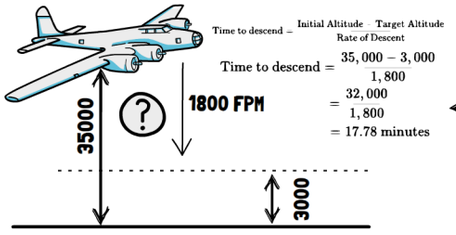
Tiramisu Recipe?

FACTS

VS



$$\text{Time to descend} = \frac{\text{Initial Altitude} - \text{Target Altitude}}{\text{Rate of Descent}}$$

$$\text{Time to descend} = \frac{35,000 - 3,000}{1,800} = \frac{32,000}{1,800} = 17.78 \text{ minutes}$$

o1, DeepSeek R1, gemini-thinking-exp, Grok 3, Claude 3.7 with extended thinking

Structured, multi-step thinking

REASONING MODELS

Chain of Thought

- INFERENCE-TIME SCALING
- PURE REINFORCEMENT LEARNING
- SUPERVISED FINE TUNING + REINFORCEMENT LEARNING
- DISTILLATION

INFERENCE-TIME SCALING

Improve LLM performance at inference time

GENERATE LONGER RESPONSES

GOOGLE, 2022

Chain-of-Thought Prompting Elicits Reasoning in Large Language Models

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: Let's think step by step.

There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls.

Chain of Thought

Generate multiple answers

LLM

Select best

- **MAJORITY VOTING**: Most frequent / consensus answer chosen
- **BEAM SEARCH**: Maintain multiple candidate sequences and select most promising iteratively

STANDARD LLM TRAINING PIPELINE

PRE-TRAINING

Massive Unlabled Datasets → Next-Word Prediction → FUNDATION BASE MODELS

Learn representations of language

SUPERVISED FINE-TUNING

Curated SFT Dataset

Translate this to Italian [Example text] [Example translation]

Summarize the following [Example text] [Example summary]

INSTRUCTIONS / TASKS

Conversational workflows

REINFORCEMENT LEARNING

Prompt → LLM → Answer → Reward

REWARD MODEL: Human preference

PPO Proximal Policy Optimization

Refine: Follow policy that maximizes reward

DEEPSEEK, 2025

Skip STF and use only RL

DEEPSEEK R1-ZERO

DeepSeek-v3 (671B) foundation model

SKIP SFT

PROMPT TEMPLATE

A conversation between User and Assistant. The user asks a question, and the Assistant solves it. The assistant first thinks about the reasoning process in the mind and then provides the user with the answer.

The reasoning process and answer are enclosed within <think> </think> and <answer> </answer> tags, respectively, i.e., <think> reasoning process </think> <answer> answer here </answer>.

User: prompt. Assistant:

GRPO: Group Relative Policy Optimization

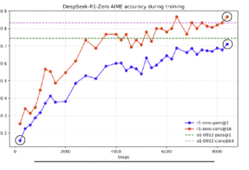
More efficient and cost-effective than PPO and DPO

RULE-BASED REWARDS

Maths

Coding

- **ACCURACY**: Does code complie? Is the maths solution correct?
- **FORMAT**: Does the model use <think> and <answer> tags correctly?



Model	AIME 2024	
	pass@1	cons@64
OpenAI-o1-mini	63.6	80.0
OpenAI-o1-0912	74.4	83.3
DeepSeek-R1-Zero	71.0	86.7

DEEPSEEK R1

FIX R1-Zero Drawbacks

- Poor readability
- Language mixing

"To solve this, we need to 'consider' the '关键因素' which is the key factor in this situation."

Cleaned-up CoT examples generated with DeepSeek R1-Zero

Cold-start STF dataset

REJECTION SAMPLING

- ✓ Only select high-quality samples
- ✗ Reject mixed-language, poorly formatted or unstructured responses

600k CoT Examples

+

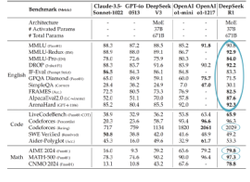
200k Non-Reasoning Examples

1ST ROUND OF REINFORCEMENT LEARNING

FINAL ROUND OF REINFORCEMENT LEARNING

Multiple reinforcement strategies to target both reasoning and alignement with human preferences

- **CONSISTENCY**: Reduce language mixing
- **REWARD MODEL**: Predict which responses humans prefer



Benchmark	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
English MMLU	88.3	87.2	88.5	85.2	91.8	90.8
English MMLU-Redux	88.9	88.0	89.1	86.7	-	92.9
English MMLU-Pro	78.0	72.6	75.9	80.3	-	84.0
English DROP	88.3	83.7	91.6	83.9	90.2	92.2
English IF-Eval	86.5	84.3	86.1	84.8	-	83.3
English GPQA Diamond	65.0	49.9	59.1	60.0	75.7	71.5
English SimpleQA	28.4	38.2	24.9	7.0	47.0	30.1
English FRAMES	72.5	80.5	73.3	76.9	-	82.5
English AlpacaEval2.0	52.0	51.1	70.0	57.8	-	87.6
English ArenaHard	85.2	80.4	85.5	92.0	-	92.3
Code LiveCodeBench	38.9	32.9	36.2	53.8	63.4	65.9
Code Codeforces	20.3	23.6	58.7	93.4	96.6	96.3
Code Codeforces	717	759	1134	1820	2061	2029
Code SWE Verified	50.8	38.8	42.0	41.6	48.9	49.2
Code Aider-Polyglot	45.3	16.0	49.6	32.9	61.7	53.3
Math AIME 2024	16.0	9.3	39.2	63.6	79.2	79.8
Math MATH-500	78.3	74.6	90.2	90.0	96.4	97.3
Math CNMO 2024	13.1	10.8	43.2	67.6	-	78.8

DISTILLATION

Teacher model

Student model

Knowledge Distillation

Llama 8B
Llama 70B
Qwen 1.5B
Qwen 7B
Qwen 14B
Qwen 32B

Supervised Fine Tuning

DeepSeek-R1-Distill-Llama 8B
DeepSeek-R1-Distill-Llama 70B
DeepSeek-R1-Distill-Qwen 1.5B
DeepSeek-R1-Distill-Qwen 7B
DeepSeek-R1-Distill-Qwen 14B
DeepSeek-R1-Distill-Qwen 32B

Model	AIME 2024		MATH-500	GPQA Diamond	LiveCode Bench	CodeForces
	pass@1	cons@64	pass@1	pass@1	pass@1	rating
GPT-4o-0513	9.3	13.4	74.6	49.9	32.9	759
Claude-3.5-Sonnet-1022	16.0	26.7	78.3	65.0	38.9	717
OpenAI-o1-mini	63.6	80.0	90.0	60.0	53.8	1820
QwQ-32B-Preview	50.0	60.0	90.6	54.5	41.9	1316
DeepSeek-R1-Distill-Qwen-1.5B	28.9	52.7	83.9	33.8	16.9	954
DeepSeek-R1-Distill-Qwen-7B	55.5	83.3	92.8	49.1	37.6	1189
DeepSeek-R1-Distill-Qwen-14B	69.7	80.0	93.9	59.1	53.1	1481
DeepSeek-R1-Distill-Qwen-32B	72.6	83.3	94.3	62.1	57.2	1691
DeepSeek-R1-Distill-Llama-8B	50.4	80.0	89.1	49.0	39.6	1205
DeepSeek-R1-Distill-Llama-70B	70.0	86.7	94.5	65.2	57.5	1633

Based on: https://magazine.sebastianraschka.com/p/understanding-reasoning-llms

AHEAD OF AI — By Sebastian Raschka, PhD

YouTube youtube.com/@donatocapitella