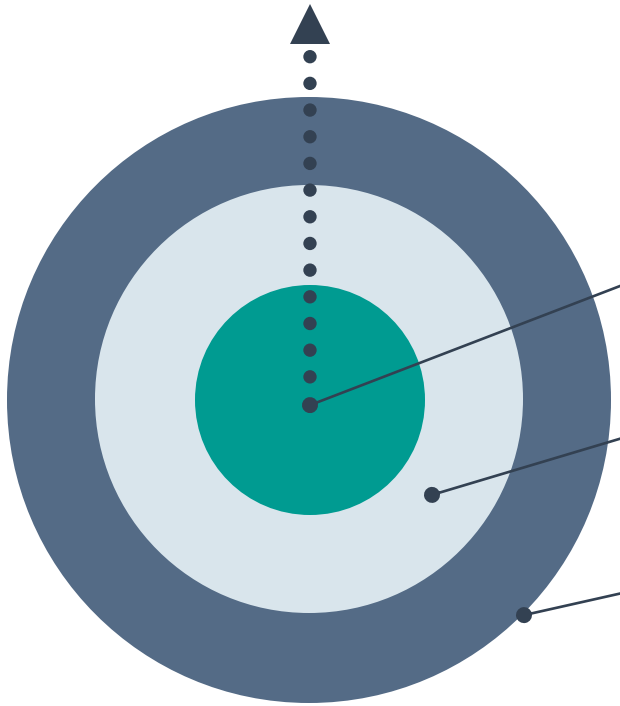
Hochschule Konstanz
Fakultät Informatik

DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning

Abschlusspräsentation, Seminar: Advanced Topics in Data Analysis and Deep Learning

Lukas Epple

Agenda

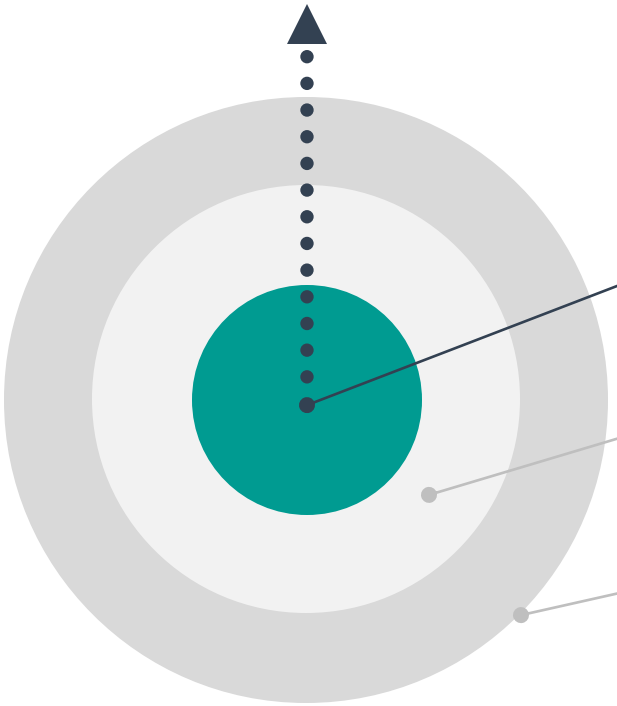


WHY ?
Warum wurde das Thema überhaupt genauer betrachtet?

HOW ?
Wie wurde vorgegangen?

WHAT ?
Was waren die Ergebnisse?

Agenda



WHY ?
Warum wurde das Thema überhaupt genauer betrachtet?

HOW ?
Wie wurde vorgegangen?

WHAT ?
Was waren die Ergebnisse?

Warum Reasoning?

Für bestimmte Aufgaben ist die Korrektheit der Rückgabe viel wichtiger als die Geschwindigkeit

„Think Fast“

Schnelle Antworten

Günstiger, geeignet für einfache oder kreative Aufgaben, wo die logische Korrektheit nicht so relevant ist.

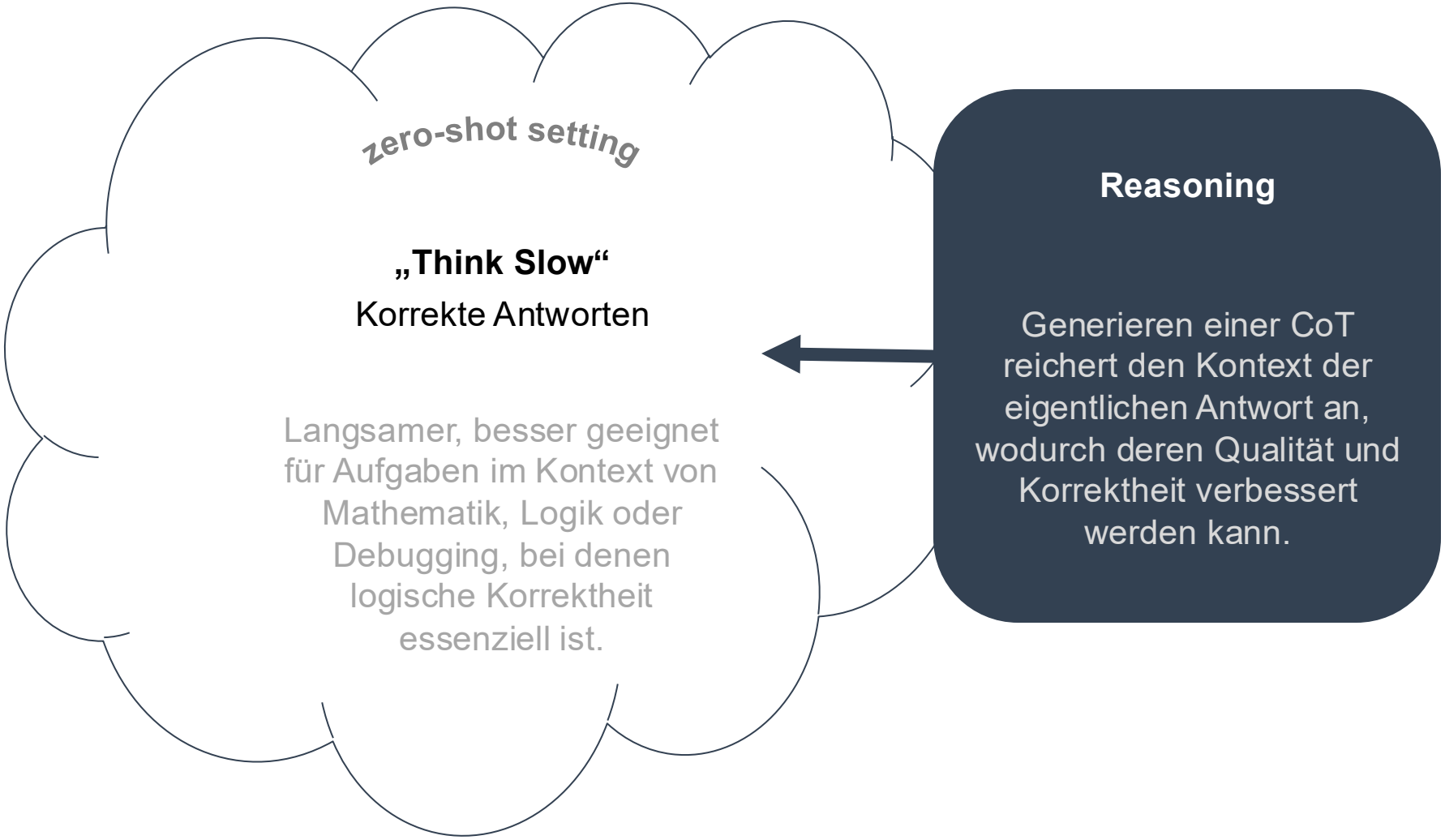
„Think Slow“

Korrekte Antworten

Langsamer, besser geeignet für Aufgaben im Kontext von Mathematik, Logik oder Debugging, bei denen logische Korrektheit essenziell ist.

Warum Reasoning?

Durch generierte Gedankenkette kann die Token-Vorhersage der Antwort präzisiert werden



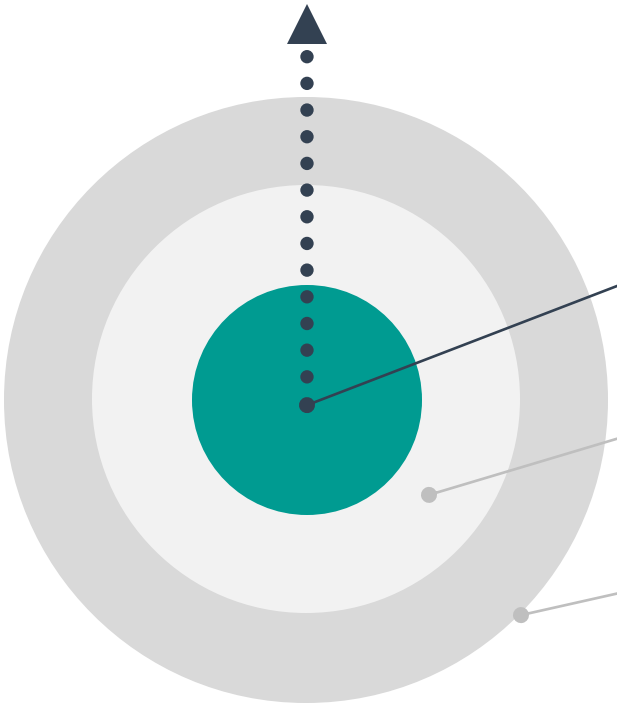
Warum Reinforcement Learning?

Autonomes Training ermöglicht schnelleres Training und effektivere Skalierung

+ **Autonomes Training** möglich
+ Deutlich **schneller und besser Skalierbar** für „große Aufgaben“
+ Keine riesige Menge **vordefinierter Daten** notwendig

  (Vollumfängliche Trainingsdaten erstellen ist ohnehin nicht trivial: curse of dimensionality)
+ Intrinsisch motivierter Denkprozess (nicht nur „**Nachahmen**“)

Agenda

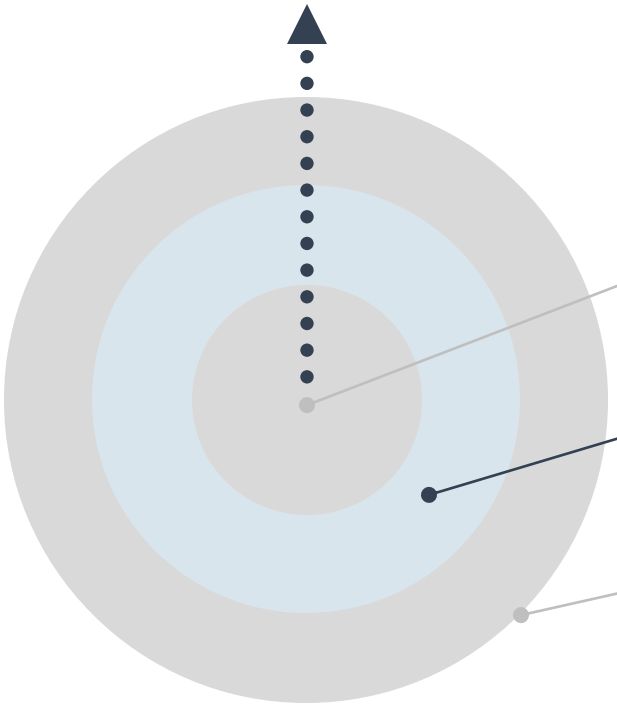


WHY ?
Warum wurde das Thema überhaupt genauer betrachtet?

HOW ?
Wie wurde vorgegangen?

WHAT ?
Was waren die Ergebnisse?

Agenda



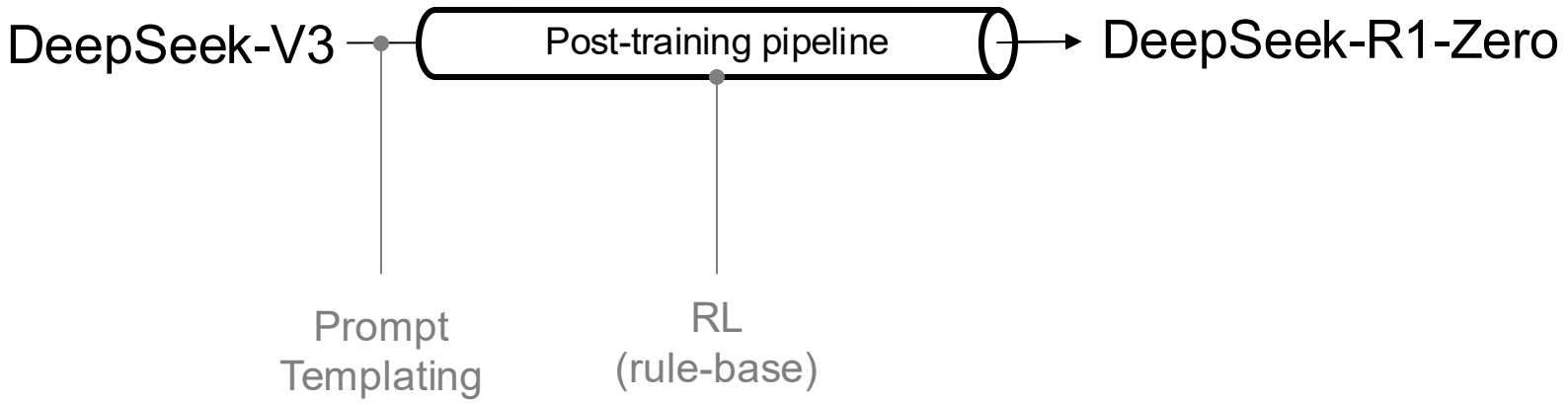
WHY ?
Warum wurde das Thema überhaupt genauer betrachtet?

HOW ?
Wie wurde vorgegangen?

WHAT ?
Was waren die Ergebnisse?

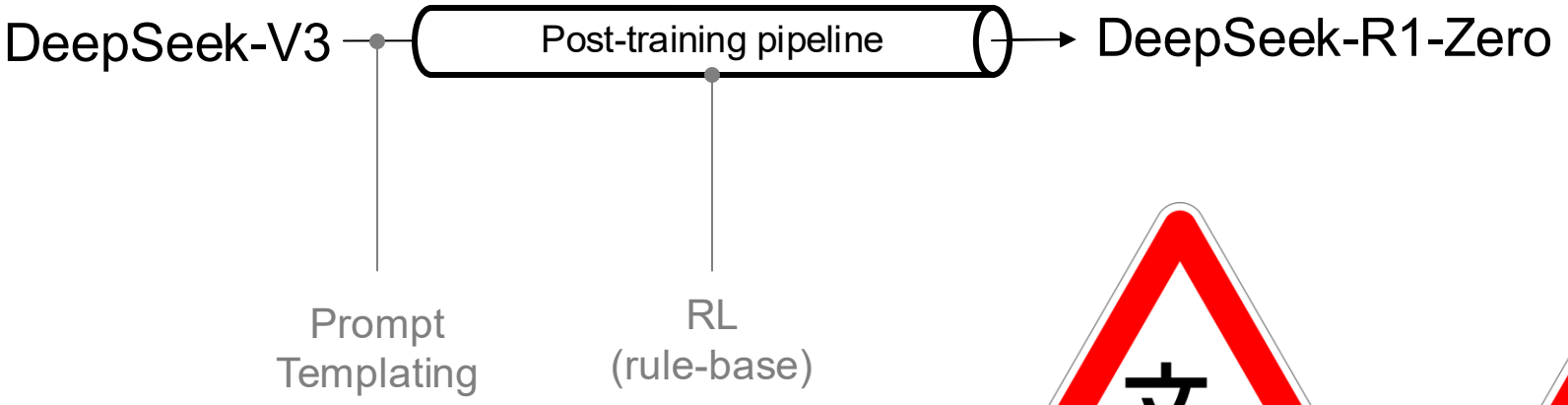
„Pure reinforcement learning“

Post-training des Mixture of Experts Model



„Pure reinforcement learning“

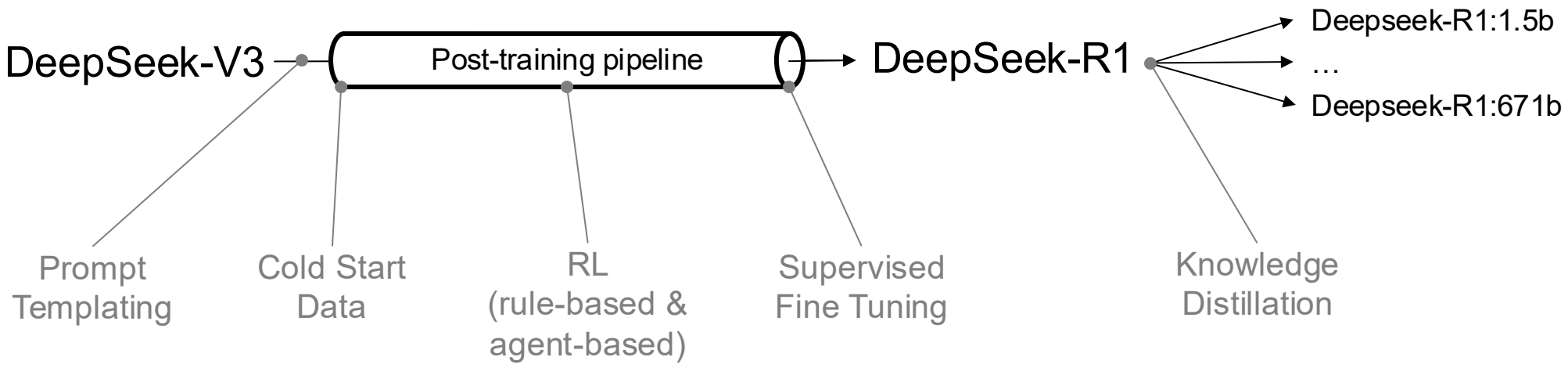
Post-training des Mixture of Experts Model



https://creazilla.com/media/icon/3231827/translate
https://www.rueden.de/wp-content/uploads/2019/06/traffic-sign-160676_1280.png

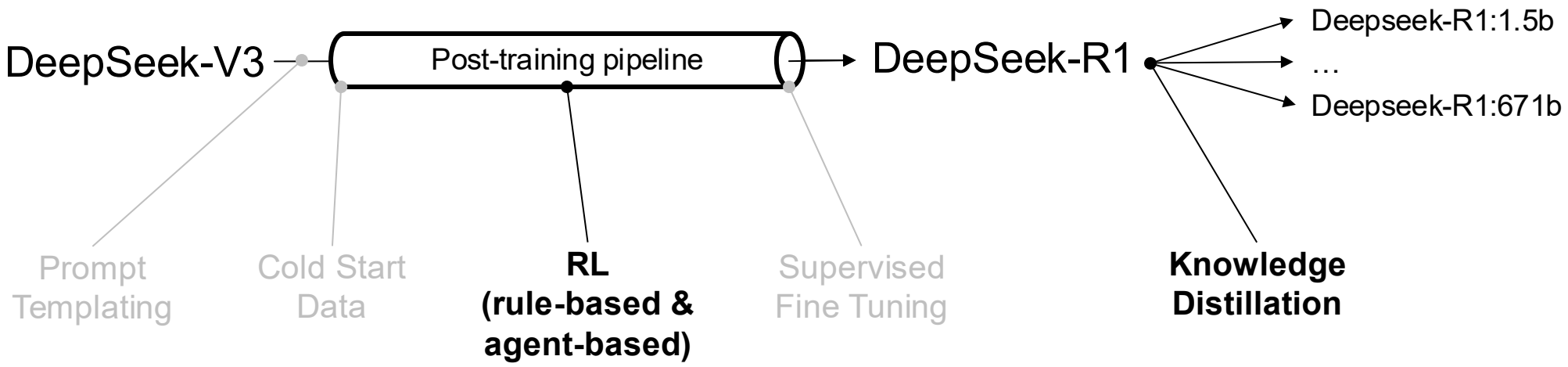
Finale Post-training Pipeline für das Training von DeepSeek-R1

Beheben der Probleme durch Cold Start Data, mehr RL und SFT



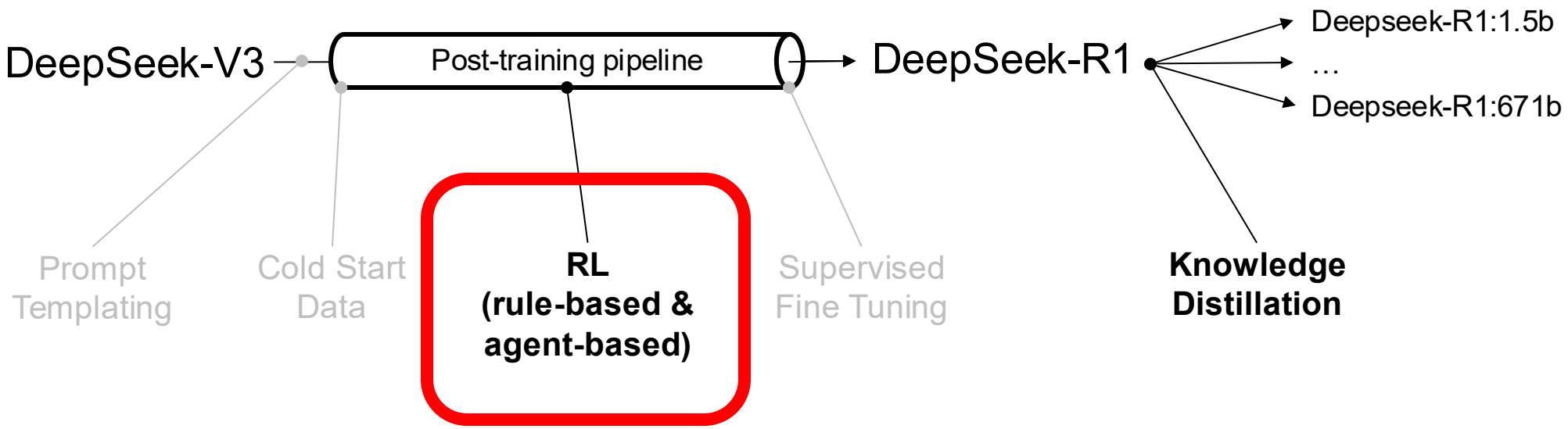
Finale Post-training Pipeline für das Training von DeepSeek-R1

Beheben der Probleme durch Cold Start Data, mehr RL und SFT



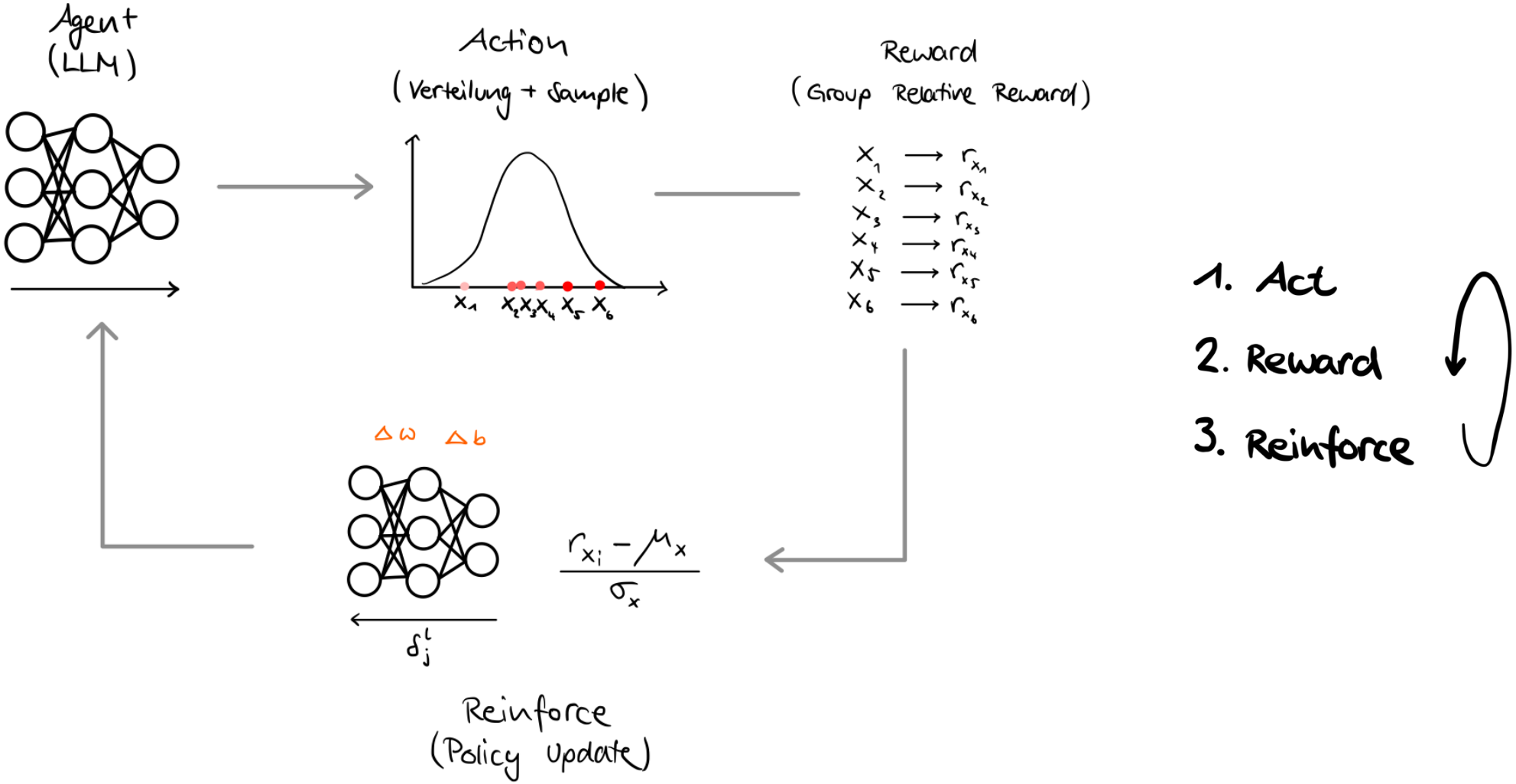
Finale Post-training Pipeline für das Training von DeepSeek-R1

Beheben der Probleme durch Cold Start Data, mehr RL und SFT



Reinforcement Learning für LLMs

Group Relative Policy Optimization (GRPO): Maximieren der Anzahl überdurchschnittlich guter Samples



1. Act
2. Reward
3. Reinforce

Group Relative Policy Optimization (GRPO)

Maximierung der erzielten Advantages

2.2.1. *Reinforcement Learning Algorithm*

Group Relative Policy Optimization In order to save the training costs of RL, we adopt Group Relative Policy Optimization (GRPO) (Shao et al., 2024), which foregoes the critic model that is typically the same size as the policy model, and estimates the baseline from group scores instead. Specifically, for each question q, GRPO samples a group of outputs $\{o_1, o_2, \cdots, o_G\}$ from the old policy $\pi_{\theta_{old}}$ and then optimizes the policy model π_θ by maximizing the following objective:

$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}[q \sim P(Q), \{o_i\}_{i=1}^{G} \sim \pi_{\theta_{old}}(O|q)]$$

$$\frac{1}{G}\sum_{i=1}^{G}\left(\min\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}A_i, \text{clip}\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}, 1-\varepsilon, 1+\varepsilon\right)A_i\right) - \beta\mathbb{D}_{KL}\left(\pi_\theta||\pi_{ref}\right)\right), \tag{1}$$

$$\mathbb{D}_{KL}\left(\pi_\theta||\pi_{ref}\right) = \frac{\pi_{ref}(o_i|q)}{\pi_\theta(o_i|q)} - \log\frac{\pi_{ref}(o_i|q)}{\pi_\theta(o_i|q)} - 1, \tag{2}$$

where ε and β are hyper-parameters, and A_i is the advantage, computed using a group of rewards $\{r_1, r_2, \ldots, r_G\}$ corresponding to the outputs within each group:

$$A_i = \frac{r_i - \text{mean}(\{r_1, r_2, \cdots, r_G\})}{\text{std}(\{r_1, r_2, \cdots, r_G\})}. \tag{3}$$

Entwickelt in: https://arxiv.org/abs/2402.03300 (Apr. 2024)

Anschauliches Beispiel: https://huggingface.co/docs/trl/main/en/grpo_trainer

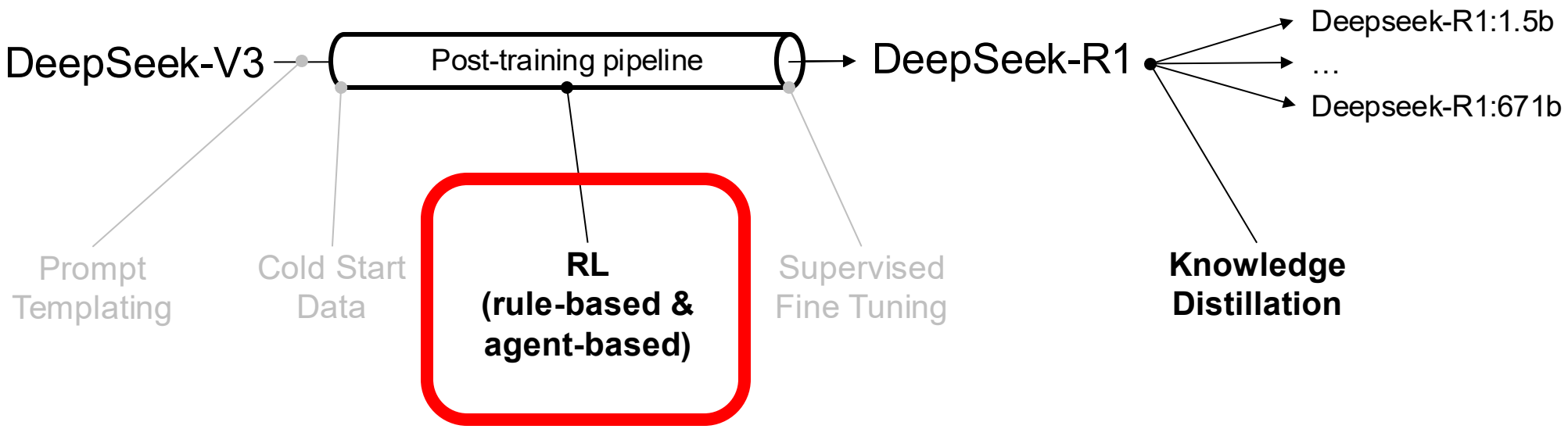
Wie wird der Reward berechnet?

DeepSeek-R1 verwendet eine Kombination aus regelbasierten Rewards und Rewardmodellen für unterschiedliche Aspekte

RL Stage	Belohnungsmodell	Fokus	Rewards
Stage 1 (R1 & R1-Zero)	Rule-based Reward	Korrektheit & Lesbarkeit	Korrektheit, Formatierung
Stage 2 (R1)	Reward Model	Menschliche Präferenzen	Helpfulness und Harmlessness

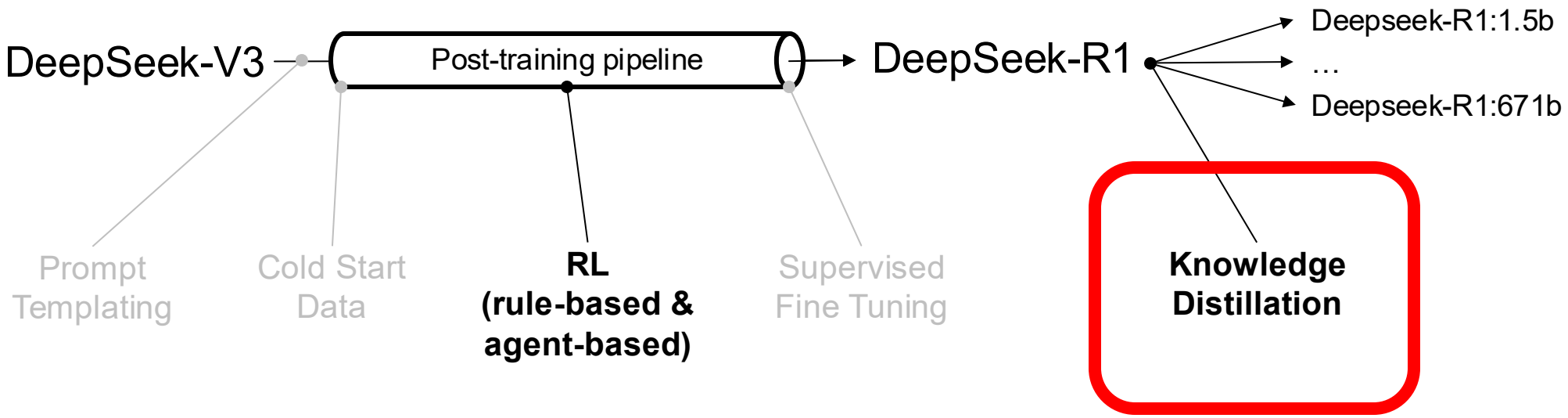
Finale Post-training Pipeline für das Training von DeepSeek-R1

Beheben der Probleme durch Cold Start Data, mehr RL und SFT



Finale Post-training Pipeline für das Training von DeepSeek-R1

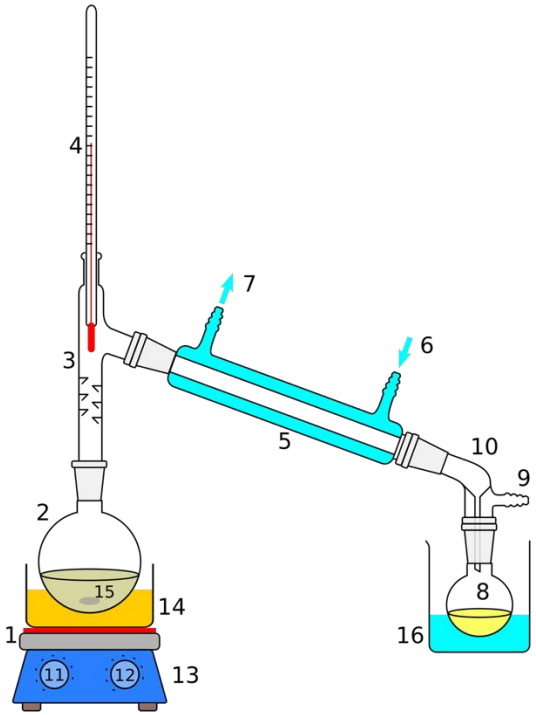
Beheben der Probleme durch Cold Start Data, mehr RL und SFT



Distillation

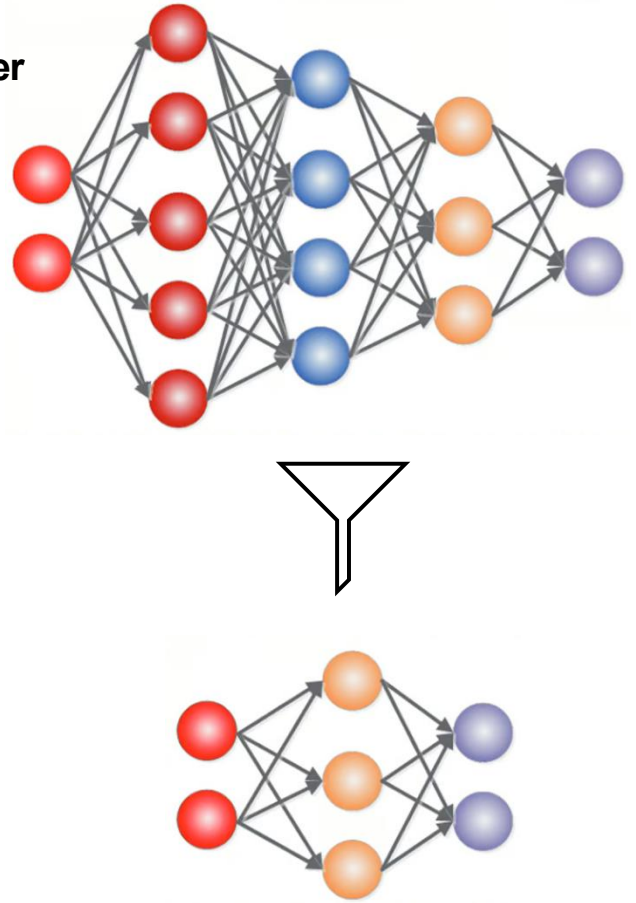
Modellkomprimierung durch Verdichtung

- Viel Flüssigkeit
- Viel Alkohol



- Weniger Flüssigkeit
- Fast gleich viel Alkohol

- Viele Parameter
- Viel „Wissen“

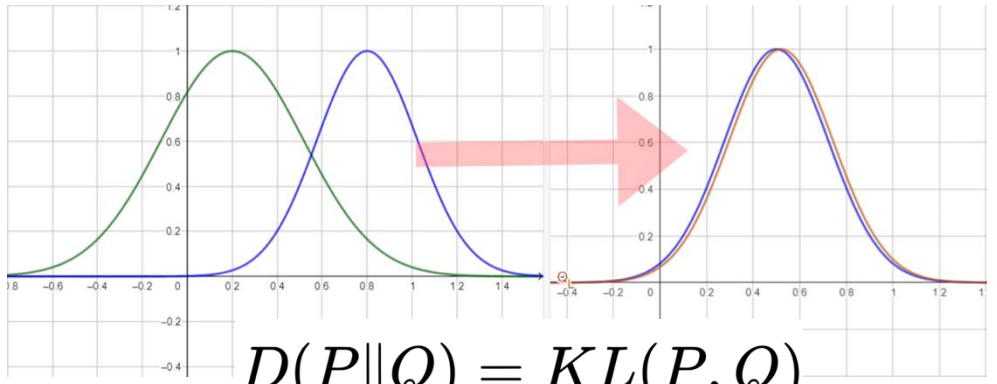


- Weniger Parameter
- Fast gleich viel „Wissen“

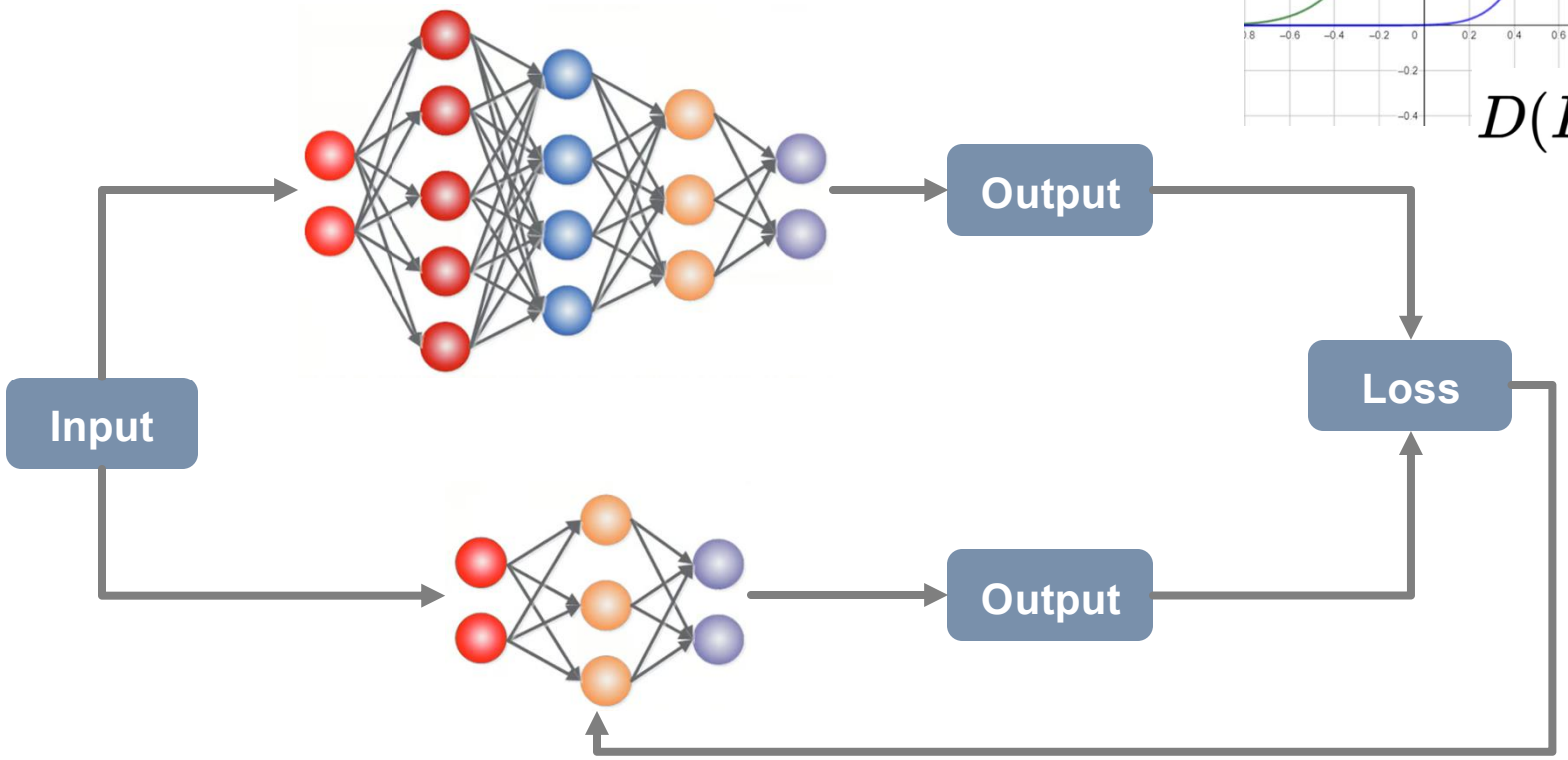
https://www.chem.ucla.edu/~harding/IGOC/D/distillation01.png
https://media.datacamp.com/legacy/image/upload/v1724950676/image_ca49121d8a.png

Wie Model-Distillation funktioniert

Nachahmen eines „Vorbilds“ durch einen „Schüler“

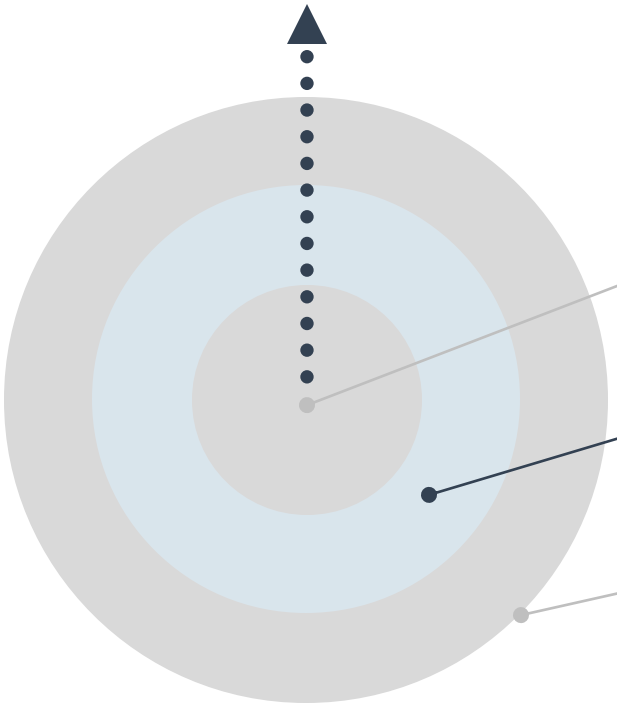


$$D(P\|Q) = KL(P,Q)$$



https://www.youtube.com/watch?v=vyJy-0zBSQ0
https://de.wikipedia.org/wiki/Kullback-Leibler-Divergenz

Agenda

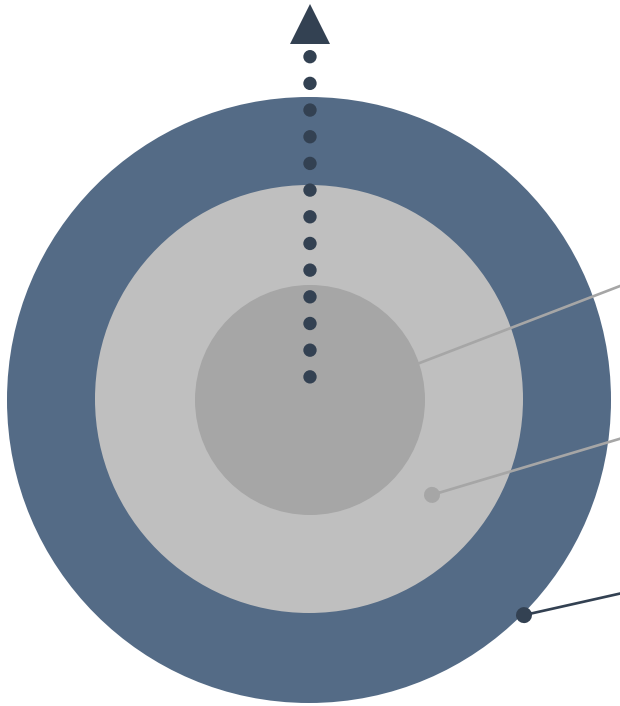


WHY ?
Warum wurde das Thema überhaupt genauer betrachtet?

HOW ?
Wie wurde vorgegangen?

WHAT ?
Was waren die Ergebnisse?

Agenda



WHY ?
Warum wurde das Thema überhaupt genauer betrachtet?

HOW ?
Wie wurde vorgegangen?

WHAT ?
Was waren die Ergebnisse?

Research Erkenntnisse (DeepSeek-R1-Zero)

Open Validation von RL als effektive Methode für autonomes Training

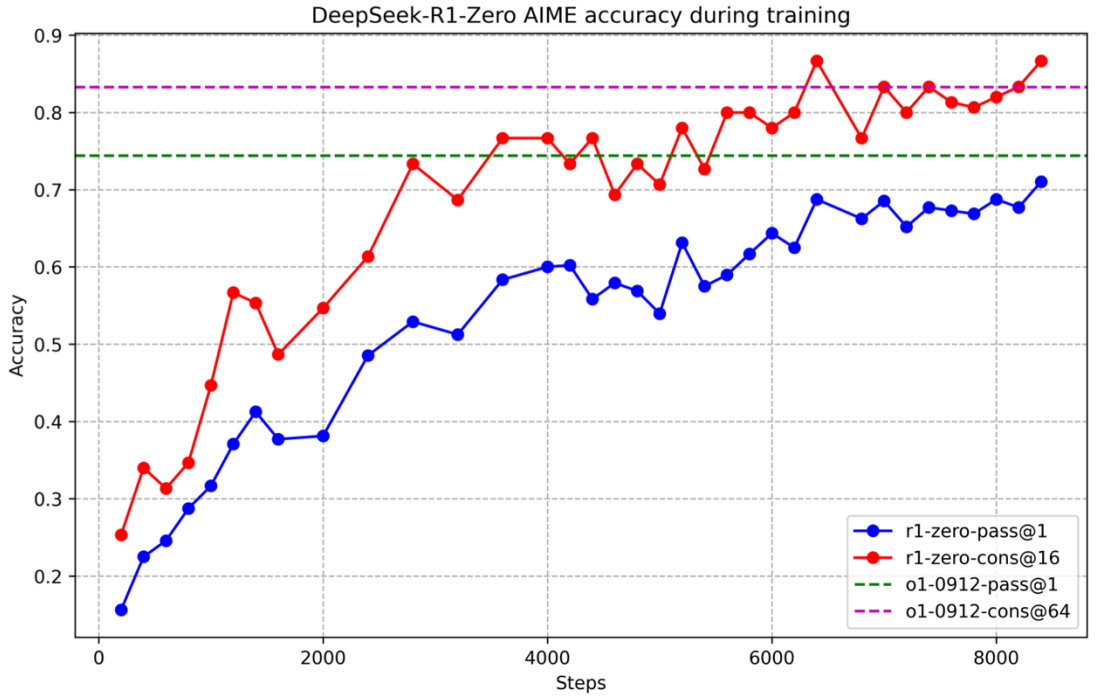


Figure 2 | AIME accuracy of DeepSeek-R1-Zero during training. For each question, we sample 16 responses and calculate the overall average accuracy to ensure a stable evaluation.

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 $\cdots$

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

Rearrange to isolate the inner square root term:

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

$\ldots$

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 $\cdots$

We started with the equation:

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

First, let's square both sides:

Model	AIME 2024		MATH-500	GPQA Diamond	LiveCode Bench	CodeForces
	pass@1	cons@64	pass@1	pass@1	pass@1	rating
OpenAI-o1-mini	63.6	80.0	90.0	60.0	53.8	1820
OpenAI-o1-0912	74.4	83.3	94.8	77.3	63.4	1843
DeepSeek-R1-Zero	71.0	86.7	95.9	73.3	50.0	1444

Table 2 | Comparison of DeepSeek-R1-Zero and OpenAI o1 models on reasoning-related benchmarks.

DeepSeek-R1 Evaluation

DeepSeek-R1 als Artefakt mit beeindruckender Outputqualität

	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
English	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**
	IF-Eval (Prompt Strict)	**86.5**	84.3	86.1	84.8	-	83.3
	GPQA Diamond (Pass@1)	65.0	49.9	59.1	60.0	**75.7**	71.5
	SimpleQA (Correct)	28.4	38.2	24.9	7.0	**47.0**	30.1
	FRAMES (Acc.)	72.5	80.5	73.3	76.9	-	**82.5**
	AlpacaEval2.0 (LC-winrate)	52.0	51.1	70.0	57.8	-	**87.6**
	ArenaHard (GPT-4-1106)	85.2	80.4	85.5	92.0	-	**92.3**
Code	LiveCodeBench (Pass@1-COT)	38.9	32.9	36.2	53.8	63.4	**65.9**
	Codeforces (Percentile)	20.3	23.6	58.7	93.4	**96.6**	96.3
	Codeforces (Rating)	717	759	1134	1820	**2061**	2029
	SWE Verified (Resolved)	**50.8**	38.8	42.0	41.6	48.9	49.2
	Aider-Polyglot (Acc.)	45.3	16.0	49.6	32.9	**61.7**	53.3
Math	AIME 2024 (Pass@1)	16.0	9.3	39.2	63.6	79.2	**79.8**
	MATH-500 (Pass@1)	78.3	74.6	90.2	90.0	96.4	**97.3**
	CNMO 2024 (Pass@1)	13.1	10.8	43.2	67.6	-	**78.8**
Chinese	CLUEWSC (EM)	85.4	87.9	90.9	89.9	-	**92.8**
	C-Eval (EM)	76.7	76.0	86.5	68.9	-	**91.8**
	C-SimpleQA (Correct)	55.4	58.7	**68.0**	40.3	-	63.7

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

Distillation Results

DeepSeek-R1 als Artefakt mit beeindruckender Outputqualität

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

Table 5 | Comparison of DeepSeek-R1 distilled models and other comparable models on reasoning-related benchmarks.

Ausblick: Erneute RL stage nach der Distillation

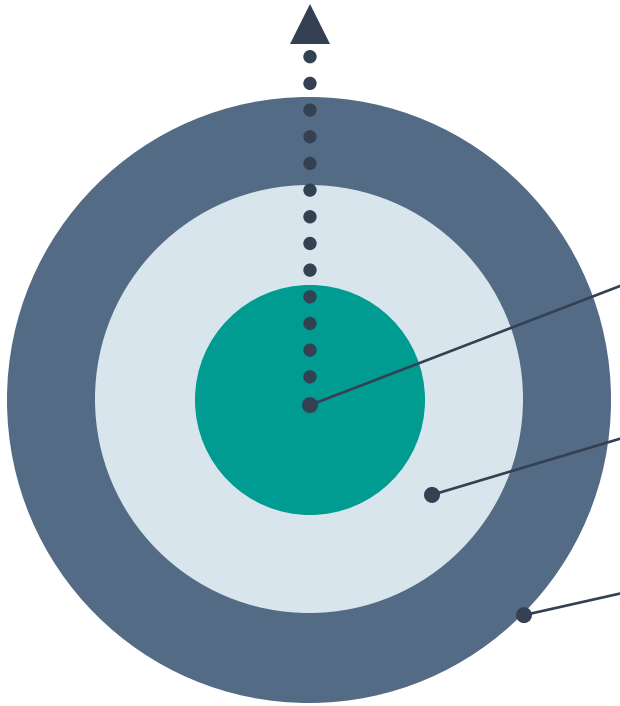
Kritik und Ausblick

- **"Pure RL":** Für herausragende Ergebnisse waren viele weitere Schritte notwendig.
  Die offene Validierung fand jedoch trotzdem umfangreich statt.

- **Limitierungen bei Benchmark-Vergleichen:** Abhängigkeit von veröffentlichten Benchmark Werten für OpenAI-Modellen aufgrund eingeschränkter Zugänglichkeit in China.

- **Ausbessern von Unzulänglichkeiten:**
  - In manchen Bereichen schlechter als das Basismodell (complex role-playing, JSON output,…)
  - Language Mixing bei Sprachen außer Englisch und Chinesisch
  - Few-shot prompting verschlechtert die Outputqualität
- **Anwendung von RL auf die Distilled Models**

Agenda



WHY ?
Warum wurde das Thema überhaupt genauer betrachtet?

HOW ?
Wie wurde vorgegangen?

WHAT ?
Was waren die Ergebnisse?

Demo

Lokales Verwenden des 8b distilled models mit Ollama

deepseek-r1

DeepSeek's first-generation of reasoning models with comparable performance to OpenAI-o1, including six dense models distilled from DeepSeek-R1 based on Llama and Qwen.

1.5b 7b 8b 14b 32b 70b 671b

42M Pulls Updated 2 months ago

8b 29 Tags `ollama run deepseek-r1:8b`

Updated 3 months ago		28f8fd6cdc67 · 4.9GB
model	arch **llama** · parameters **8.03B** · quantization **Q4_K_M**	4.9GB
params	{ "stop": ["<｜begin_of_sentence｜>", "<｜end_of_sentence｜>",	148B
template	{{- if .System }}{{ .System }}{{ end }} {{- range $i, $_ := .…	387B
license	MIT License Copyright (c) 2023 DeepSeek Permission is hereby …	1.1kB

https://ollama.com/library/deepseek-r1:8b

Hochschule Konstanz
Fakultät Informatik

Vielen Dank für die Aufmerksamkeit!