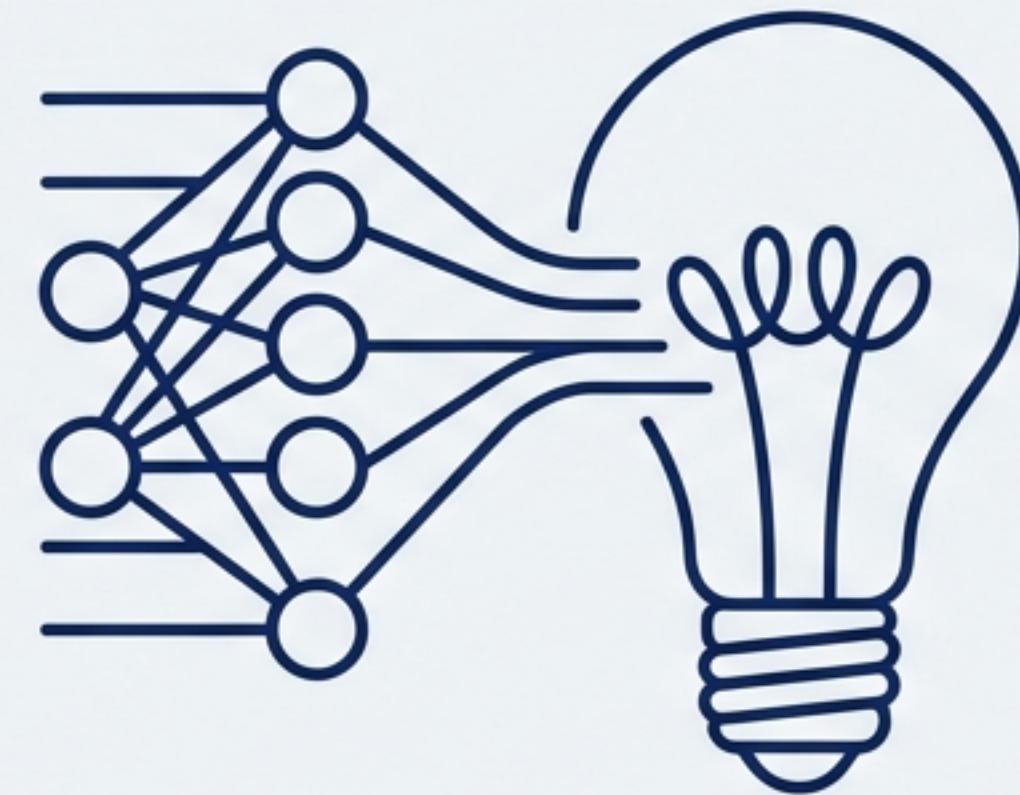
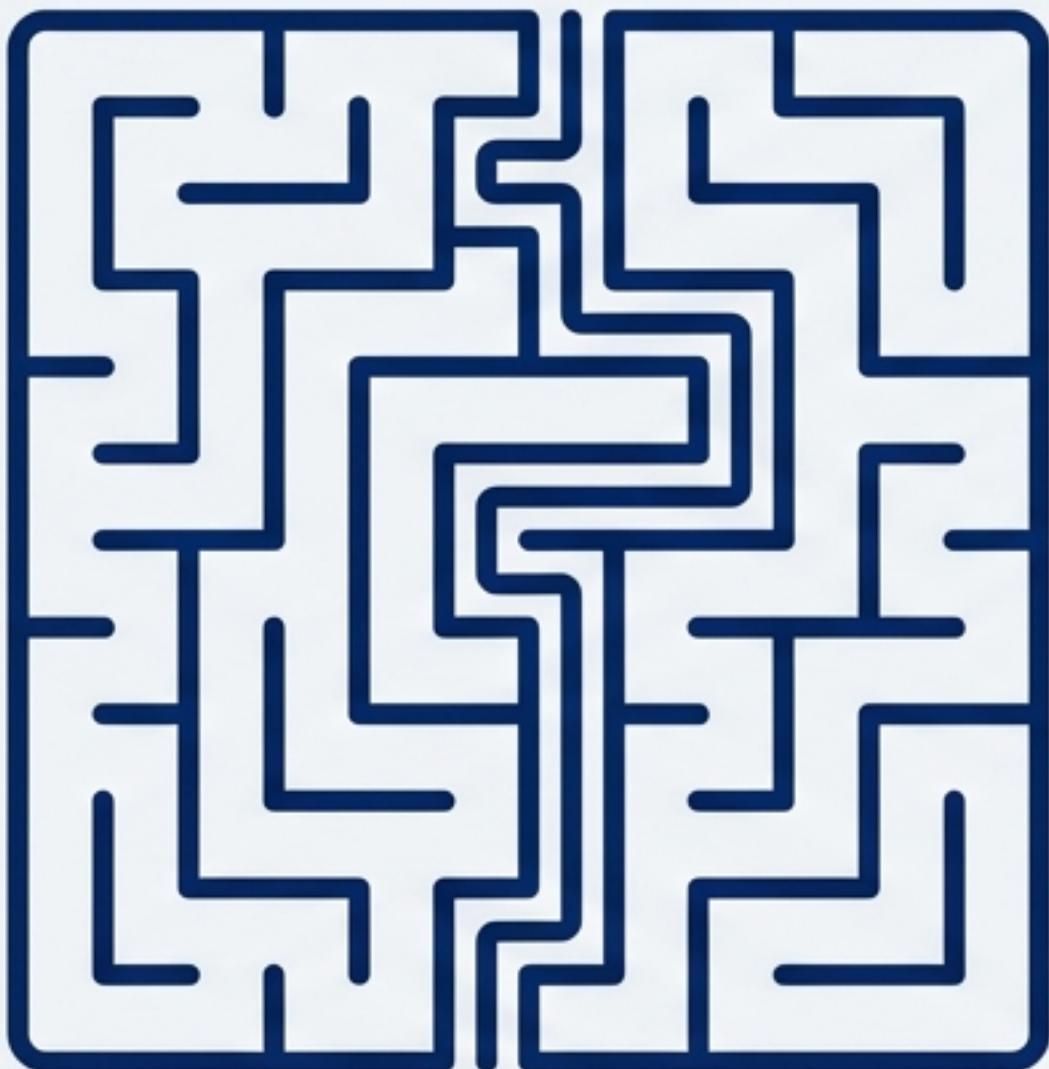


DeepSeek R1: Nowa Era Rozumowania AI

Jak Reinforcement Learning uczy modele myślenia, a nie tylko powtarzania.

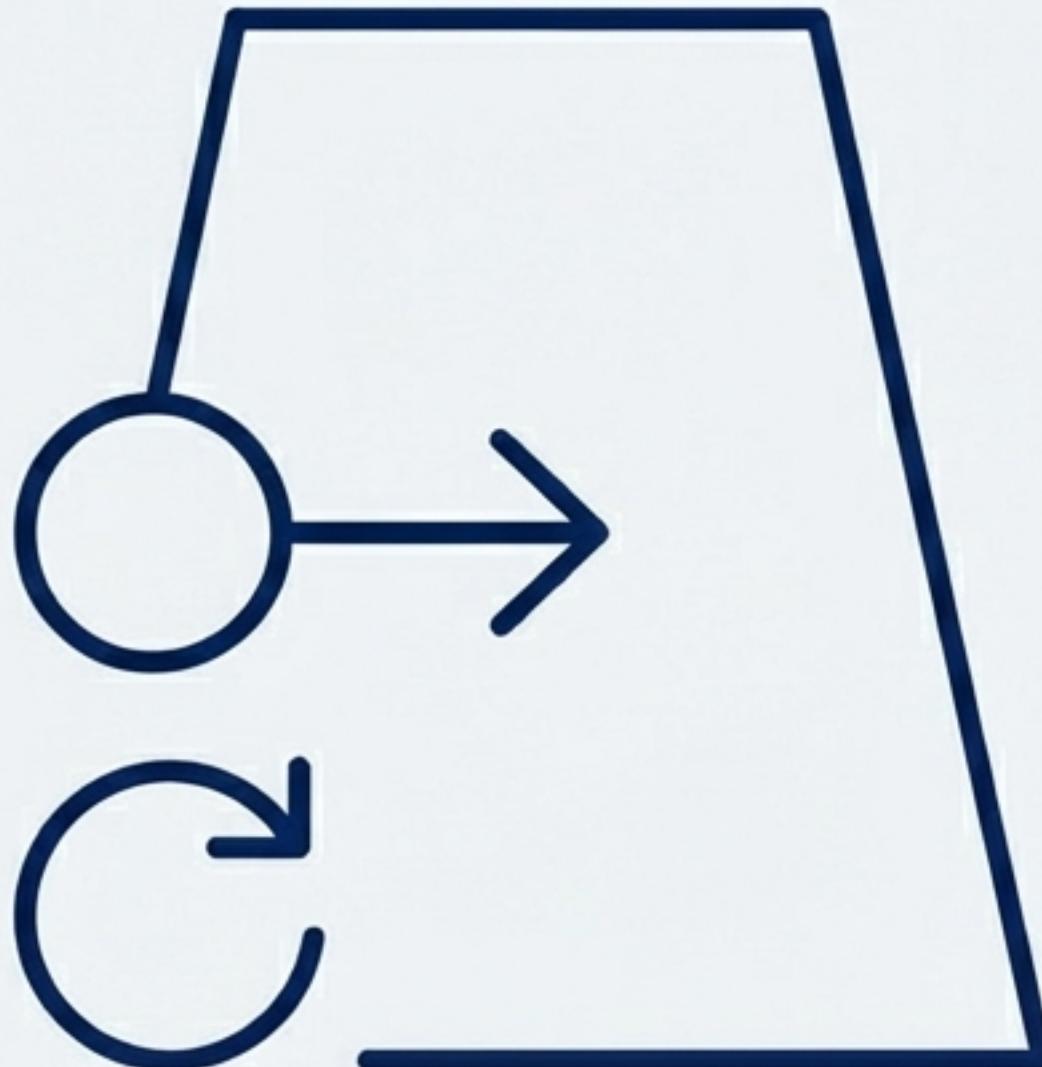


Wyzwanie: Ku Prawdziwemu Rozumowaniu



- **Problem fundamentalny:** Uczenie maszyn prawdziwego, głębokiego rozumowania, a nie tylko odtwarzania zapamiętanych faktów.
- **Kontekst rynkowy:** Do tej pory modele z serii O1 od OpenAI dominowały w zaawansowanych zadaniach wymagających myślenia.
- **Ambicja DeepSeek AI:** Osiągnąć i przewyższyć poziom state-of-the-art (SOTA) za pomocą czystego Reinforcement Learning (RL).
- **Kluczowe pytanie badawcze:** Czy model może nauczyć się rozumować wyłącznie metodą prób i błędów, bez nadzorowanych przykładów?

Eksperyment: R1-Zero – Czysta Siła RL



- **Rewolucyjne podejście:** Brak wstępnego etapu Supervised Fine-Tuning (SFT). Czysty RL zastosowany bezpośrednio na surowym modelu bazowym.
- **Podejście 'czystej karty' (blank slate):** Model uczy się od zera, jak rozwiązywać problemy, bez gotowych przykładów krok po kroku.
- **Analogia:** To jak uczenie gry w szachy poprzez nagradzanie dobrych ruchów, a nie pokazywanie partii mistrzów.
- **Znaczenie:** Pierwszy otwarty dowód koncepcji, że zdolności rozumowania mogą wyłonić się w LLM wyłącznie dzięki RL.

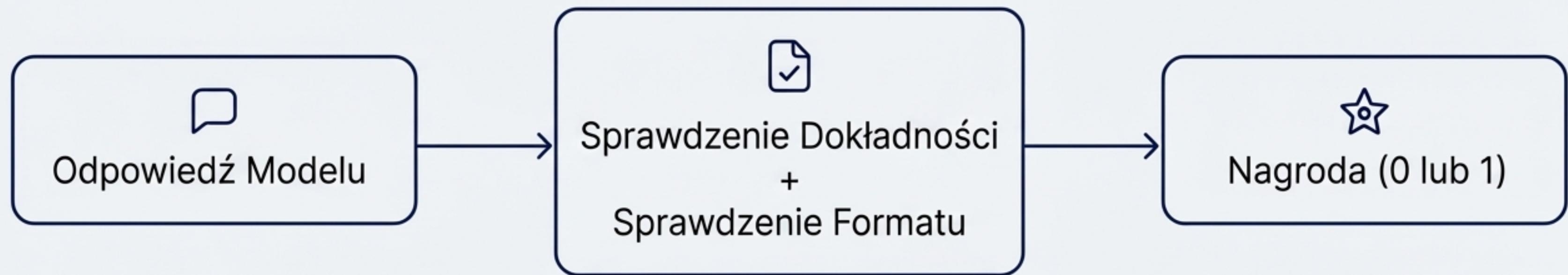
Mechanizm Nagród: Proste Zasady, Złożone Efekty

Nagrody za dokładność (Accuracy Rewards)

Binarne (0/1). Weryfikacja poprawności odpowiedzi w matematyce lub komplikacja i zaliczenie testów w zadaniach programistycznych.

Nagrody za format (Format Rewards)

Wymuszenie zgodności z instrukcjami, np. umieszczanie całego procesu myślowego w tagach <think>`.



Przełom: Spektakularne Wyniki R1-Zero

- 1. Benchmark AIME 2024 (Pass@1):**
Niesamowity skok dokładności z 15.6% do 71.0%.
- 2. Z głosowaniem większościowym (cons@64):** Wynik rośnie do 86.7%, przewyższając OpenAI O1-preview (79.2%).
- 3. Organiczny wzrost 'czasu na myślenie':** Model sam nauczył się, że trudniejsze problemy wymagają dłuższych, bardziej złożonych odpowiedzi.





Wyłonione Zachowania: Samo-ewolucja i Refleksja

Model samodzielnie, bez programowania, rozwinął zaawansowane strategie meta-poznawcze:

- **Samo-ewolucja (Self-evolution):** Spontaniczne generowanie coraz dłuższych odpowiedzi dla bardziej złożonych problemów.
- **Refleksja (Reflection):** Niezaprogramowane wracanie do poprzednich kroków rozumowania w celu ich ponownej oceny i weryfikacji.
- **Eksploracja alternatywnych ścieżek:** Porzucanie pierwotnego podejścia na rzecz zupełnie nowej strategii w trakcie rozwiązywania problemu.

'Moment Aha!': Przełom w Samo-korekcie



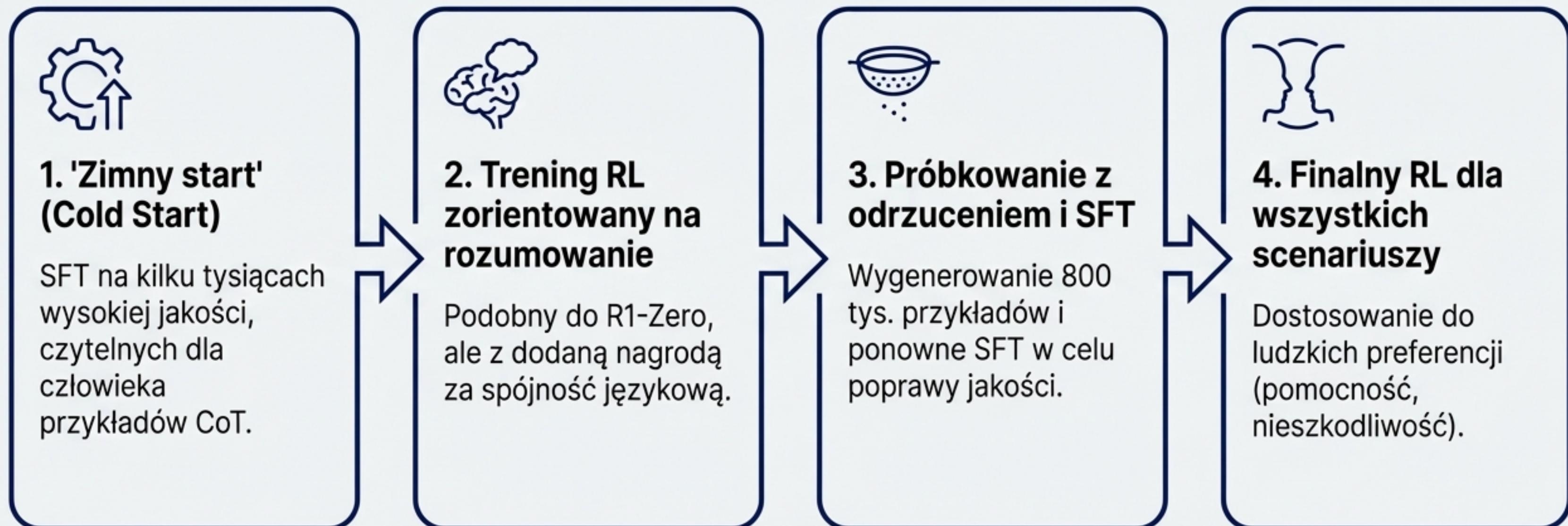
> "Wait, wait. Wait. That's an aha moment I can flag here. Let's reevaluate this step-by-step..."

(PL: "Czekaj, czekaj. Czekaj. To jest moment aha, który mogę tu oznaczyć. Oceńmy to jeszcze raz. Krok po kroku...")

- **Historyczne znaczenie:** Pierwszy udokumentowany przypadek spontanicznej samo-korekty i kwestionowania własnego toku myślenia w modelu trenowanym czystym RL.
- Przełomowy dowód, że złożone procesy myślowe mogą wyłonić się bez ludzkiego nadzoru.

Inżynieria Doskonałości: Pipeline DeepSeek R1

****Problem z R1-Zero**:** Genialny, ale chaotyczny. Problemy z czytelnością, mieszanym językiem (language mixing) i brakiem ogólnych zdolności konwersacyjnych.



Efekt Końceny: R1 na Szczycie Benchmarków



AIME 2024 (Pass@1)

79.8%

Nieznacznie przewyższa
01 (79.2%)

MATH-500 (Pass@1)

97.3%

Na równi z 01 (96.4%)

Codeforces (Percentyl)

96.3%

Poziom ekspercki,
porównywalny z 01

Model zachował przy tym silne zdolności ogólne i konwersacyjne, w przeciwieństwie do R1-Zero.

Dzielenie się Wiedzą: Destylowanie Rozumowania



- **DeepSeek-R1-Distill-Qwen-7B:** Przewyższa znacznie większe modele (np. GPT-4o) w zadaniach rozumowania.
- **DeepSeek-R1-Distill-Qwen-32B:** Znacząco przewyższa OpenAI 01-mini.
- **Kluczowe odkrycie:** Destylowanie wiedzy jest znacznie skuteczniejsze i tańsze obliczeniowo niż trenowanie małych modeli od zera za pomocą RL.

Zapiski z Laboratorium: Ograniczenia i Ślepe Zaułki

Nieuudane Eksperymenty



- **Process Reward Model (PRM):**
Nagradzanie każdego kroku okazało się zbyt trudne do zdefiniowania i skalowania.
- **Monte Carlo Tree Search (MCTS):**
Przestrzeń języka jest nieskończonymi większa niż plansza do gry w Go, co uniemożliwiło skuteczną eksplorację.

Obecne Ograniczenia R1



- Mniejsza wszechstronność w zadaniach ogólnych (np. function calling) niż DeepSeek V3.
- Problem mieszania języków dla zapytań spoza angielskiego/chińskiego.
- Wysoka wrażliwość na sposób formułowania promptów (zero-shot działa lepiej niż few-shot).