

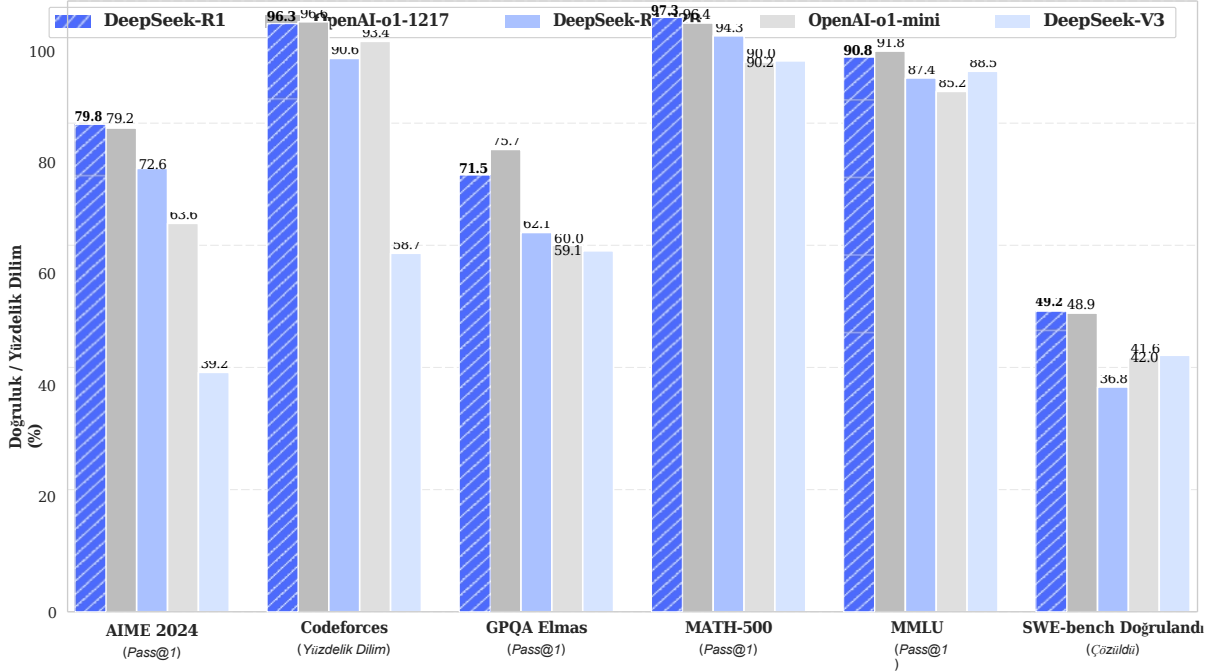
DeepSeek-R1: Takviyeli Öğrenme Yoluyla LLM'lerde Muhakeme Yeteneğinin Teşvik Edilmesi

DeepSeek-AI

research@deepseek.com

Özet

İlk nesil muhakeme modellerimiz olan DeepSeek-R1-Zero ve DeepSeek-R1'i tanıtıyoruz. Bir ön adım olarak süper videde fine-tuning (SFT) olmadan büyük ölçekli takviyeli öğrenme (RL) yoluyla eğitilen bir model olan DeepSeek-R1-Zero, dikkate değer muhakeme yetenekleri sergiliyor. RL sayesinde DeepSeek-R1-Zero doğal olarak çok sayıda güçlü ve ilgi çekici muhakeme davranışıyla ortaya çıkmaktadır. Ancak, zayıf okunabilirlik ve dil karışımı gibi zorluklarla karşılaşmaktadır. Bu sorunları ele almak ve muhakeme performansını daha da artırmak için, RL'den önce çok aşamalı eğitim ve soğuk başlangıç verilerini içeren DeepSeek-R1'i tanıtıyoruz. DeepSeek-R1, muhakeme görevlerinde OpenAI-o1-1217 ile karşılaştırılabilir bir performans elde etmektedir. Araştırma topluluğunu desteklemek için DeepSeek-R1-Zero, DeepSeek-R1 ve DeepSeek-R1'den Qwen ve Llama temelinde damıtılmış altı yoğun modeli (1.5B, 7B, 8B, 14B, 32B, 70B) açık kaynak olarak sunuyoruz.



Şekil 1| DeepSeek-R1'in kıyaslama performansı.

İçindekiler

1 Giriş	3
1.1 Katkılar	4
1.2 Değerlendirme Sonuçlarının Özeti	4
2 Yaklaşım	5
2.1 Genel Bakış	5
2.2 DeepSeek-R1-Zero: Temel Model Üzerinde Takviyeli Öğrenme	5
2.2.1 Takviyeli Öğrenme Algoritması	5
2.2.2 Ödül Modellemesi	6
2.2.3 Eğitim Şablonu	6
2.2.4 DeepSeek-R1-Zero'nun Performansı, Kendi Kendine Evrim Süreci ve Aha Anı	6
2.3 DeepSeek-R1: Soğuk Başlatma ile Takviyeli Öğrenme	9
2.3.1 Soğuk Çalıştırma	9
2.3.2 Muhakeme Odaklı Pekiştirmeli Öğrenme	10
2.3.3 Reddetme Örnekleme ve Denetimli İnce Ayar	10
2.3.4 Tüm Senaryolar için Takviyeli Öğrenme	11
2.4 Damıtma: Küçük Modelleri Muhakeme Yeteneği ile Güçlendirin	11
3 Deney	11
3.1 DeepSeek-R1 Değerlendirmesi	13
3.2 Damıtılmış Model Değerlendirmesi	14
4 Tartışma	14
4.1 Distilasyon ve Takviyeli Öğrenme	14
4.2 Başarısız Girişimler	15
5 Sonuç, Sınırlamalar ve Gelecek Çalışmalar	16
A Katkılar ve Teşekkür	20

1. Giriş

Son yıllarda, Büyük Dil Modelleri (LLM'ler) hızlı bir yineleme ve evrim geçirmekte (Anthropic, Google, 2024; 2024, 2024aile ; OpenAI) ve Yapay Genel Zeka (AGI) aradaki farkı giderek azaltmaktadır.

Son zamanlarda, eğitim sonrası, tam eğitim hattının önemli bir bileşeni olarak ortaya çıkmıştır. Ön eğitime kıyasla nispeten daha az hesaplama kaynağı gerektirirken, muhakeme görevlerinde doğruluğu artırdığı, sosyal değerlerle uyumlu olduğu ve kullanıcı tercihlerine uyum sağladığı gösterilmiştir. Muhakeme yetenekleri bağlamında, OpenAI'nin o1 (OpenAI, 2024b) serisi modelleri, Düşünce Zinciri muhakeme sürecinin uzunluğunu artırarak çıkarım zamanı ölçeklendirmesini sunan ilk model olmuştur. Bu yaklaşım matematik, kodlama ve bilimsel akıl yürütme gibi çeşitli yürütme görevlerinde önemli gelişmeler sağlamıştır. Bununla birlikte, etkili test zamanı ölçeklendirmesinin zorluğu araştırma topluluğu için açık bir soru olmaya devam etmektedir. Önceki birkaç çalışma, sürece dayalı ödül modelleri (Lightman vdKumar ..., , , 2023),) 2023; Uesato vd2022; 2024; 2024Wang .takviyeli öğrenme ((Feng , , , vdvd.vd., 2024ve Monte Carlo Ağaç Arama ve Işın Arama gibi arama algoritmaları Trinh ; Xin vd2024) . dahil olmak üzere çeşitli yaklaşımları araştırmıştırAncak, bu yöntemlerden hiçbiri OpenAI'nin o1 serisi modelleriyle karşılaştırılabilir bir genel muhakeme performansı elde edememiştir.

Bu makalede, saf takviyeli öğrenme (RL) kullanarak dil modeli muhakeme yeteneklerini geliştirmeye yönelik ilk adımı atıyoruz. Amacımız, LLM'lerin herhangi bir denetimli veri olmadan muhakeme yeteneklerini geliştirme potansiyelini keşfetmek ve saf bir RL süreci yoluyla kendi kendine evrimleşmelerine odaklanmaktır. Özellikle, DeepSeek-V3-Base'i temel model olarak kullanıyoruz ve akıl yürütmede model performansını iyileştirmek için RL çerçevesi olarak GRPO'yu (Shao ve ark., 2024) kullanıyoruz. Eğitim sırasında, DeepSeek-R1-Zero doğal olarak çok sayıda güçlü ve ilginç muhakeme davranışıyla ortaya çıktı. Binlerce RL adımından sonra DeepSeek-R1-Zero, muhakeme ölçütlerinde süper performans sergilemektedir. Örneğin, AIME 2024'teki pass@1 puanı %15,6'dan %71,0'a yükselir ve çoğunluk oylamasıyla puan daha da artarak OpenAI-o1-0912'nin performansıyla eşleşen %86,7'ye yükselir.

Ancak DeepSeek-R1-Zero, zayıf okunabilirlik ve dil karışımı gibi zorluklarla karşılaşmaktadır. Bu sorunları ele almak ve muhakeme performansını daha da artırmak için, az miktarda soğuk başlangıç verisi ve çok aşamalı bir eğitim hattı içeren DeepSeek-R1'i tanıtıyoruz. Özellikle, DeepSeek-V3-Base modeline ince ayar yapmak için binlerce soğuk başlangıç verisi toplayarak başlıyoruz. Bunu takiben, DeepSeek-R1- Zero gibi muhakeme odaklı RL gerçekleştiriyoruz. RL sürecinde yakınsamaya yaklaşıldığında, RL kontrol noktasında reddetme örnekleme yoluyla yeni SFT verileri oluşturuyor, DeepSeek-V3'ün yazma, olgusal KG ve öz biliş gibi alanlardaki denetimli verileriyle birleştiriyor ve ardından DeepSeek-V3-Base modelini yeniden eğitiyoruz. Yeni verilerle ince ayar yapıldıktan sonra, kontrol noktası tüm senaryolardan gelen istemleri dikkate alarak ek bir RL sürecinden geçer. Bu adımlardan sonra, OpenAI-o1-1217 ile eşit performans elde eden DeepSeek-R1 olarak bir kontrol noktası elde ettik.

DeepSeek-R1'den daha küçük yoğun modellere damıtmayı daha fazla araştırıyoruz. Temel model olarak Qwen2.5- 32B (Qwen, 2024b) kullanıldığında, DeepSeek-R1'den doğrudan damıtma, üzerinde RL uygulamasından daha iyi performans göstermektedir. Bu, daha büyük temel modeller tarafından keşfedilen muhakeme modellerinin muhakeme yeteneklerini geliştirmek için çok önemli olduğunu göstermektedir. Damıtılmış Qwen ve Llama (Dubey ., 2024) serilerini açık kaynak olarak sunuyoruz. Özellikle, damıtılmış 14B modelimiz, son teknoloji ürünü açık kaynaklı QwQ-32B-Preview'(Qwenden , 2024a) büyük bir farkla daha iyi performans gösteriyor ve damıtılmış 32B ve 70B modelleri, yoğun modeller arasında akıl yürütme ölçütlerinde yeni bir rekor kırıyor.

1.1. Katkılar

Eğitim Sonrası: Temel Model Üzerinde Büyük Ölçekli Takviye Öğrenimi

- Bir ön adım olarak denetimli ince ayara (SFT) güvenmeden temel modele doğrudan RL uyguluyoruz. Bu yaklaşım, modelin karmaşık problemleri çözmek için düşünce zincirini (CoT) keşfetmesine olanak tanıyarak DeepSeek-R1-Zero'nun geliştirilmesini sağlıyor. DeepSeek-R1-Zero, kendi kendini doğrulama, yansıtma ve uzun CoT'ler üretme gibi yetenekler sergileyerek araştırma topluluğu için önemli bir kilometre taşı işaret etmektedir. Özellikle, LLM'lerin muhakeme yeteneklerinin SFT'ye ihtiyaç duymadan yalnızca RL yoluyla teşvik edilebileceğini doğrulayan ilk açık çalışmadır. Bu atılım, bu alanda gelecekteki ilerlemelerin önünü açmaktadır.
- DeepSeek-R1'i geliştirmek için kullandığımız boru hattını tanıttıyoruz. Boru hattı, gelişmiş muhakeme kalıplarını keşfetmeyi ve insan tercihleriyle uyum sağlamayı amaçlayan iki RL aşamasının yanı sıra modelin muhakeme ve muhakeme dışı yetenekleri için tohum görevi gören iki SFT aşamasını içermektedir. Boru hattının daha iyi modeller yaratarak sektöre fayda sağlayacağına inanıyoruz.

Distilasyon: Daha Küçük Modeller de Güçlü Olabilir

- Daha büyük modellerin muhakeme kalıplarının daha küçük modellere damıtılabileceğini ve bunun da küçük modeller üzerinde RL yoluyla keşfedilen muhakeme kalıplarına kıyasla daha iyi performansla sonuçlanacağını gösteriyoruz. Açık kaynak DeepSeek-R1 ve API'si, gelecekte daha küçük modelleri damıtmak için araştırma topluluğuna fayda sağlayacaktır.
- DeepSeek-R1 tarafından üretilen muhakeme verilerini kullanarak, araştırma topluluğunda yaygın olarak kullanılan birkaç yoğun modele ince ayar yaptık. Değerlendirme sonuçları, damıtılmış daha küçük yoğun modellerin kıyaslama ölçütlerinde olağanüstü iyi performans gösterdiğini ortaya koymaktadır. DeepSeek-R1-Distill-Qwen-7B, AIME 2024'te QwQ-32B-Preview'i geride bırakarak %55,5'lik bir başarı elde etti. Ayrıca, DeepSeek-R1-Distill-Qwen-32B AIME 2024'te %72,6, MATH-500'de %94,3 ve LiveCodeBench'te %57,2 puan almıştır. Bu sonuçlar önceki açık kaynaklı modellerden önemli ölçüde daha iyi performans göstermektedir ve o1-mini ile karşılaştırılabilir düzeydedir. Qwen2.5 ve Llama3 serilerine dayanan damıtılmış 1.5B, 7B, 8B, 14B, 32B ve 70B kontrol noktalarını topluluğa açık kaynak olarak sunuyoruz.

1.2. Değerlendirme Sonuçlarının Özeti

- **Muhakeme görevleri:** (1) DeepSeek-R1, AIME 2024'te %79,8 Pass@1 skoru elde ederek OpenAI-o1-1217'yi biraz geride bırakmıştır. MATH-500'de, OpenAI-o1-1217 ile eşit performans göstererek ve diğer modellerden önemli ölçüde daha iyi performans göstererek %97,3'lük etkileyici bir puan elde eder. (2) Kodlama ile ilgili görevlerde DeepSeek-R1, Codeforces'ta 2.029 Elo derecesine ulaşarak yarışmadaki insan katılımcıların %96,3'ünden daha iyi performans göstererek kod yarışması görevlerinde uzmanlık seviyesini göstermektedir. Mühendislikle ilgili görevlerde DeepSeek-R1, DeepSeek-V3'ten biraz daha iyi performans göstererek geliştiricilere gerçek dünya görevlerinde yardımcı olabilir.
- **Bilgi birikimi:** MMLU, MMLU-Pro ve GPQA Diamond gibi ölçütlerde DeepSeek-R1 olağanüstü sonuçlar elde ederek MMLU'da %90,8, MMLU-Pro'da %84,0 ve GPQA Diamond'da %71,5'lik skorlarla DeepSeek-V3'ten önemli ölçüde daha iyi performans göstermiştir. Bu kıyaslamalarda performansı OpenAI-o1-1217'nin biraz altında olsa da DeepSeek-R1 diğer kapalı kaynak modellerini geride bırakarak eğitim görevlerindeki rekabet üstünlüğünü ortaya koymaktadır. Gerçeklere dayalı kıyaslama ölçütü SimpleQA'da DeepSeek-R1, DeepSeek-V3'ten daha iyi performans göstererek gerçeklere dayalı sorguları işleme kapasitesini ortaya koymuştur. Benzer bir eğilim, OpenAI-o1'in bu kıyaslamada 4o'yu geçtiği yerde de gözlenmektedir.

- **Diğerleri:** DeepSeek-R1 ayrıca yaratıcı yazma, genel soru yanıtlama, düzenleme, özetleme ve daha fazlası dahil olmak üzere çok çeşitli görevlerde de başarılıdır. AlpacaEval 2.0'da %87,6'lık etkileyici bir uzunluk kontrollü kazanma oranı ve ArenaHard'da %92,3'lük bir kazanma oranı elde ederek sınav odaklı olmayan sorguları akıllıca ele alma konusundaki güçlü yeteneğini ortaya koymuştur. Ayrıca DeepSeek-R1, uzun bağlamli kıyaslamalarında DeepSeek-V3'ten önemli ölçüde daha iyi performans göstererek uzun bağlamli anlama gerektiren görevlerde olağanüstü performans sergilemektedir.

2. Yaklaşım

2.1. Genel Bakış

Önceki çalışmalar, model performansını artırmak için büyük miktarda denetimli veriye dayanıyordu. Bu çalışmada, soğuk başlangıç olarak denetimli ince ayar (SFT) kullanmadan bile büyük ölçekli takviyeli öğrenme (RL) yoluyla muhakeme yeteneklerinin önemli ölçüde geliştirilebileceğini gösteriyoruz. Ayrıca, az miktarda soğuk başlangıç verisinin dahil edilmesiyle performans daha da artırılabilir. Aşağıdaki bölümlerde şunları sunuyoruz: (1) RL'yi herhangi bir SFT verisi olmadan doğrudan temel modele uygulayan DeepSeek-R1-Zero ve (2) Binlerce uzun Düşünce Zinciri (CoT) örneği ile ince ayarlanmış bir kontrol noktasından başlayarak RL uygulayan DeepSeek-R1. 3) DeepSeek-R1'deki muhakeme yeteneğini küçük yoğun modellere damıtın.

2.2. DeepSeek-R1-Zero: Temel Model Üzerinde Takviyeli Öğrenme

Takviyeli öğrenme, önceki çalışmalarımızda da görüldüğü gibi, muhakeme görevlerinde önemli bir etkinlik göstermiştir (Shao ve ark., 2024; Wang ve ark., 2023). Ancak, bu çalışmalar büyük ölçüde toplanması zaman alan denetimli verilere dayanmaktadır. Bu bölümde, LLM'lerin **herhangi bir denetimli veri olmadan** muhakeme yetenekleri geliştirme potansiyelini araştırıyoruz ve saf bir pekiştirmeli öğrenme süreci yoluyla kendi kendine evrimleşmelerine odaklanıyoruz. RL algoritmamıza kısa bir genel bakışla başlıyoruz, ardından bazı heyecan verici sonuçların sunumunu yapıyoruz ve bunun topluluğa değerli içgörüler sağlamasını umuyoruz.

2.2.1. Takviyeli Öğrenme Algoritması

Grup Göreceli Politika Optimizasyonu RL'nin eğitim maliyetlerinden tasarruf etmek için, Grup Göreceli Politika Optimizasyonunu (GRPO) (2024) benimsiyoruz (Shao ve ark., . . .). tipik olarak politika modeliyle aynı boyuttadır ve bunun yerine grup puanlarından temel çizgiyi tahmin eder. Özellikle, her bir soru için q , GRPO eski politikadan $\{o_1, o_2, \dots, o_G\}$ bir grup çıktıyı $\pi_{\theta_{old}}$ örnekler ve ardından aşağıdaki hedefi maksimize ederek politika modelini π_{θ} optimize eder:

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

$$\sum_{i=1}^G \min \frac{\pi_{\theta}(o_i|q)}{\pi_{\theta_{old}}(o_i|q)} A_i, \text{klip} \frac{\pi_{\theta}(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}, 1 - \epsilon, 1 + \epsilon A_i \beta_{KL} \pi_{\theta} || \pi_{ref}, \quad (1)$$

$$D_{KL} \pi_{\theta} || \pi_{ref} = \sum_{o_i|q} \pi_{\theta}(o_i|q) \log \frac{\pi_{\theta}(o_i|q)}{\pi_{ref}(o_i|q)} \quad (2)$$

Burada ϵ ve β hiper-parametrelerdir ve A_i bir grup ödül $\{r_1, r_2, \dots, r_G\}$ kullanılarak hesaplanan avantajdır. r_G her bir grup içindeki çıktılara karşılık gelir:

$$A_i = \frac{r_i - \text{mean}(\{r_1, r_2, \dots, r_G\})}{\text{std}(\{r_1, r_2, \dots, r_G\})} \quad (3)$$

Kullanıcı ve Asistan arasında bir konuşma. Kullanıcı bir soru sorar ve Asistan bunu çözer. Asistan önce akıl yürütme sürecini zihinde düşünür ve ardından kullanıcıya cevabı verir. Muhakeme süreci ve cevap `<think>` `</think>` ve `<cevap>` `</cevap>` etiketleri, sırasıyla, yani, `<düşün>` burada akıl yürütme süreci `</düşün>` `<cevap>` buraya cevap ver `</cevap>`. Kullanıcı: **komut istemi**. Asistan:

Tablo 1 | DeepSeek-R1-Zero için şablon. **istemi**, eğitim sırasında belirli bir muhakeme sorusuyla değiştirilecektir.

2.2.2. *Ödül Modellemesi*

Ödül, RL'nin optimizasyon yönüne karar veren eğitim sinyalinin kaynağıdır. DeepSeek-R1-Zero'yu eğitmek için, esas olarak iki tür ödülün oluşan kural tabanlı bir ödül sistemi benimsiyoruz:

- **Doğruluk ödülleri:** Doğruluk ödülü modeli yanıtın doğru olup olmadığını değerlendirir. Örneğin, deterministik sonuçlara sahip matematik problemleri söz konusu olduğunda, modelin nihai cevabı belirli bir formatta (örneğin, bir kutu içinde) sağlaması gerekir, bu da doğruluğun güvenilir kural tabanlı doğrulanmasını sağlar. Benzer şekilde, LeetCode problemleri için, önceden tanımlanmış test durumlarına dayalı geri bildirim üretmek üzere bir derleyici kullanılabilir.
- **Biçim ödülleri:** Doğruluk ödül modeline ek olarak, modeli düşünme sürecini '`<think>`' ve '`</think>`' etiketleri arasına koymaya zorlayan bir biçim ödül modeli kullanıyoruz.

DeepSeek-R1-Zero'yu geliştirirken sonuç veya süreç nöral ödül modelini uygulamıyoruz, çünkü nöral ödül modelinin büyük ölçekli takviye öğrenme sürecinde ödül hacklemesinden muzdarip olabileceğini ve ödül modelinin yeniden eğitilmesinin ek eğitim kaynaklarına ihtiyaç duyduğunu ve tüm eğitim hattını karmaşıklaştırdığını görüyoruz.

2.2.3. *Eğitim Şablonu*

DeepSeek-R1-Zero'yu eğitmek için, temel modeli belirtilen talimatlarımıza uyması için yönlendiren basit bir şablon tasarlayarak başlıyoruz. Tablo 1', , de gösterildiği gibiyse şablon DeepSeek-R1-Zero'nun önce bir muhakeme süreci üretmesini ardından da nihai cevabı vermesini gerektiriyor. RL süreci boyunca modelin doğal ilerleyişini doğru bir şekilde gözlemleyebilmemizi sağlamak için yansıtıcı akıl yürütmeyi zorunlu kılmak veya belirli problem çözme stratejilerini teşvik etmek gibi içeriğe özgü önyargılardan kaçınarak kısıtlamalarımızı kasıtlı olarak bu yapısal formatla sınırlandırıyoruz.

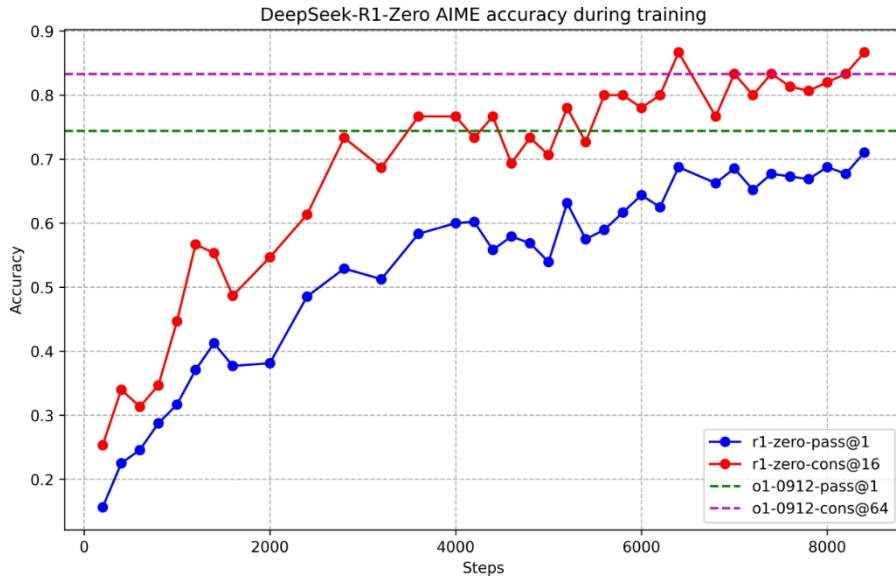
2.2.4. *DeepSeek-R1-Zero'nun Performansı, Kendi Kendine Evrim Süreci ve Aha Anı*

DeepSeek-R1-Zero'nun Performansı Şekil 2, DeepSeek-R1-Zero'nun RL eğitim süreci boyunca AIME 2024 kıyaslama ölçütü üzerindeki performans yörüngesini göstermektedir. Görüldüğü üzere DeepSeek-R1-Zero, RL eğitimi ilerledikçe performansında istikrarlı ve tutarlı bir artış göstermektedir. Özellikle, AIME 2024'teki ortalama pass@1 puanı önemli bir artış göstererek başlangıçtaki %15,6'dan etkileyici bir %71,0'a sıçramış ve OpenAI-o1-0912 ile karşılaştırılabilir performans seviyelerine ulaşmıştır. Bu önemli gelişme, RL algoritmamızın modelin performansını zaman içinde optimize etmedeki etkinliğini vurgulamaktadır.

Tablo 2, DeepSeek-R1-Zero ve OpenAI'nin o1-0912 modelleri arasında muhakeme ile ilgili çeşitli kıyaslamalarda karşılaştırmalı bir analiz sunmaktadır. Bulgular, RL'nin aşağıdakileri güçlendirdiğini ortaya koymaktadır

Model	AIME 2024		MATH-500	GPQA Elmas	LiveCode Bench	CodeForces
	pass@1	cons@64	pass@1	pass@1	pass@1	derecelendirme
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

Tablo 2 | DeepSeek-R1-Zero ve OpenAI o1 modellerinin muhakeme ile ilgili kıyaslamalarda karşılaştırılması.

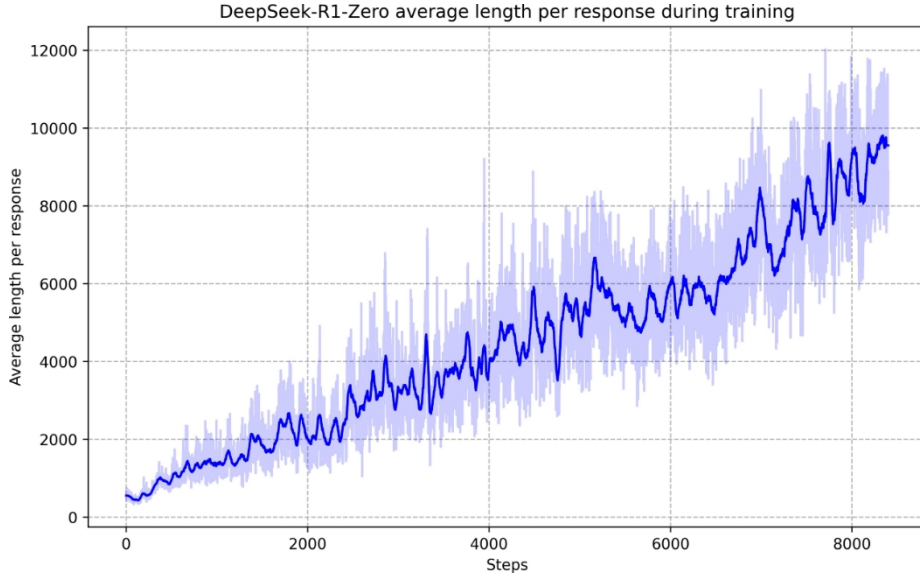


Şekil 2 Eğitim sırasında DeepSeek-R1-Zero'nun AIME doğruluğu. Her soru için 16 yanıt örnekleniyor ve istikrarlı bir değerlendirme sağlamak için genel ortalama doğruluk hesaplanıyor.

DeepSeek-R1-Zero, herhangi bir denetimli ince ayar verisine ihtiyaç duymadan sağlam muhakeme yeteneklerine ulaşmıştır. Bu, modelin yalnızca RL yoluyla etkili bir şekilde öğrenme ve genelleme yeteneğinin altını çizdiği için kayda değer bir başarıdır. Ayrıca, DeepSeek-R1-Zero'nun performansı çoğunluk oylaması uygulanarak daha da artırılabilir. Örneğin, AIME kıyaslamasında çoğunluk oylaması kullanıldığında, DeepSeek-R1-Zero'nun performansı %71,0'den %86,7'ye yükselerek OpenAI-o1-0912'nin performansını aşmaktadır. DeepSeek-R1-Zero'nun hem çoğunluk oylamasıyla hem de çoğunluk oylaması olmadan bu kadar rekabetçi bir performans elde edebilmesi, güçlü temel yeteneklerini ve akıl yürütme görevlerinde daha fazla ilerleme potansiyelini vurgulamaktadır.

DeepSeek-R1-Zero'nun Kendi Kendine Evrim Süreci DeepSeek-R1-Zero'nun kendi kendine evrim süreci, RL'nin bir modeli muhakeme yeteneklerini otonom olarak geliştirmek için nasıl yönlendirebileceğinin büyüleyici bir göstergesidir. RL'yi doğrudan temel modelden başlatarak, denetimli ince ayar aşamasının etkisi olmadan modelin ilerlemesini yakından izleyebiliyoruz. Bu yaklaşım, özellikle karmaşık muhakeme görevlerini yerine getirme kabiliyeti açısından modelin zaman içinde nasıl geliştiğine dair net bir görünüm sağlar.

Şekil 3', .te gösterildiği gibi DeepSeek-R1-Zero'nun düşünme süresi tutarlı bir iyileşme göstermektedir



Şekil 3 DeepSeek-R1-Zero'nun RL süreci boyunca eğitim setindeki ortalama yanıt uzunluğu. DeepSeek-R1-Zero doğal olarak muhakeme görevlerini daha fazla düşünme süresiyle çözmeyi öğrenir.

Eğitim süreci boyunca ment. Bu gelişme harici ayarlamaların sonucu değil, modelin kendi içindeki bir gelişmedir. DeepSeek-R1-Zero, genişletilmiş test zamanı yararlanarak giderek karmaşıklaşan muhakeme görevlerini çözme becerisini doğal olarak kazanır. Bu hesaplama, modelin düşünce süreçlerini daha derinlemesine keşfetmesine ve iyileştirmesine olanak tanıyan yüzlerce ila binlerce muhakeme belirteci üretme arasında değişir.

Bu kendi kendine evrimin en dikkat çekici yönlerinden biri, test zamanı hesaplaması arttıkça sofistike davranışların ortaya çıkmasıdır. Modelin önceki adımlarını tekrar gözden geçirdiği ve yeniden değerlendirdiği yansıtma gibi davranışlar ve problem çözmeye yönelik alternatif yaklaşımların keşfi kendiliğinden ortaya çıkar. Bu davranışlar açıkça programlanmaz, bunun yerine modelin pekiştirmeli öğrenme ortamıyla etkileşiminin bir sonucu olarak ortaya çıkar. Bu spontane gelişim, DeepSeek-R1-Zero'nun muhakeme yeteneklerini önemli ölçüde geliştirerek daha zorlu görevlerin üstesinden daha yüksek verimlilik ve doğrulukla gelmesini sağlar.

DeepSeek-R1-Zero'nun Aha Anı DeepSeek-R1-Zero'nun eğitimi sırasında gözlemlenen özellikle ilgi çekici bir olgu, bir "aha anının" ortaya çıkmasıdır. Tablo 3'te gösterildiği gibi bu an, modelin ara bir versiyonunda meydana gelmektedir. Bu aşamada DeepSeek-R1-Zero, ilk yaklaşımını yeniden değerlendirerek bir probleme daha fazla düşünme süresi ayırmayı öğrenir. Bu davranış sadece modelin büyüyen muhakeme yeteneklerinin bir kanıtı değil, aynı zamanda pekiştirmeli öğrenmenin nasıl beklenmedik ve sofistike sonuçlara yol açabileceğinin büyüleyici bir örneğidir.

Bu an sadece model için değil, aynı zamanda onun davranışını gözlemleyen araştırmacılar için de bir "aha anı". Pekiştirmeli öğrenmenin gücünün ve güzelliğinin altını çiziyor: modele bir problemi nasıl çözeceğini açıkça öğretmek yerine, ona sadece doğru teşvikleri sağlıyoruz ve o da otonom olarak gelişmiş problem çözme stratejileri geliştiriyor. Bu "aha anı", RL'nin yapay sistemlerde yeni zeka seviyelerinin kilidini açma potansiyelinin güçlü bir hatırlatıcısı olarak hizmet ediyor ve gelecekte daha otonom ve uyarlanabilir modellerin önünü açıyor.

Soru: Eğer $a > 1$ ise, o zaman gerçek çözümlerin toplamı	$\sqrt{a-x} + \sqrt{a+x} = x$	Yanıt'a
eşittir: <düşün>		
Denklemleri çözmek için	$\sqrt{a-x} + \sqrt{a+x} = x$	her ikisinin de karesini alarak başlayalım - - -
	$\sqrt{a-x} + \sqrt{a+x} = x^2 \Rightarrow \sqrt{a-x} = x^2 - \sqrt{a+x}$	
İç karekök terimini izole etmek için yeniden düzenleyin:		
	$(a-x)^2 = a+x \Rightarrow a^2 - 2ax + x^2 = a+x \Rightarrow x^4 - 2ax^2 - x + (a^2 - a) = 0$	
Bekle, Bekle, bekle. Burada işaretleyebileceğim bir aha anı var.		
Doğru toplamın olup belirlemek için bunu adım adım yeniden değerlendiririm Denklemlerle başladık:		--
		-
İlk olarak, her iki tarafın da karesini alalım:		
	$a - \sqrt{a+x} = x^2 \Rightarrow \sqrt{a+x} = a - x^2$	
Sonra, her iki tarafın da karesini alarak denklemleri yeniden ele alabiliriz:		

Tablo 3 DeepSeek-R1-Zero'nun bir ara versiyonunun ilginç bir "aha anı". Model antropomorfik bir ton kullanarak yeniden düşünmeyi öğreniyor. Bu bizim için de bir aha anıdır ve pekiştirmeli öğrenmenin gücüne ve güzelliğine tanık olmamızı sağlar.

DeepSeek-R1-Zero'nun Dezavantajı DeepSeek-R1-Zero güçlü muhakeme yetenekleri sergilemesine ve otonom olarak beklenmedik ve güçlü muhakeme davranışları geliştirmesine rağmen, çeşitli sorunlarla karşı karşıyadır. Örneğin, DeepSeek-R1-Zero zayıf okunabilirlik ve dil karıştırma gibi zorluklarla mücadele etmektedir. Muhakeme süreçlerini daha okunabilir hale getirmek ve bunları açık toplulukla paylaşmak için, insan dostu soğuk başlangıç verileriyle RL kullanan bir yöntem olan DeepSeek-R1'i keşfediyoruz.

2.3. DeepSeek-R1: Soğuk Başlatma ile Takviyeli Öğrenme

DeepSeek-R1-Zero'nun umut verici sonuçlarından esinlenerek iki doğal soru ortaya çıkmaktadır: 1) Soğuk bir başlangıç olarak az miktarda yüksek kaliteli veri dahil edilerek muhakeme performansı daha da geliştirilebilir veya yakınsama hızlandırılabilir mi? 2) Sadece net ve tutarlı Düşünce Zincirleri (CoT) üretmekle kalmayıp aynı zamanda güçlü genel yetenekler sergileyen kullanıcı dostu bir model nasıl eğitebiliriz? Bu soruları ele almak için DeepSeek-R1'i eğitmek üzere bir işlem hattı tasarladık. boru hattı, aşağıdaki gibi özetlenen dört aşamadan oluşmaktadır.

2.3.1. Soğuk Çalıştırma

DeepSeek-R1-Zero'dan farklı olarak, temel modelden RL eğitiminin erken dengesiz soğuk başlangıç aşamasını önlemek için, DeepSeek-R1 için ilk RL aktörü olarak modele ince ayar yapmak için az miktarda uzun CoT verisi oluşturuyor ve topluyoruz. Bu tür verileri toplamak için çeşitli yaklaşımlar araştırdık: örnek olarak uzun bir CoT ile birkaç atışlı yönlendirme kullanmak, modellerden yansıtma ve doğrulama ile ayrıntılı yanıtlar üretmelerini doğrudan istemek, DeepSeek-R1-Zero çıktılarını okunabilir bir formatta toplamak ve sonuçları insan açıklayıcılar tarafından sonradan işleyerek iyileştirmek.

Bu çalışmada, RL'nin başlangıç noktası olarak DeepSeek-V3-Base'e ince ayar yapmak için binlerce soğuk başlangıç verisi topluyoruz. DeepSeek-R1-Zero ile karşılaştırıldığında, soğuk başlangıç verilerinin avantajları

dahil:

- Okunabilirlik: DeepSeek-R1-Zero'nun önemli bir sınırlaması, içeriğinin genellikle okumaya uygun olmamasıdır. Yanıtlar birden fazla dili karıştırabilir veya kullanıcılar için yanıtları vurgulamak için markdown biçimlendirmesinden yoksun olabilir. Buna karşılık, DeepSeek-R1 için soğuk başlangıç verileri oluştururken, her yanıtın sonunda bir özet içeren ve okuyucu dostu olmayan yanıtları filtreleyen okunabilir bir model tasarlıyoruz. Burada çıktı formatını şu şekilde tanımlıyoruz
|special_token|<reasoning_process>|special_token|<summary>, burada akıl yürütme süreci sorgunun CoT'sidir ve özet akıl yürütme sonuçlarını özetlemek için kullanılır.
- Potansiyel: İnsan öncülleri ile soğuk başlangıç verileri için modeli dikkatlice tasarlayarak, DeepSeek-R1-Zero'ya karşı daha iyi performans gözlemliyoruz. Yinelemeli eğitimin akıl yürütme modelleri için daha iyi bir yol olduğuna inanıyoruz.

2.3.2. Muhakeme Odaklı Pekiştirmeli Öğrenme

DeepSeek-V3-Base'i soğuk başlangıç verileri üzerinde ince ayar yaptıktan sonra, DeepSeek-R1-Zero'da kullanılan aynı büyük ölçekli takviyeli öğrenme eğitim sürecini uyguluyoruz. Bu aşama, özellikle kodlama, matematik, bilim ve mantık muhakemesi gibi net çözümleri olan iyi tanımlanmış problemleri içeren muhakeme yoğun görevlerde modelin muhakeme yeteneklerini geliştirmeye odaklanmaktadır. Eğitim sürecinde, CoT'un özellikle RL istemleri birden fazla dil içerdiğinde sıklıkla dil karışımı sergilediğini gözlemliyoruz. Dil karışımı sorununu hafifletmek için, RL eğitimi sırasında CoT'deki hedef dil kelimelerinin oranı olarak hesaplanan bir dil tutarlılığı ödülü sunuyoruz. Ablasyon deneyleri, bu tür bir hizalamanın modelin performansında hafif bir düşüşe neden olduğunu gösterse de, bu ödül insan tercihleriyle uyumludur ve daha okunabilir hale getirir. Son , muhakeme görevlerinin doğruluğunu ve dil tutarlılığı ödülünü, nihai oluşturmak için doğrudan toplayarak birleştiriyoruz. Daha sonra, muhakeme görevlerinde yakınsama elde edene kadar ince ayarlı model üzerinde RL eğitimi uyguluyoruz.

2.3.3. Reddetme Örneklemesi ve Denetimli İnce Ayar

Muhakeme odaklı RL yakınsadığında, bir sonraki tur için SFT (Supervised Fine-Tuning) verilerini toplamak üzere ortaya çıkan kontrol noktasını kullanırız. Öncelikle akıl yürütmeye odaklanan ilk soğuk başlangıç verilerinin aksine, bu aşama modelin yazma, rol yapma ve diğer genel amaçlı görevlerdeki yeteneklerini geliştirmek için diğer alanlardan gelen verileri içerir. Özellikle, aşağıda açıklandığı gibi verileri oluşturuyor ve modele ince ayar yapıyoruz.

Akıl yürütme **verileri** Yukarıdaki RL eğitiminden elde edilen kontrol noktasından ret örneklemesi yaparak akıl yürütme istemlerini seçiyor ve akıl yürütme yörüngeleri oluşturuyoruz. Önceki aşamada, yalnızca kural tabanlı ödüller kullanılarak değerlendirilebilecek verileri dahil ettik. Ancak bu aşamada, veri kümesini, bazıları temel gerçeği ve model tahminlerini değerlendirme için DeepSeek-V3'e besleyerek üretken bir ödül modeli kullanan ek verileri dahil ederek genişletiyoruz. Ek olarak, model çıktısı bazen kaotik ve okunması zor olduğundan, karışık diller, uzun paragraflar ve kod blokları içeren düşünce zincirlerini filtreledik. Her komut istemi için birden fazla yanıt örnekliyor ve yalnızca doğru olanları saklıyoruz. Toplamda, yaklaşık 600 bin akıl yürütme ile ilgili eğitim örneği topluyoruz.

Muhakeme **dışı veriler** Yazma, olgusal KG, kendi kendine bilme ve çeviri gibi muhakeme dışı veriler için DeepSeek-V3 işlem hattını benimsiyor ve DeepSeek-V3'ün SFT veri kümesinin bazı kısımlarını yeniden kullanıyoruz. Muhakeme gerektirmeyen belirli görevler için, DeepSeek-V3'ü soruyu sorarak yanıtlamadan önce potansiyel bir düşünce zinciri oluşturması için çağırıyoruz. Ancak, "hello" gibi daha basit sorgular için yanıt olarak bir CoT sağlamıyoruz. Sonuç olarak, muhakeme ile ilgisi olmayan toplamda yaklaşık 200 bin eğitim örneği topladık.

DeepSeek-V3-Base'i, yaklaşık 800 bin örnekten oluşan yukarıdaki küratörlü veri setini kullanarak iki dönem için ince ayar yapıyoruz.

2.3.4. Tüm Senaryolar için Takviyeli Öğrenme

Modeli insan tercihleriyle daha da uyumlu hale getirmek için, modelin yardımseverliğini ve zararsızlığını geliştirmeyi ve aynı zamanda muhakeme yeteneklerini geliştirmeyi amaçlayan ikincil bir takviye öğrenme aşaması uyguluyoruz. Spesifik olarak, modeli ödül sinyalleri ve çeşitli uyarı dağılımlarının bir kombinasyonunu kullanarak eğitiyoruz. Muhakeme verileri için, matematik, kod ve mantıksal muhakeme alanlarında öğrenme sürecini yönlendirmek için kural tabanlı ödüller kullanan DeepSeek-R1-Zero'da özetlenen metodolojiye bağlı kalıyoruz. Genel veriler için, karmaşık ve incelikli senaryolarda insan tercihlerini yakalamak için ödül modellerine başvuruyoruz. DeepSeek-V3 boru hattını temel alıyoruz ve benzer bir tercih çifti ve eğitim ipucu dağılımını benimsiyoruz. Yardımseverlik için, yalnızca nihai özete odaklanıyoruz ve değerlendirmenin, temel muhakeme sürecine müdahaleyi en aza indirirken yanıtın kullanıcı için faydasını ve alaka düzeyini vurgulamasını sağlıyoruz. Zararsızlık için, oluşturma sürecinde ortaya çıkabilecek potansiyel riskleri, önyargıları veya zararlı içeriği belirlemek ve azaltmak için hem muhakeme süreci hem de özet dahil olmak üzere modelin tüm yanıtını değerlendiriyoruz. Sonuç olarak, ödül sinyallerinin ve çeşitli veri dağılımlarının entegrasyonu, yardımseverliğe ve zararsızlığa öncelik verirken muhakemede üstün olan bir modeli eğitmemizi sağlıyor.

2.4. Damıtma: Küçük Modelleri Muhakeme Yeteneği ile Güçlendirin

Daha verimli küçük modelleri DeepSeek-R1 gibi muhakeme yetenekleriyle donatmak için, Qwen (Qwen, 2024b) ve Llama (AI@Meta, 2024) gibi açık kaynaklı modellere, §2.3.3' te ayrıntılı olarak açıklandığı gibi DeepSeek-R1 ile küratörlüğünü yaptığımız 800 bin örneği kullanarak doğrudan ince ayar yaptık. Bulgularımız, bu basit damıtma yönteminin daha küçük modellerin muhakeme yeteneklerini önemli ölçüde geliştirdiğini göstermektedir. Burada kullandığımız temel modeller Qwen2.5-Math-1.5B, Qwen2.5-Math-7B, Qwen2.5- 14B, Qwen2.5-32B, Llama-3.1-8B ve Llama-3.3-70B-Instruct'tir. Llama-3.3'ü seçmemizin nedeni, muhakeme yeteneğinin Llama-3.1'den biraz daha iyi olmasıdır.

Damıtılmış modeller için yalnızca SFT uyguluyoruz ve RL'nin dahil edilmesi model performansını önemli ölçüde artırabilecek olsa da RL aşamasını dahil etmiyoruz. Buradaki birincil amacımız, damıtma tekniğinin etkinliğini göstermek ve RL aşamasının keşfini daha geniş araştırma topluluğuna bırakmaktır.

3. Deney

Benchmarklar Modelleri MMLU (Hendrycks vd., 2020), MMLU-Redux (Gema, 2024), MMLU-Pro ((Li vd., Wang vd., Huang vd., C-Eval, 2023) ve CMMLU (2023)), IFEval (Zhou vd., 2023), GPQA Diamond ((OpenAI/Rein SimpleQA C-SimpleQA He vd.)SWE-Bench Verified (Open,

2024d), Aider ¹, LiveCodeBench (Jain vd., 2024) (2024-08 - 2025-01, Codeforces ((MAA, ²), Çin Ulusal Lise Matematik Olimpiyatı (CNMO 2024)⁽³⁾ ve American Invitational Mathematics Examination 2024 (AIME) 2024.)Standart ölçütlere ek olarak, modellerimizi LLM'leri jüri olarak kullanan açık uçlu üretim görevleri üzerinde de değerlendiriyoruz. Özellikle, ikili karşılaştırmalar için GPT-4-Turbo-1106'yı jüri olarak kullanan AlpacaEval 2.0 (Dubois ., 2024) ve Arena-Hard'vd., ın (Li 2024) orijinal konfigürasyonlarına bağlı kalıyoruz. Burada, uzunluk yanlılığından kaçınmak için değerlendirmeye yalnızca nihai özeti besliyoruz. Damıtılmış modeller için AIME 2024, MATH-500, GPQA Diamond, Codeforces ve LiveCodeBench üzerindeki temsili sonuçları rapor ediyoruz.

Değerlendirme İstemleri DeepSeek-V3'teki kurulumu takiben, MMLU, DROP, GPQA Diamond ve SimpleQA gibi standart kıyaslamalar simple-evals çerçevesindeki istemler kullanılarak değerlendirilir. MMLU-Redux için, sıfır atış ortamında Zero-Eval istem formatını (Lin, 2024) benimsiyoruz. MMLU-Pro, C-Eval ve CLUE-WSC açısından, orijinal istemler az çekim olduğundan, istemi sıfır çekim ayarına göre biraz değiştiriyoruz. Az çekimdeki CoT, DeepSeek-R1'in performansına zarar verebilir. Diğer veri kümeleri, yaratıcıları tarafından sağlanan varsayılan istemlerle orijinal değerlendirme protokollerini takip eder. Kod ve matematik kıyaslamaları için HumanEval-Mul veri kümesi sekiz ana programlama dilini (Python, Java, C++, C#, JavaScript, TypeScript, PHP ve Bash) kapsamaktadır. LiveCodeBench üzerindeki model performansı, Ağustos 2024 ile Ocak 2025 arasında toplanan verilerle CoT formatı kullanılarak değerlendirilmiştir. Codeforces veri seti, 10 Div.2 yarışmasından alınan problemler ve uzmanlar tarafından hazırlanmış test kullanılarak değerlendirilmiş, ardından beklenen derecelendirmeler ve rakiplerin yüzdeleri hesaplanmıştır. SWE-Bench onaylı sonuçlar aracısız çerçeve aracılığıyla elde edilir (Xia vd., 2024). AIDER ile ilgili kıyaslamalar "diff" formatı kullanılarak ölçülür. DeepSeek-R1 çıktıları her bir kıyaslama için maksimum 32.768 token ile sınırlandırılmıştır.

Temeller DeepSeek-V3, Claude-Sonnet-3.5-1022, GPT-4o-0513, OpenAI-o1-mini ve OpenAI-o1-1217 dahil olmak üzere çeşitli güçlü temellere karşı kapsamlı değerlendirmeler yapıyoruz. OpenAI-o1-1217 API'sine erişim Çin anakarasında zor olduğundan, performansını resmi raporlara dayanarak bildiriyoruz. Damıtılmış modeller için açık kaynaklı QwQ-32B-Preview modelini de karşılaştırıyoruz (Qwen, 2024a).

Değerlendirme Kurulumu Modeller için maksimum üretim uzunluğunu 32.768 token olarak belirledik. Uzun çıkışlı muhakeme modellerini değerlendirmek için açgözlü kod çözme kullanmanın daha yüksek tekrarlamaya oranlarına ve farklı kontrol noktalarında önemli değişkenliğe neden olduğunu gördük. Bu nedenle, pass@k değerlendirmesini varsayılan olarak kullanıyoruz (Chen ve ark., 2021) ve sıfır olmayan bir sıcaklık kullanarak pass@1'i rapor ediyoruz. Özellikle, her soru için k yanıtı (test seti boyutuna bağlı olarak tipik olarak 4 ila 64 arasında) oluşturmak için 0,6 örnekleme sıcaklığı ve 0,95 üstü değeri kullanıyoruz. Pass@1 daha sonra şu şekilde hesaplanır

$$\text{pass}@1 = \frac{1}{k} \sum_{i=1}^k p_i,$$

Burada p_i i -inci yanıtın doğruluğunu ifade eder. Bu yöntem daha güvenilir performans tahminleri sağlamaktadır. AIME 2024 için, cons@64 olarak gösterilen 64 örnek kullanarak konsensüs (çoğunluk oyu) sonuçlarını da rapor ediyoruz (Wang vd., 2022).

⁽¹⁾<https://aider.chat>

⁽²⁾<https://codeforces.com>

⁽³⁾<https://www.cms.org.cn/Home/comp/comp/cid/12.html>

3.1. DeepSeek-R1 Değerlendirmesi

Karşılaştırma Ölçütü (Metrik)	Claude-3.5- GPT-4o DeepSeek			OpenAI OpenAI		DeepSeek R1	
	Sonnet-1022	0513	V3	o1-mini	o1-1217		
Mimarlık	-	-	Çevre ve Şehircilik Bakanlığı	-	-	Çevre ve Şehircilik Bakanlığı	
# Etkinleştirilmiş Parametreler	-	-	37B	-	-	37B	
# Toplam Parametreler	-	-	671B	-	-	671B	
İngilizce	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 atış 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 Elmas (Pass@1)	65.0	49.9	59.1	60.0	75.7	71.5
	SimpleQA (Doğru)	28.4	38.2	24.9	7.0	47.0	30.1
	ÇERÇEVELER (Acc.)	72.5	80.5	73.3	76.9	-	82.5
	AlpacaEval2.0 (LC-winsrate)	52.0	51.1	70.0	57.8	-	87.6
	ArenaHard (GPT-4-1106)	85.2	80.4	85.5	92.0	-	92.3
Kod	LiveCodeBench (Pass@1-COT)	38.9	32.9	36.2	53.8	63.4	65.9
	Kod Güçleri (Yüzdelik)	20.3	23.6	58.7	93.4	96.6	96.3
	Codeforces (Değerlendirme)	717	759	1134	1820	2061	2029
	SWE Doğrulandı (Çözüldü)	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
Matematik	AIME 2024 (Pass@1)	16.0	9.3	39.2	63.6	79.2	79.8
	MATH-500 (Geçer@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
Çin C-Eval (EM)	CLUEWSC (EM)	85.4	87.9	90.9	89.9	-	92.8
	Çin C-Eval (EM)	76.7	76.0	86.5	68.9	-	91.8
	C-SimpleQA (Doğru)	55.4	58.7	68.0	40.3	-	63.7

Tablo 4| DeepSeek-R1 ile diğer temsili modeller arasındaki karşılaştırma.

MMLU, MMLU-Pro ve GPQA Diamond gibi eğitim odaklı bilgi ölçütleri için DeepSeek-R1, DeepSeek-V3'e kıyasla üstün performans göstermektedir. Bu iyileşme öncelikle, büyük ölçekli pekiştirmeli öğrenme yoluyla önemli kazanımların elde edildiği STEM ile ilgili sorulardaki gelişmiş doğruluğa bağlanmaktadır. Ayrıca DeepSeek-R1, uzun bağlama bağlı bir KG görevi olan FRAMES'te üstünlük göstererek güçlü belge analizi yeteneklerini ortaya koymuştur. Bu, yapay zeka odaklı arama ve veri analizi görevlerinde muhakeme modellerinin potansiyelini vurgulamaktadır. Gerçeklere dayalı SimpleQA kıyaslamasında DeepSeek-R1, DeepSeek-V3'ten daha iyi performans göstererek gerçeklere dayalı sorguları işleme kabiliyetini ortaya koymaktadır. Benzer bir eğilim, OpenAI-o1'in bu kıyaslamada GPT-4o'yu geride bıraktığı durumlarda da gözlemlenmektedir. Bununla birlikte, DeepSeek-R1, Çin SimpleQA kıyaslamasında DeepSeek-V3'ten daha kötü performans göstermektedir, bunun başlıca nedeni güvenlik RL'sinden sonra belirli sorguları yanıtlamayı reddetme eğilimidir. Güvenlik RL'si olmadan, DeepSeek-R1 %70'in üzerinde bir doğruluğa ulaşabilir.

DeepSeek-R1, bir modelin format talimatlarını takip etme becerisini değerlendirmek için tasarlanmış bir ölçüt olan IF-Eval'de de etkileyici sonuçlar veriyor. Bu gelişmeler, denetimli ince ayar (SFT) ve RL eğitiminin son aşamalarında talimat takip verilerinin dahil edilmesiyle ilişkilendirilebilir. Ayrıca, AlpacaEval2.0 ve ArenaHard'da kayda değer bir performans gözlemlenmiştir, bu da DeepSeek-R1'in yazma görevlerinde ve açık alan soru cevaplama güçlü olduğunu göstermektedir. DeepSeek-V3'ten önemli ölçüde daha iyi performans göstermesi, yalnızca

muhakeme yeteneklerini artırmakla kalmayıp aynı zamanda farklı alanlardaki performansı da geliştiren büyük ölçekli RL'nin genelleme avantajlarının altını çizmektedir. Ayrıca, DeepSeek-R1 tarafından üretilen özet uzunlukları, ArenaHard'da ortalama 689 belirteç ve AlpacaEval 2.0'da 2.218 karakter ile kısadır. Bu da şunu göstermektedir

DeepSeek-R1, GPT tabanlı değerlendirmeler sırasında uzunluk yanlılığının ortaya çıkmasını önleyerek birden fazla görevde sağlamlığını daha da sağlamlaştırır.

Matematik görevlerinde DeepSeek-R1, OpenAI-o1-1217 ile eşit performans göstererek diğer modelleri büyük bir farkla geride bırakıyor. Benzer bir eğilim, LiveCodeBench ve Codeforces gibi kodlama algoritması görevlerinde de gözlemlenmekte olup, muhakeme odaklı modeller bu kıyaslamalara hakimdir. Mühendislik odaklı kodlama görevlerinde OpenAI-o1-1217, Aider'da DeepSeek-R1'den daha iyi performans gösterirken SWE Verified'da karşılaştırılabilir bir performans elde ediyor. DeepSeek-R1'in mühendislik performansının bir sonraki sürümde artacağına inanıyoruz, çünkü ilgili RL eğitim verilerinin miktarı şu anda çok sınırlı kalmaktadır.

3.2. Damıtılmış Model Değerlendirmesi

Model	AIME 2024		MATH 500	GPQA Elmas	LiveCode Bench	CodeForces
	pass@1	cons@64	pass@1	pass@1	pass@1	derecelendirme
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-İnceleme	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

Tablo 5 | DeepSeek-R1 damıtılmış modellerinin ve diğer karşılaştırılabilir modellerin muhakeme ile ilgili kıyaslamalarda karşılaştırılması.

Tablo 5'te gösterildiği gibi, DeepSeek-R1'in çıktılarını damıtmak, verimli DeepSeek-R1-7B'nin (yani DeepSeek-R1-Distill-Qwen-7B, aşağıda benzer şekilde kısaltılmıştır) GPT-4o-0513 gibi akıllı yürütmeyen modellerden daha iyi performans sağlar. DeepSeek-R1-14B tüm değerlendirme ölçütlerinde QwQ-32B-Preview'i geçerken, DeepSeek-R1-32B ve DeepSeek-R1-70B çoğu kıyaslamada o1-mini'yi önemli ölçüde aşmaktadır. Bu sonuçlar distillasyon'un güçlü potansiyelini göstermektedir. Ayrıca, bu damıtılmış modellere RL uygulamanın önemli kazanımlar sağladığını gördük. Bunun daha fazla araştırılması gerektiğine inanıyoruz ve bu nedenle burada yalnızca basit SFT ile damıtılmış modellerin sonuçlarını sunuyoruz.

4. Tartışma

4.1. Distilasyon ve Takviyeli Öğrenme

Bölüm 3.2'de, DeepSeek-R1'i damıtarak küçük modelin etkileyici sonuçlar elde edebileceğini görebiliyoruz. Ancak hala bir soru var: Model, damıtma olmadan makalede tartışılan büyük ölçekli RL eğitimi yoluyla karşılaştırılabilir performans elde edebilir mi?

Bu soruyu yanıtlamak için, matematik, kod ve STEM verilerini kullanarak Qwen-32B-Base üzerinde büyük ölçekli RL eğitimi gerçekleştirdik, 10K'dan fazla adım için eğitim yaptık ve DeepSeek-R1-Zero-Qwen-32B'yi elde ettik. Tablo 6'da gösterilen deneysel sonuçlar, 32B temel modelinin büyük ölçekli eğitimden sonra

Model	AIME 2024		MATH-500	GPQA Elmas	LiveCodeBench
	pass@1	cons@64	pass@1	pass@1	pass@1
QwQ-32B-İnceleme	50.0	60.0	90.6	54.5	41.9
DeepSeek-R1-Zero-Qwen-32B	47.0	60.0	91.6	55.0	40.2
DeepSeek-R1-Distill-Qwen-32B	72.6	83.3	94.3	62.1	57.2

Tablo 6| Damıtılmış ve RL Modellerinin Akıl Yürütme ile İlgili Üzerindeki Karşılaştırması.

RL eğitimi, QwQ-32B-Preview ile eşit performans elde eder. Bununla birlikte, DeepSeek-R1'den damıtılan DeepSeek-R1-Distill-Qwen-32B, tüm kıyaslamalarda DeepSeek-R1-Zero-Qwen-32B'den önemli ölçüde daha iyi performans göstermektedir.

Bu nedenle iki sonuç çıkarabiliriz: Birincisi, daha güçlü modellerin daha küçük modellere damıtılması mükemmel sonuçlar verirken, bu makalede bahsedilen büyük ölçekli RL'ye dayanan daha küçük modeller muazzam hesaplama gücü gerektirir ve damıtma performansına bile ulaşamayabilir. İkinci olarak, damıtma stratejileri hem ekonomik hem de etkili olsa da, zeka sınırlarının ötesine geçmek için daha güçlü temel modeller ve daha büyük ölçekli takviye öğrenme gerekebilir.

4.2. Başarısız Girişimler

DeepSeek-R1'i geliştirmenin ilk aşamalarında, yol boyunca başarısızlıklar ve aksiliklerle de karşılaştık. Burada başarısızlık deneyimlerimizi içgörü sağlamak için paylaşıyoruz, ancak bu, bu yaklaşımların etkili muhakeme modelleri geliştirmede yetersiz olduğu anlamına gelmez.

Süreç Ödül Modeli (PRM) PRM, (Lightman modeli muhakeme görevlerini çözmeye yönelik daha iyi yaklaşımlara yönlendirmek için makul bir yöntemdir vdUesato vd, 2023; ., Wang vd., 2022; 2023). Bununla birlikte, uygulamada, PRM'nin nihai başarısını engelleyebilecek üç ana sınırlaması vardır. İlk olarak, genel muhakemede ince taneli bir adımı açıkça tanımlamak zordur. İkincisi, mevcut ara adımın doğru olup olmadığını belirlemek zorlu bir görevdir. Modeller kullanarak yapılan otomatik açıklama tatmin edici sonuçlar vermeyebilirken, manuel açıklama da ölçek elverişli değildir. Üçüncüsü, model tabanlı bir PRM kullanılmaya , kaçınılmaz olarak ödülün hacklenmesine yol açar (Gao ., 2022) ve ödül modelinin yeniden eğitilmesi ek eğitim kaynaklarına ihtiyaç duyar ve tüm eğitim hattını karmaşıklaştırır. Sonuç olarak, PRM, vdmodel tarafından üretilen en iyi N yanıtlarını yeniden sıralama veya yönlendirmeli aramaya yardımcı olma konusunda iyi bir yetenek gösterse de ., 2024), (Snell deneylerimizdeki büyük ölçekli pekiştirmeli öğrenme süreci sırasında getirdiği ek hesaplama ek yüküne kıyasla avantajları sınırlıdır.

Monte Carlo Tree Search (MCTS) AlphaGo (Silver vd., 2017b) ve AlphaZero'(Sil- dan ver vd., 2017a) esinlenerek, test zamanı hesaplama ölçeklenebilirliğini geliştirmek için Monte Carlo Tree Search (MCTS) kullanmayı araştırdık. Bu yaklaşım, modelin çözüm uzayını sistematik olarak keşfetmesini sağlamak için cevapları daha küçük parçalara ayırmayı içerir. Bunu kolaylaştırmak için modelden, arama için gerekli olan belirli muhakeme adımlarına karşılık gelen birden fazla etiket oluşturmasını istiyoruz. Eğitim için, ilk olarak önceden eğitilmiş bir değer modeli tarafından yönlendirilen MCTS aracılığıyla yanıtları bulmak için toplanan istemleri kullanıyoruz. Daha sonra, ortaya çıkan soru-cevap çiftlerini hem aktör modelini hem de değer modelini eğitmek için kullanıyoruz ve süreci yinelemeli olarak iyileştiriyoruz.

Ancak bu yaklaşım, eğitimi ölçeklendirirken çeşitli zorluklarla karşılaşır. İlk olarak, arama uzayının nispeten iyi tanımlandığı satrancın aksine, token üretimi

üstel olarak daha büyük arama uzayı. Bunu ele almak için her düğüm için bir maksimum uzatma limiti belirledik, ancak bu modelin yerel optimumlarda takılıp kalmasına yol açabilir. İkinci olarak, değer modeli, arama sürecinin her adımına rehberlik ettiği için üretim kalitesini doğrudan etkiler. İnce taneli bir değer modelini eğitmek doğası gereği zordur, bu da modelin iteratif olarak gelişmesini zorlaştırır. AlphaGo'nun temel başarısı, performansını aşamalı olarak artırmak için bir değer modelinin eğitilmesine dayanırken, token üretiminin karmaşıklığı nedeniyle bu ilkenin bizim düzenimizde tekrarlanması zor görünmektedir.

Sonuç olarak, MCTS önceden eğitilmiş bir değer modeliyle eşleştirildiğinde çıkarım sırasında performansı artırabilirken, kendi kendine arama yoluyla model performansını yinelemeli olarak artırmak önemli bir zorluk olmaya devam etmektedir.

5. Sonuç, Sınırlamalar ve Gelecek Çalışmalar

Bu çalışmada, pekiştirmeli öğrenme yoluyla model muhakeme yeteneklerini geliştirme yolculuğumuzu paylaşıyoruz. DeepSeek-R1-Zero, soğuk başlangıç verilerine dayanmayan saf bir RL yaklaşımını temsil eder ve çeşitli görevlerde güçlü performans elde eder. DeepSeek-R1, yinelemeli RL ince ayarının yanı sıra soğuk başlangıç verilerinden yararlanarak daha güçlüdür. Sonuç olarak DeepSeek-R1, bir dizi görevde OpenAI-o1-1217 ile karşılaştırılabilir performans elde eder.

Muhakeme yeteneğini küçük yoğun modellere damıtmayı daha fazla araştırıyoruz. Öğretmen model olarak DeepSeek-R1'i kullanarak 800 bin eğitim örneği oluşturuyor ve birkaç küçük yoğun ince ayar yapıyoruz. Sonuçlar umut verici: DeepSeek-R1-Distill-Qwen-1.5B, AIME'de %28,9 ve MATH'de %83,9 ile matematik kıyaslamalarında GPT-4o ve Claude-3.5-Sonnet'ten daha iyi performans gösteriyor. Diğer yoğun modeller de etkileyici sonuçlar elde ederek aynı temel kontrol noktalarına dayanan diğer komut ayarlı modellerden önemli ölçüde daha iyi performans göstermiştir.

Gelecekte, DeepSeek-R1 için aşağıdaki yönlerde araştırmalara yatırım yapmayı planlıyoruz.

- **Genel Yetenekler:** Şu anda DeepSeek-R1'in yetenekleri, fonksiyon çağırma, çoklu dönüş, karmaşık rol yapma ve JSON çıktısı gibi görevlerde DeepSeek-V3'ün gerisinde kalmaktadır. İleriye dönük olarak, bu alanlardaki görevleri geliştirmek için CoT'den ne kadar uzun süre yararlanılabileceğini araştırmayı planlıyoruz.
- **Dil Karışımı:** DeepSeek-R1 şu anda Çince ve İngilizce için optimize edilmiştir, bu da diğer dillerdeki sorguları işlerken dil karışımı sorunlarına neden olabilir. Örneğin, DeepSeek-R1, sorgu İngilizce veya Çince dışında bir dilde olsa bile muhakeme ve yanıtlar için İngilizceyi kullanabilir. Gelecekteki güncellemelerde bu sınırlamayı ele almayı hedefliyoruz.
- **Yönlendirme Mühendisliği:** DeepSeek-R1'i değerlendirirken, istemlere karşı hassas olduğunu gözlemledik. Az sayıda komut istemi sürekli olarak performansını düşürmektedir. Bu nedenle, kullanıcıların sorunu doğrudan tanımlamalarını ve en iyi sonuçlar için sıfır çekim arayarı kullanarak çıktı biçimini belirtmelerini öneririz.
- **Yazılım Mühendisliği Görevleri:** RL sürecinin verimliliğini etkileyen uzun değerlendirme süreleri nedeniyle, büyük ölçekli RL yazılım mühendisliği görevlerinde yaygın olarak uygulanmamıştır. Sonuç olarak DeepSeek-R1, yazılım mühendisliği kıyaslamalarında DeepSeek-V3'e göre büyük bir gelişme gösterememiştir. Gelecek sürümler, yazılım mühendisliği verileri üzerinde reddetme örnekleme uygulayarak veya verimliliği artırmak için RL süreci sırasında eşzamansız değerlendirmeler ekleyerek bu sorunu çözecektir.

Referanslar

AI@Meta. Llama 3.1 model kartı, 2024. URL https://github.com/meta-llama/llama-models/blob/main/models/llama3_1/MODEL_CARD.md.

Antropik. Claude 3.5 sone, 2024. URL <https://www.anthropic.com/news/claude-3-5-sonnet>.

M. Chen, J. Tworek, H. Jun, Q. Yuan, H. P. de Oliveira Pinto, J. Kaplan, . Edwards, Y. Burda, N. Joseph, G. Brockman, A. Ray, R. Puri, G. Krueger, M. Petrov, H. Khlaaf, G. Sastry, P. Mishkin, B. Chan, S. Gray, N. Ryder, M. Pavlov, A. Power, L. Kaiser, M. Bavarian, C. Winter, P. Tillet, F. P. Such, D. Cummings, M. Plappert, F. Chantzis, E. Barnes, A. Herbert-Voss, W. H. Guss, A. Nichol, A. Paino, N. Tezak, J. Tang, I. Babuschkin, S. Balaji, S. Jain, W. Saunders, C. Hesse, A. N. Carr, J. Leike, J. Achiam, V. Misra, E. Morikawa, A. Radford, M. Knight, M. Brundage, M. Murati, K. Mayer, P. Welinder, B. McGrew, D. Amodei, S. McCandlish, I. Sutskever ve W. Zaremba. Kod üzerinde eğitilmiş büyük dil modellerinin değerlendirilmesi. *CoRR*, abs/2107.03374, 2021. URL <https://arxiv.org/abs/2107.03374>.

A. Dubey, . Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, vd. lama 3 model sürüsü. *arXiv ön baskı arXiv:2407.21783*, 2024.

Y. Dubois, B. Galambosi, P. Liang ve T. B. Hashimoto. Uzunluk kontrollü alpakaeval: Otomatik değerlendiricileri debias etmenin basit bir yolu. *arXiv ön baskı arXiv:2404.04475*, 2024.

X. Feng, Z. Wan, M. Wen, S. M. McAleer, Y. Wen, W. Zhang ve J. Wang. Alphazero benzeri ağaç arama, büyük dil modeli kod çözme ve eğitimine rehberlik edebilir, 2024. URL <https://arxiv.org/abs/2309.17179>.

L. Gao, J. Schulman ve J. Hilton. Ödül modeli aşırı optimizasyonu için ölçeklendirme yasaları, 2022. URL <https://arxiv.org/abs/2210.10760>.

A. P. Gema, J. O. J. Leang, . Hong, A. Devoto, A. C. M. Mancino, R. Saxena, X. He, Y. Zhao, X. Du, M. R. G. Madani, C. Barale, R. McHardy, J. Harris, J. Kaddour, E. van Krieken ve P. Minervini. Mmlu ile işimiz bitti mi? *CoRR*, abs/2406.04127, 2024. URL <https://doi.org/10.48550/arXiv.2406.04127>.

Google. Yeni nesil modelimiz: Gemini 1.5, 2024. URL <https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024>.

Y. He, S. Li, J. Liu, Y. Tan, W. Wang, H. Huang, X. Bu, H. Guo, C. Hu, B. Zheng, et al. Chi-nese simpleqa: Büyük dil modelleri için Çince olgusal değerlendirme. *arXiv ön baskı arXiv:2411.07140*, 2024.

D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song ve J. Steinhardt. Devasa çoklu görev dil anlayışının ölçülmesi. *arXiv ön baskı arXiv:2009.03300*, 2020.

Y. Huang, Y. Bai, Z. Zhu, J. Zhang, J. Zhang, T. Su, J. Liu, C. Lv, Y. Zhang, J. Lei, . C-Eval: Temel modeller için çok seviyeli, çok disiplinli bir Çince değerlendirme paketi. *arXiv preprint :2305.08322*, 2023.

N. Jain, K. Han, A. Gu, W. Li, F. Yan, T. Zhang, S. Wang, A. Solar-Lezama, K. Sen ve I. Stoica. Livecodebench: Kod için büyük dil modellerinin bütünsel ve kontaminasyonsuz değerlendirmesi. *CoRR*, abs/2403.07974, 2024. URL <https://doi.org/10.48550/arXiv.2403.07974>.

- S. Krishna, K. Krishna, A. Mohananeey, S. Schwarcz, A. Stambler, S. Upadhyay ve M. Faruqi. Gerçek, getir ve akıl yürüt: Geri getirme destekli üretimin birleşik bir değerlendirmesi. CoRR, abs/2409.12941, 2024. doi: 10.48550/ARXIV.2409.12941. URL <https://doi.org/10.48550/arXiv.2409.12941>.
- A. Kumar, V. Zhuang, R. Agarwal, Y. Su, J. D. Co-Reyes, A. Singh, K. Baumli, S. Iqbal, C. Bishop, R. Roelofs, . Takviyeli öğrenme yoluyla kendi kendini düzeltmek için dil modellerini eğitme. arXiv ön baskı arXiv:2409.12917, 2024.
- H. Li, Y. Zhang, F. Koto, Y. Yang, H. Zhao, Y. Gong, N. Duan ve T. Baldwin. CMMLU: Measuring massive multitask language understanding in Chinese. arXiv ön baskı arXiv:2306.09212, 2023.
- T. Li, W.-L. Chiang, E. Frick, L. Dunlap, T. Wu, B. Zhu, J. E. Gonzalez ve I. Stoica. Kitle kaynaklı verilerden yüksek kaliteli kıyaslamalara: Arena-hard ve benchbuilder pipeline. arXiv ön baskı arXiv:2406.11939, 2024.
- H. Lightman, V. Kosaraju, Y. Burda, H. Edwards, B. Baker, T. Lee, J. Leike, J. Schulman, I. Sutskever, ve K. Cobbe. Adım adım doğrulayalım. arXiv ön baskı arXiv:2305.20050, 2023.
- B. Y. Lin. ZeroEval: Dil Modellerini Değerlendirmek için Birleşik Bir Çerçeve, Temmuz 2024. URL <https://github.com/WildEval/ZeroEval>.
- MAA. American invitational mathematics examination - aime. American Invitational Mathematics Examination - AIME 2024 içinde, Şubat 2024. URL <https://maa.org/math-yarismalar/amerikan-istihdamsal-matematik-sinavi-aime>.
- OpenAI. Merhaba GPT-4o, 2024a. URL <https://openai.com/index/hello-gpt-4o/>.
- OpenAI. llms ile akıl yürütmeyi öğrenme, 2024b. URL <https://openai.com/index/learnin-g-to-reason-with-llms/>.
- OpenAI. SimpleQA ile tanışın, 2024c. URL <https://openai.com/index/introducing-simpleqa/>.
- OpenAI. SWE-bench ile tanışın doğrulanmış swe-bench'in insan tarafından doğrulanmış bir alt kümesini yayınlıyoruz, bu daha fazla, 2024d. URL <https://openai.com/index/introducing-swe-bench-dogrulandi/>.
- Qwen. Qwq: Bilinmeyen sınırları üzerine derinlemesine düşünün, 2024a. URL <https://qwenlm.github.io/blog/qwq-32b-preview/>.
- Qwen. Qwen2.5: Vakıf modellerinden oluşan bir parti, 2024b. URL <https://qwenlm.github.io/blog/qwen2.5>.
- D. Rein, B. L. Hou, A. C. Stickland, J. Petty, R. Y. Pang, J. Dirani, J. Michael ve S. R. Bowman. GPQA: Lisansüstü düzeyde bir google-proof soru-cevap kıyaslaması. arXiv ön baskı arXiv:2311.12022, 2023.
- Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, M. Zhang, Y. Li, Y. Wu ve D. Guo. Deepseekmath: Açık dil modellerinde matematiksel akıl yürütmenin sınırlarını zorlamak. arXiv ön baskı arXiv:2402.03300, 2024.
- D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, T. P. Lillicrap, K. Simonyan ve D. Hassabis. Genel bir pekiştirmeli öğrenme algoritması ile kendi kendine oynayarak satranç ve shogi'de ustalaşmak. CoRR, abs/1712.01815, 2017a. URL <http://arxiv.org/abs/1712.01815>.

- D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. P. Lillicrap, F. Hui, L. Sifre, G. van den Driessche, T. Graepel ve D. Hassabis. İnsan bilgisi olmadan go oyununda ustalaşmak. *Nat.*, 550(7676):354-359, 2017b. doi: 10.1038/NATURE24270. URL <https://doi.org/10.1038/nature24270>.
- C. Snell, J. Lee, K. Xu ve A. Kumar. Ilm test süresi hesaplamasını en iyi şekilde ölçeklendirmek, model parametrelerini ölçeklendirmekten daha etkili olabilir, 2024. URL <https://arxiv.org/abs/2408.03314>.
- T. Trinh, Y. Wu, Q. Le, H. He ve T. Luong. İnsan gösterileri olmadan olimpiyat geometrisi çözme. *Nature*, 2024. doi: 10.1038/s41586-023-06747-5.
- J. Uesato, N. Kushman, R. Kumar, F. Song, N. Siegel, L. Wang, A. Creswell, G. Irving ve I. Higgins. Süreç ve sonuç tabanlı geri bildirim ile matematik kelime problemlerini çözme. *arXiv ön baskı arXiv:2211.14275*, 2022.
- P. Wang, L. Li, Z. Shao, R. Xu, D. Dai, Y. Li, D. Chen, Y. Wu ve Z. Sui. Math-shepherd: A label-free step-by-step verifier for llms in mathematical reasoning. *arXiv preprint arXiv:2312.08935*, 2023.
- X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, ve D. Zhou. Öz tutarlılık, dil modellerinde düşünce zinciri muhakemesini geliştirir. *arXiv ön baskı arXiv:2203.11171*, 2022.
- Y. Wang, X. Ma, G. Zhang, Y. Ni, A. Chandra, S. Guo, W. Ren, A. Arulraj, X. He, Z. Jiang, T. Li, M. Ku, K. Wang, A. Zhuang, R. Fan, X. Yue ve W. Chen. Mmlu-pro: Daha sağlam ve zorlu bir çok görevli dil anlama ölçütü. *CoRR*, abs/2406.01574, 2024. URL <https://doi.org/10.48550/arXiv.2406.01574>.
- C. S. Xia, Y. Deng, S. Dunn ve L. Zhang. Ajansız: Ilm tabanlı yazılım mühendisliği araçlarının gizemini çözme. *arXiv ön baskı*, 2024.
- H. Xin, Z. Z. Ren, J. Song, Z. Shao, W. Zhao, H. Wang, B. Liu, L. Zhang, X. Lu, Q. Du, W. Gao, Q. Zhu, D. Yang, Z. Gou, Z. F. Wu, F. Luo ve C. Ruan. Deepseek-prover-v1.5: Takviyeli öğrenme ve monte-carlo ağaç arama için kanıt asistanı geri bildiriminden yararlanma, 2024. URL <https://arxiv.org/abs/2408.08152>.
- J. Zhou, T. Lu, S. Mishra, S. Brahma, S. Basu, Y. Luan, D. Zhou ve L. Hou. Büyük dil modelleri için yönerge takip değerlendirmesi. *arXiv ön baskı arXiv:2311.07911*, 2023.

Ekler

A. Katkılar ve Teşekkür

Çekirdek

Katılımcılar

Daya Guo
Dejian Yang
Haowei Zhang
Junxiao Song
Ruoyu Zhang
Runxin Xu
Qihao Zhu
Shirong Ma
Peiyi Wang
Xiao Bi
Xiaokang Zhang
Xingkai Yu
Yu Wu
Z.F. Wu
Zhibin Gou
Zhihong Shao
Zhuoshu Li
Ziyi Gao

Katkıda

Bulunanlar

Aixin Liu
Bing Xue
Bingxuan Wang
Bochao Wu
Bei Feng Chengda
Lu Chenggang
Zhao Chengqi
Deng Chong Ruan
Damai Dai
Deli Chen
Dongjie Ji
Erhang Li
Fangyun Lin
Fucong Dai Fuli
Luo* Guangbo
Hao Guanting
Chen Guowei
Li
H. Zhang
Hanwei Xu
Honghui Ding
Huazuo Gao
Hui Qu

Hui Li
Jianzhong Guo
Jiashi Li
Jingchang Chen
Jingyang Yuan
Jinhao Tu
Junjie Qiu
Junlong Li
J.L. Cai
Jiaqi Ni
Jian Liang
Jin Chen
Kai Dong
Kai Hu*
Kaichao You
Kaige Gao
Kang Guan
Kexin Huang
Kuai Yu Lean
Wang
Lecong Zhang
Liang Zhao
Litong Wang
Liyue Zhang
Lei Xu
Leyi Xia
Mingchuan Zhang
Minghua Zhang
Minghui Tang
Mingxu Zhou Meng
Li
Miaojun Wang
Mingming Li
Ning Tian
Panpan Huang
Peng Zhang
Qiancheng Wang
Qinyu Chen
Qiushi Du
Ruiqi Ge*
Ruisong Zhang
Ruizhe Pan
Runji Wang
R.J. Chen
R.L. Jin

Ruyi Chen
Shanghao Lu
Shangyan Zhou
Shanhuang Chen
Shengfeng Ye
Shiyu Wang
Shuiping Yu
Shunfeng Zhou
Shuting Pan
S.S. Li
Shuang Zhou
Shaoqing Wu
Shengfeng Ye
Tao Yun
Tian Pei
Tianyu Sun
T. Wang
Wangding Zeng
Wen Liu
Wenfeng Liang
Wenjun Gao
Wenqin Yu*
Wentao Zhang
W.L. Xiao
Wei An
Xiaodong Liu
Xiaohan Wang
Xiaokang Chen
Xiaotao Nie Xin
Cheng
Xin Liu
Xin Xie
Xingchao Liu
Xinyu Yang
Xinyuan Li
Xuecheng Su
Xuheng Lin
X.Q. Li
Xiangyue Jin
Xiaojin Shen
Xiaosha Chen
Xiaowen Sun
Xiaoxiang Wang
Xinnan Song
Xinyi Zhou
Xianzu Wang
Xinxia Shan
Y.K. Li
Y.Q. Wang

Y.X. Wei
Yang Zhang
Yanhong Xu
Yao Li
Yao Zhao
Yaofeng Sun
Yaohui Wang
Yi Yu
Yichao Zhang
Yifan Shi
Yiliang Xiong
Ying He
Yishi Piao
Yisong Wang
Yixuan Tan
Yiyang Ma*
Yiyuan Liu
Yongqiang Guo
Yuan Ou
Yuduan Wang
Yue Gong
Yuheng Zou
Yujia He
Yunfan Xiong
Yuxiang Luo
Yuxiang You
Yuxuan Liu
Yuyang Zhou
Y.X. Zhu
Yanping Huang
Yaohui Li
Yi Zheng
Yuchen Zhu
Yunxian Ma
Ying Tang
Yukun Zha
Yuting Yan
Z.Z. Ren
Zehui Ren
Zhangli Sha
Zhe Fu
Zhean Xu
Zhenda Xie
Zhengyan Zhang
Zhewen Hao
Zhicheng Ma
Zhigang Yan Zhiyu
Wu
Zihui Gu

Zijia Zhu
Zijun Liu*
Zilin Li
Ziwei Xie
Ziyang Song
Zizheng Pan

Zhen Huang
Zhipeng Xu
Zhongyu Zhang
Zhen Zhang

Her bir rol içinde yazarlar ilk isme göre alfabetik olarak listelenmiştir. ile işaretlenmiş isimler ekibimizden ayrılan kişileri ifade etmektedir.