

Learned to reason without ever being shown how

Incentivised

@shivanivirdi

NEOSAGE

The Engine

Group Relative Policy Optimization (GRPO)

A cheaper, more stable alternative to PPO/RLHF.

- **No Critic Model**: Drops the expensive second network.
- **Group Comparison**: Samples multiple outputs for a prompt and scores them with a programmatic reward (e.g., for accuracy).
- **Advantage Calculation**: Reinforces outputs that score higher than the group average.

@shivanivirdi

NEOSAGE

The Engine

GRPO: Reinforcement Without a Critic

Trains by comparing completions within a group

1 Sample responses

Reasoning based question, eg: “What’s (3x + 2)^2 = 49?”

Completion 1:
<Think>detailed reasoning<\Think>
<answer>x1=-3, x2=1.66667 <\answer>

Completion 2:
<Think>reasoned a little<\Think>
<answer>x1=-3, x2=1.66667 <\answer>

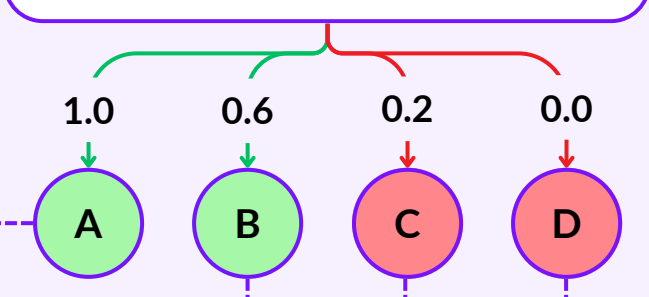
Completion 3:
<Think>reasoned a little<\Think>
<answer>x1=-2, x2=2 <\answer>

Completion 4:
<answer>x1=-2, x2=2 <\answer>

2 Score using reward function

Scoring dimensions:

- Accuracy (math answer correct)
- Format (e.g., uses <think>/<answer> as applicable)



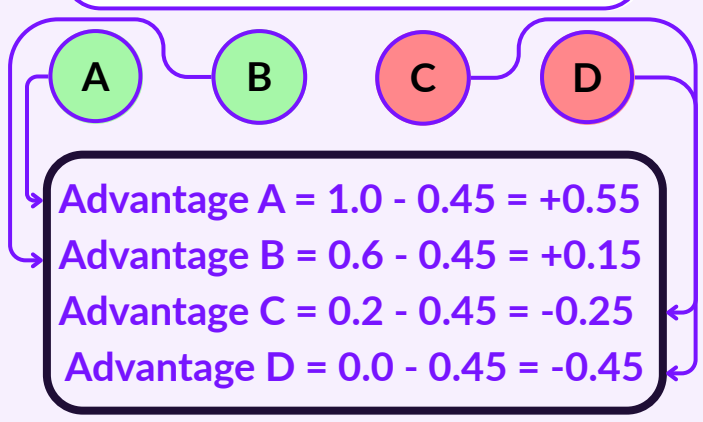
3 Calculate Group Average

Compute group reward baseline

$$\bar{R} = (1.0 + 0.6 + 0.2 + 0.0) / 4 = 0.45$$

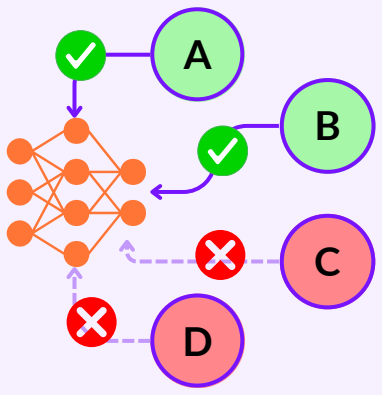
4 Compute Advantage

Compare each score to the group average



5 Policy Update

Reinforce higher-advantage responses



Model updates to increase likelihood of better-than-average responses

@shivanivirdi

NEOSAGE

From R1-Zero to R1

R1 zero had emergent reasoning, but a raw reasoner isn't a deployable system.

You need a structured, generalized, and aligned model. DeepSeek built it in four sequential stages.

DeepSeek-R1: A Four-Stage System to Sculpt Reasoning into Alignment

Cold-Start Fine-Tuning

Imput: DeepSeek R1 Zero Model

Why: R1-Zero outputs were verbose, unstructured, and mixed languages

What: Curated small dataset (few-shot, zero-shot, manually cleaned R1-Zero outputs)

fixed: Readability, Language drift, Structural inconsistency

Reinforcement Learning

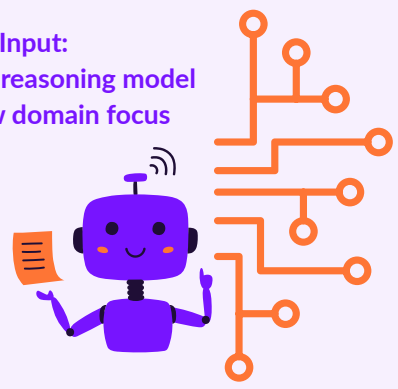
Input:
- Stable outputs from Cold-Start SFT
- Still underperforming on accuracy
- Residual multilingual mixing

What: Reinforcement using accuracy + language consistency rewards

fixed: Reasoning depth, Consistent monolingual output, Reward-based refinement without breaking format

Rejection Sampling + SFT

Input:
- Strong reasoning model
- Narrow domain focus



What: Combined ~600K self-generated reasoning examples (via rejection sampling) with ~200K general-task data for supervised fine-tuning.

fixed: Added domain generalisation, Preserved structured reasoning

RL for Alignment

Input:
- Generalised, strong reasoning model
- Still not aligned in tone, safety, helpfulness

What: Reinforced with reward models (helpfulness, harmlessness), Prompt sampling from diverse distributions

fixed: Instruction quality, Tone alignment, Safety handling in general tasks

@shivanivirdi

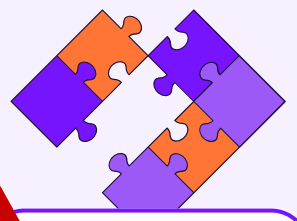
NEOSAGE

Cold Start: How DeepSeek Engineered a Stable Launch Point for RL

What Was Broken in R1-Zero?

Mixed English + Chinese

Poor readability



No human priors → unstable RL convergence

No summarised answers

Hard to train from this. Needed a human-prior stabilised base.

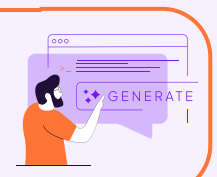
How Cold-Start Dataset Was Constructed

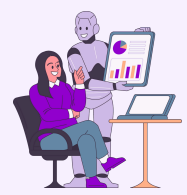
Few-shot prompting

Long CoT examples

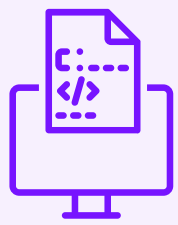
Readable R1-Zero outputs

Post-processed manually





Clean summaries



Direct prompting

Detailed answer + self-checks

Human-refined completions

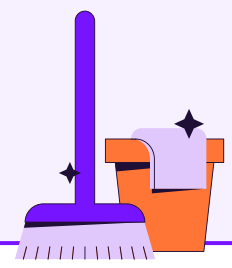
Enforced Output Pattern

Response format

<reasoning_process>detailed CoT style output<\reasoning_process>
<summary>Human readable summary<\summary>

- *Reasoning stays structured*
- *Summary is extractable*
- *Markdown- and user-friendly*

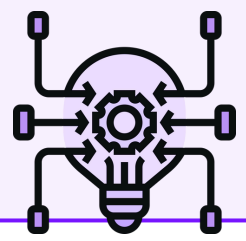
What Cold Start Fixed



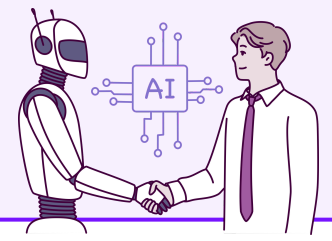
Clean rewardable outputs

Summary extractable → scoreable

Reduced language drift



Enabled RL to converge faster



Human priors gave strong inductive bias

@shivanivirdi

NEOSAGE

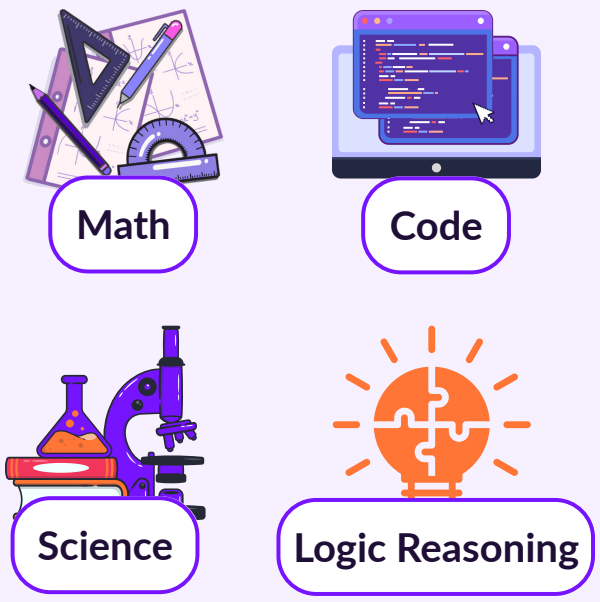
Reasoning-Oriented Reinforcement Learning

Problem: The format was stable, but reasoning accuracy could still be improved and language mixing persisted.

Solution: Applied GRPO with a composite reward function, optimizing for both task accuracy and language consistency simultaneously.

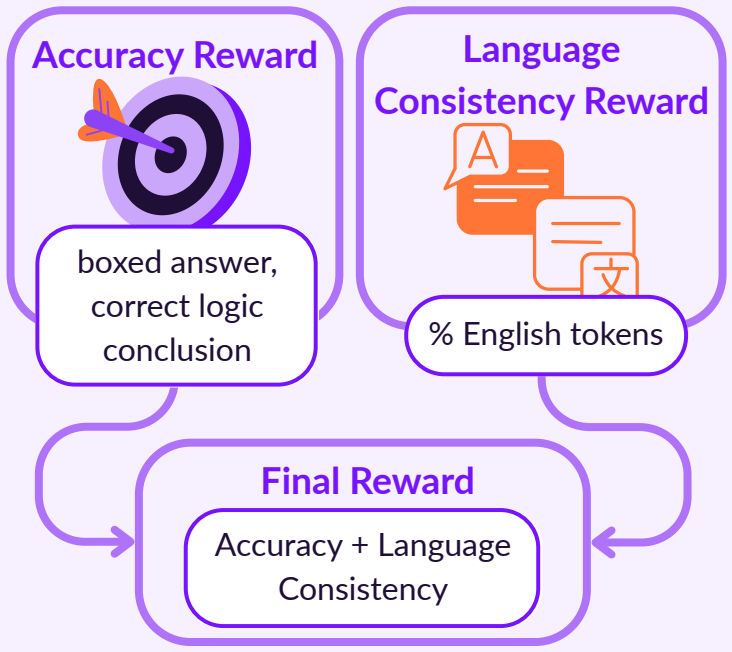
Reinforcing Reasoning While Controlling for Language Drift

Focused Task Domains



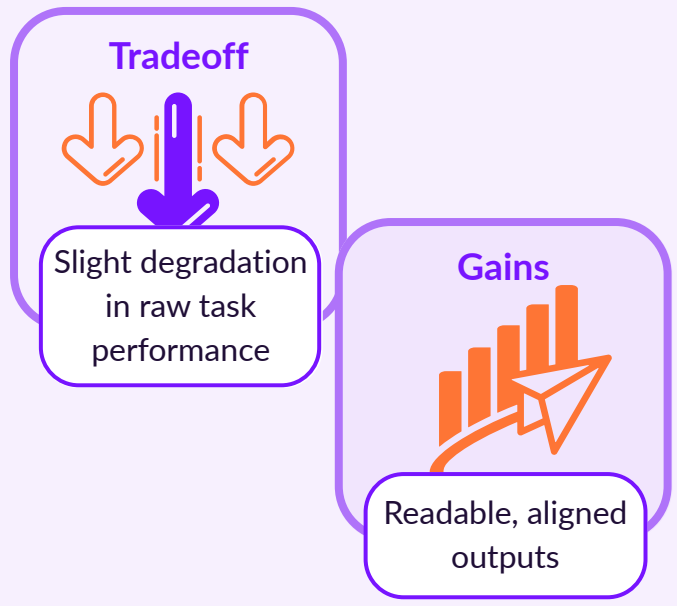
All tasks had clearly verifiable final answers — ideal for reinforcement

Reward Function Changes



Language consistency = % of target-language tokens in CoT

Tradeoff + Outcome



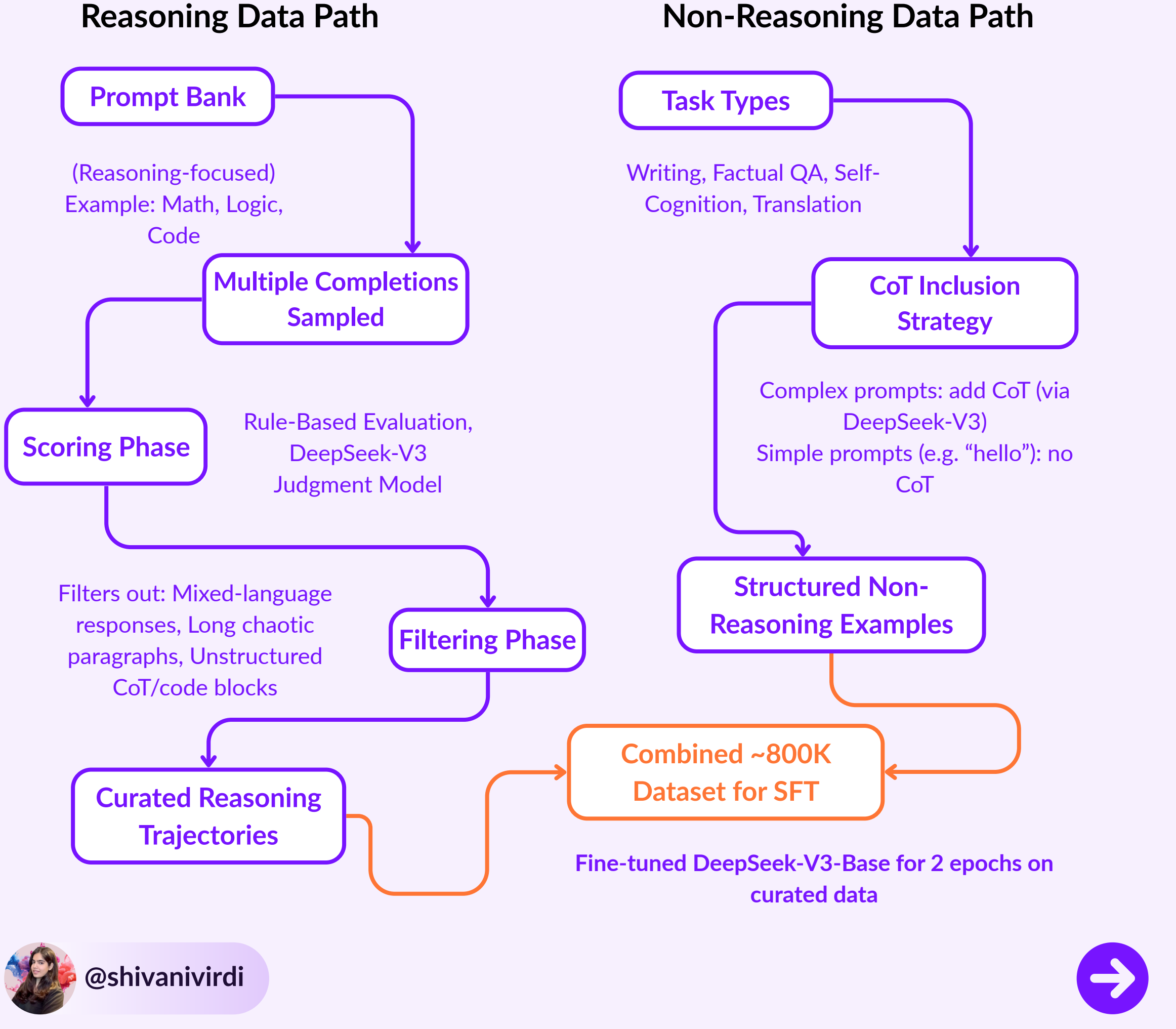
This stage didn't change how the model learned it changed what it was rewarded for.

@shivanivirdi

NEOSAGE

Rejection Sampling + Supervised Fine-Tuning

- The model was a specialist, excelling at STEM.
- It lacked general-purpose abilities like writing, summarization, and chat.



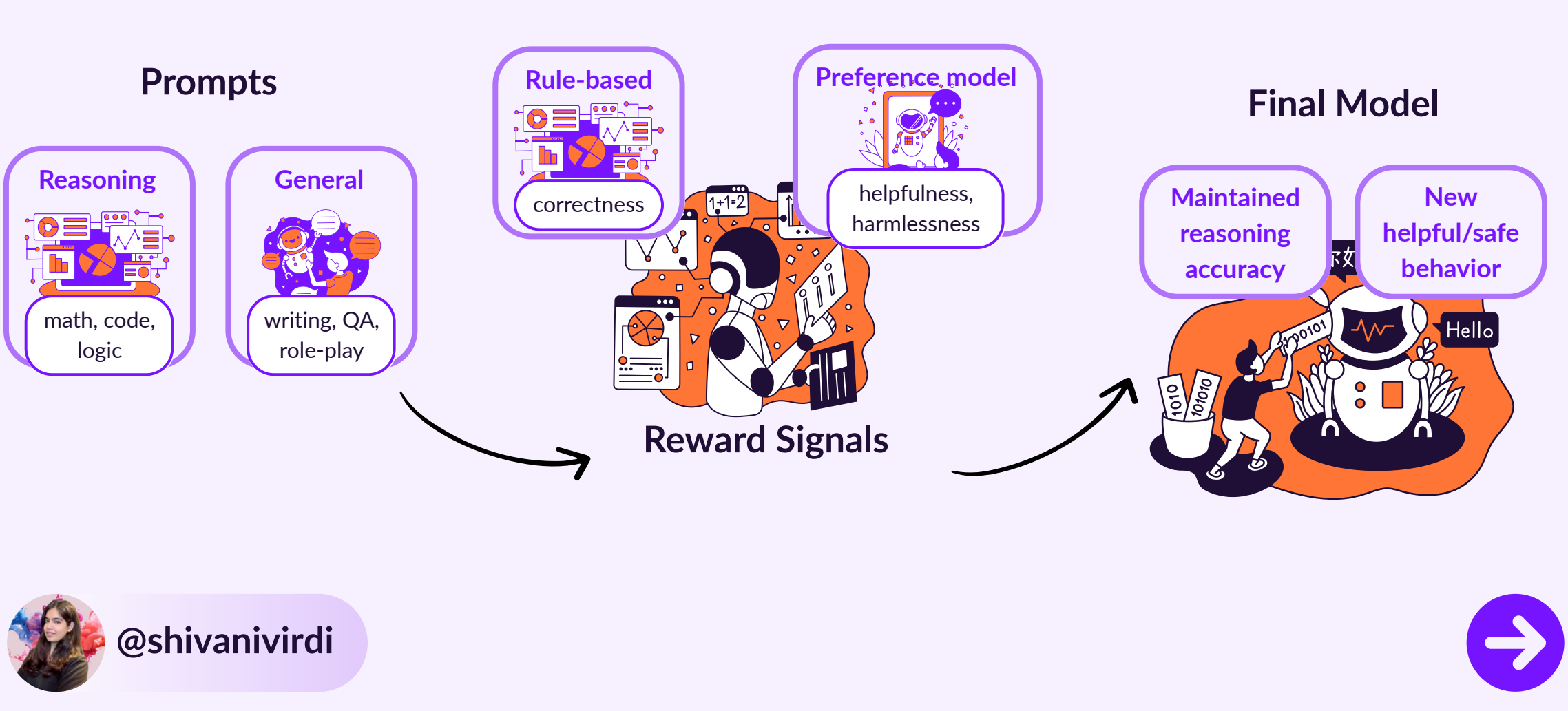
@shivanivirdi

NEOSAGE

Reinforcement for Alignment and Safety

- **Problem**: The model was capable but not yet aligned on subjective qualities like helpfulness and safety.
- **Solution**: A final RL pass using a hybrid reward system: rule-based rewards for reasoning tasks and preference models for helpfulness/harmlessness on general tasks.

Reinforcement for All Scenarios



@shivanivirdi

NEOSAGE

The Payoff: Distilling Genius

- Used the curated ~800K dataset from Stage 3.
- Fine-tuned smaller, dense models like LLaMA3 & Qwen2.5.
- No RL needed, just standard Supervised Fine-Tuning (SFT).
- Result: The smaller models inherited the emergent reasoning behaviors of the large R1 model, proving the capability is transferable.

@shivanivirdi

NEOSAGE

Final Note Think Like an Architect

This isn’t just about DeepSeek. It’s a new paradigm for building advanced AI.

Systematic Refinement: Each stage deliberately fixed a failure from the previous one.
Incentivised Behaviour: Reasoning wasn’t labelled in a dataset; the system was designed to make it the optimal strategy.
Distilled Capability: Smaller models don’t have to learn from scratch; they can inherit complex behaviours from a well-trained teacher model.

Reasoning is emergent. It cannot be taught through imitation alone

Read the full deep dive:
https://blog.neosage.io/p/inside-deepseek-r1-a-masterclass

@shivanivirdi

@shivanivirdi

Liked This?

SAVE
REPOST
FOLLOW

Read full issue on my newsletter blog.neosage.io

NEOSAGE