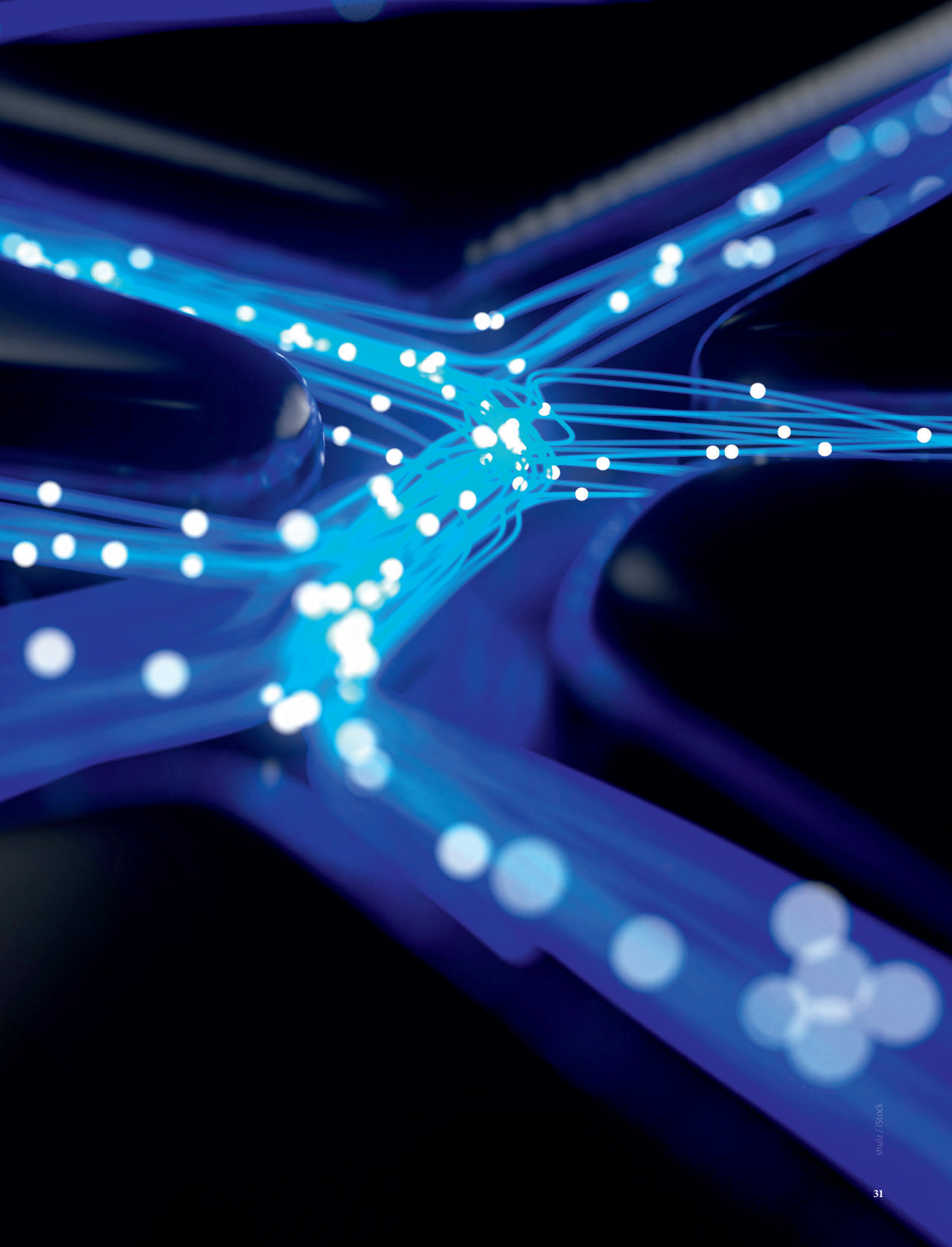
# Yapay Zekâda Paradigma Değişikliği: DeepSeek

Gürkan Caner Birer [ *Bilgisayar Mühendisi*

**Yapay zekâ teknolojileri alanında uzun zamandır süregelen ABD merkezli şirketlerin hâkimiyeti, Çinli teknoloji şirketi DeepSeek'in yeni dil modeli DeepSeek-R1 ile sekteye uğradı. Model, sektörün önde gelen aktörlerinden OpenAI'ın mantıksal modeli o1 ve Anthropic'in Claude modellerine denk bir performans sergileyerek teknoloji dünyasında büyük yankı uyandırdı.**



shulz / iStock

VCG Contributor / GettyImages

# Gelişmiş Düşünce Zinciri (Chain of Thought - CoT) Teknolojisi

Yapay zekâ modellerinin mantıksal düşünme yeteneklerinde önemli bir atılım olan Düşünce Zinciri (Chain of Thought – CoT) teknolojisi, 2024'ün son çeyreğinde OpenAI'ın o1 modeliyle birlikte önemli bir dönüm noktasına ulaştı. Bu teknoloji, yapay zekâ sistemlerinin karmaşık problemleri çözme sürecinde insan benzeri bir akıl yürütme yöntemi kullanmasını sağlıyor.

DeepSeek-R1'in özellikle matematik ve kodlama alanlarındaki üstün başarısı, akademik çevrelerin dikkatini çekti. Modelin en dikkat çekici özellikleri arasında düşünce süreçlerini şeffaf bir şekilde gösterebilmesi ve gelişmiş mantık yürütme yetenekleri yer alıyor. Bu özellikler, onu mevcut rakiplerinden belirgin şekilde ayırıyor.

Modelin piyasaya sürülmesi, sektörde ekonomik dengeler açısından da önemli etkiler yarattı. ABD merkezli teknoloji firmalarının hisse değerlerinde kayda değer düşüşler gözlemlendi. Özellikle dikkat çeken bir nokta, DeepSeek-R1'in rakiplerine kıyasla çok daha düşük maliyetle eğitilmiş olması. Bu durum, yapay zekâ altyapı yatırımlarının öngörülenden daha maliyet-etkin olabileceğini gösteriyor.

DeepSeek-R1'in sektörde yarattığı bir diğer önemli etki, benimsediği açık yaklaşımdan kaynaklanıyor. Modelin ticari kullanıma uygun lisansı, rekabetçi API fiyatlandırması ve geliştirme sürecine dair önemli teknik detayların açık kaynak olarak paylaşılması, küresel teknoloji topluluğunda büyük ilgi uyandırdı.

Bu gelişmeler, yapay zekâ teknolojilerindeki güç dengesinin yeniden şekillenebileceğine işaret ediyor. Gelin şimdi bu yenilikçi modelin teknik detaylarını ve özelliklerini daha yakından inceleyelim.

CoT teknolojisinin çalışma prensibi, özellikle çok katmanlı analiz gerektiren sorularda belirgin bir fark yaratıyor. Basit bir örnekle açıklamak gerekirse: "Türkiye'nin başkenti neresidir?" gibi doğrudan bilgi gerektiren sorular için standart bir yapay zekâ modeli yeterli olabilirken "Ankara'dan Adana'ya 60 kWh'lık tam dolu pile sahip aracımla ne kadar sürede gidebilirim?" gibi karmaşık hesaplamalar gerektiren sorular için çok daha gelişmiş bir analiz süreci gerekiyor. Bu tür sorularda model,

- İki şehir arasındaki mesafe
- Elektrikli aracın enerji tüketim değerleri
- Ortalama seyahat hızı
- Şarj istasyonlarının konumları
- Şarj süreleri

gibi çok sayıda değişkeni hesaba katarak kapsamlı bir analiz gerçekleştiriyor.

Teknolojinin uygulanmasında iki farklı yaklaşım dikkat çekiyor. OpenAI'nin o1 ve henüz yaygın kullanımda olmayan o3 modelleri, bir soruya yanıt üretirken izlediği çok katmanlı düşünme sürecinin özet bir versiyonunu kullanıcılarla paylaşmayı tercih ediyor. Bu strateji hem teknoloji güvenliğini sağlamayı hem de olası güvenlik açıklarını minimize etmeyi hedefliyor.Buna karşılık DeepSeek-R1, düşünce sürecinin tüm aşamalarını şeffaf bir şekilde kullanıcılarla paylaşıyor. Bu yaklaşım, kullanıcılara modelin mantık yürütme sürecini adım adım gözlemleme imkânı sunuyor. Şeffaflık odaklı bu strateji, özellikle iki önemli avantaj sağlıyor:

**1.** Kullanıcılar, modelin sonuca ulaşma sürecini detaylı olarak analiz edebiliyor.
**2.** Hatalı sonuçlarda kullanıcılar, modelin düşünce düşünce sürecinin herhangi bir adımındaki hatalı çıkarımları tespit ederek daha doğru yönlendirmeler yapabiliyor.

Özellikle dikkat çekici bir nokta, DeepSeek-R1'in bazen nihai cevabı hatalı olsa bile düşünce sürecinde ortaya koyduğu ara analizlerin kullanıcılara değerli içgörüler sağlayabilmesi. Bu özellik, modelin eğitim ve araştırma alanlarındaki potansiyel kullanımını önemli ölçüde artırıyor.

# Açık Kaynak Yaklaşımı ve Teknik Yenilikler

DeepSeek'in benimsediği şeffaflık odaklı yaklaşım, yapay zekâ alanında önemli bir paradigma değişiminin öncüsü olarak değerlendiriliyor. Şirketin paylaşım stratejisi, geleneksel anlamda tam bir açık kaynak yaklaşımı sergilemese de sektördeki diğer aktörlere kıyasla benzeri görülmemiş bir şeffaflık sunuyor. Bu yaklaşımın en somut göstergeleri:

- Model ağırlıklarının paylaşılması (açık ağırlık olarak adlandırılıyor)
- GitHub üzerinden erişilebilen kapsamlı kod tabanı
- Akademik makalelerle detaylandırılan eğitim süreçleri
- Esnek ticari kullanım lisansı
- Yeni model geliştirmede serbest kullanım imkânı

DeepSeek, bu stratejinin bir parçası olarak üç farklı model geliştirdi:

**1. DeepSeek-R1-Zero:**
2024'ün aralık ayında tanıtılan 671 milyar parametreli DeepSeek-V3 temel modelinin pekiştirmeli öğrenme (RL) ile geliştirilmiş versiyonu. Modelin ürettiği cevapların kalitesinin insanlar tarafından yapılan değerlendirmeye göre artırılmaması yani insan geri bildirimine dayalı ince ayar (SFT) aşaması içermemesi geri bildirimine

piranka / iStock

dayalı denetimli ince ayar (SFT) aşamasını içermemesi nedeniyle "soğuk çalıştırma" eğitimi olarak nitelendiriliyor.

**2. DeepSeek-R1:**
R1-Zero modelinin daha ileri bir versiyonu olan bu model, pekiştirmeli öğrenmeye ek olarak insan geri bildirimlerini içeren SFT aşamaları ve geliştirilmiş pekiştirmeli öğrenme süreçleri ile güçlendirildi.

**3. DeepSeek-R1-Distill:**
Bu model, önceki SFT verilerini kullanarak Qwen ve Llama modellerinin mantık yürütme kapasitelerini geliştirmeyi hedefliyor. Klasik damıtma yönteminden farklı olarak 671 milyar parametreli DeepSeek-R1'den elde edilen çıktılar, Llama (8B ve 70B versiyonları) ve Qwen (1.5B-30B versiyonları) gibi modellerin akıl yürütme yeteneklerini geliştirmek için kullanıldı.

# Damıtma Teknolojisi ve Sektörel Etkileri

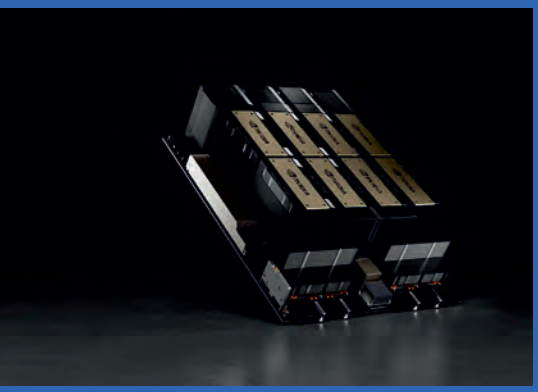
Damıtma teknolojisi, bir modelin bilgi ve yeteneklerini başka bir modele aktarmanın sistematik bir yöntemi olarak öne çıkıyor. Bu süreçte öğretmen modeline girdiler iletiliyor, elde edilen çıktılar kaydediliyor ve bu veriler öğrenci modelinin eğitiminde kullanılıyor.

GPT-4'ten GPT-4 Turbo'ya geçiş, bu teknolojinin başarılı örneklerinden biri. Damıtma işlemi, özellikle şirketlerin kendi geliştirdiği modeller arasında daha verimli sonuçlar veriyor. Bununla birlikte API veya sohbet arayüzleri üzerinden de -daha kısıtlı olsa da- uygulanabiliyor.

Bu teknoloji, sektörde önemli etik ve ekonomik tartışmaları da beraberinde getiriyor. Çeşitli modellerin hizmet şartlarıyla çelişmesine rağmen, damıtma sürecinin engellenmesi oldukça güç. IP yasakları ve hız sınırlamaları gibi teknik önlemler, tam bir koruma sağlamakta yetersiz kalıyor. Bu durum, özellikle büyük yatırımlarla geliştirilen öncü modeller için ciddi bir dezavantaj oluşturuyor. Zira geliştirilen teknolojiler kolaylıkla taklit edilebiliyor. Nitekim GPT-4 kalitesindeki modellerin sayısındaki artış, bu durumun somut bir göstergesi olarak değerlendiriliyor. Bir başka deyişle devasa bütçeler ayrılarak ortaya çıkarılan bir modeli öğretmen olarak kullanan diğer geliştiriciler, çok daha ucuza benzer kalitede modeller üretebiliyor. DeepSeek'in kendisi R1'i öğretmen olarak kullanıp açık ağırlıklı olarak sunulan Llama gibi modelleri çok daha "akıllı" hâle getirmiş. Elbette OpenAI gibi rakip firmalar DeepSeek'in ChatGPT ile eğitildiğinden de kuşkulanıyor ama bununla ilgili somut bir kanıt yok. DeepSeek, yayınladığı modellerin damıtma da dâhil başka modellerin üretilmesinde kullanılmasına izin veriyor.

DeepSeek'in sunduğu bir diğer önemli yenilik, tüm eğitim süreci boyunca 8 bitlik virgüllü sayıların (FP8) kullanılmasına izin veren, gelişmiş karma hassasiyetli eğitim çerçevesidir. Günümüzde çoğu yapay zekâ laboratuvarı, yapay nöronların hesaplamalarında

“tam hassasiyetli” 32 bitlik sayılar (FP32) kullanıyor. Bu yaklaşım, yüksek hassasiyet sağlarken önemli miktarda bellek tüketimine yol açıyor. DeepSeek’in FP8 teknolojisi ise bilimsel notasyon kullanarak sınırlı bit sayısıyla geniş sayı aralıklarını kapsama, hem mikroskobik hem de makroskobik değerleri etkin şekilde temsil etme ve geleneksel 8 bitlik sistemlerin 256 farklı değer sınırlamasını aşma özelliklerini başarıyla bir araya getiriyor. Normalde 32 bit ile virgüllü sayıları daha ayrıntılı tutulabilir. 8 bit ile küsuratın bir kısmından vazgeçilir bu da hassasiyetin düşmesi anlamına gelir. FP8 bellekten tasarruf ederken hassasiyetten sınırlı oranda fedakârlık eder ve yapay zekâ ile gerçekleştirilen birçok görev için yeterli doğruluk sağlar.



Özellikle yüksek kapasiteli işlemcilere erişimin sınırlı olduğu durumlarda bu tür verimlilik odaklı çözümler büyük önem taşıyor. Çünkü Nvidia H100 işlemcilerinin Çin’e satılmasına izin verilmiyor, bu nedenle DeepSeek daha düşük kapasiteli işlemcilerde çalışabilmek için birçok yenilik geliştirdi.

DeepSeek, sayıları aktivasyonlar için küçük karolara ve ağırlıklar için bloklara ayıran ve ağdaki kilit noktalarda stratejik olarak yüksek hassasiyetli hesaplamalar kullanan akıllı bir sistem geliştirerek bu sorunu çözdü. Diğer laboratuvarların benimsediği “eğit ve sıkıştır” yaklaşımı, süreç sonunda belirli bir kalite kaybına neden oluyor. Buna karşın DeepSeek’in yerel FP8 teknolojisi, doğrudan düşük bellek kullanımıyla eğitim yapıyor, performanstan ödün vermeden bellek optimizasyonu sağlıyor ve sıkıştırma sonrası kalite kaybı sorununu ortadan kaldırıyor. Binlerce GPU’da eğitim yaparken GPU başına bellek gereksinimlerindeki bu önemli azalma, genel olarak çok daha az GPU’ya ihtiyaç duyulması anlamına gelir.

Modern dil modellerine mantıksal düşünme yetenekleri kazandırma çabaları, yapay zekâ araştırmalarının en önemli odak noktalarından biri hâline geldi. ChatGPT 4o gibi gelişmiş modellere bu yeteneklerin kazandırılmasında kullanılan tekniklerin bir kısmı ticari sır olarak korunurken R1’in başarıyla uyguladığı ve akademik literatürde yer alan çeşitli yöntemler yapay zekâ teknolojilerinin gelişimine katkıda bulunuyor. Bu yaklaşımlardan en dikkat çekici olanı, modellerin düşünme süresini artırmaya yönelik etkileşim süresi optimizasyonu (inference-time scaling). İnsan bilişsel süreçlerine benzer şekilde yapay zekâ modellerine de daha fazla düşünme süresi tanındığında daha tutarlı ve mantıklı sonuçlar üretebiliyorlar. Kullanıcıların sorularına “adım adım düşün” veya “çözüm sürecini detaylandır” gibi yönlendirmeler eklemeleri, modellerin daha kapsamlı ve doğru yanıtlar üretmesini sağlıyor.

Ayrıca aynı soruyu tekrarlayarak en sık karşılaşılan yanıtı seçme stratejisi de yaygın olarak kullanılıyor.

Ancak DeepSeek-R1, bu geleneksel yaklaşımlardan farklı bir yol izleyerek pekiştirmeli öğrenme teknolojisini merkeze alan yenilikçi bir strateji geliştirdi. R1-Zero'nun geliştirilme sürecinde standart V3 modelinin pekiştirmeli öğrenme ile eğitilmesi temel alındı. Bu süreçte model, doğru yanıtlar ve mantık yürütme adımlarını gösterme başarısına göre ödüllendirildi veya cezalandırıldı. Bu basit ama etkili yaklaşım, modelin akıl yürütme yeteneklerini geliştirmede şaşırtıcı derecede başarılı oldu.

DeepSeek ekibi, R1-Zero'nun performansını daha da ileriye taşımak için kapsamlı bir veri seti oluşturdu. V3 ve R1-Zero modellerinin ürettiği 800.000 katmandan oluşan düşünce zinciri örneği kullanılarak geliştirilen R1, selefi R1-Zero'nun yeteneklerini önemli ölçüde aşmayı başardı.

Şirketin inovasyon yolculuğu burada durmadı. DeepSeek ekibi, R1'in kazandığı yetenekleri açık kaynaklı Llama ve Qwen 2.5 modellerine aktarmayı başardı. Damıtma yöntemiyle gerçekleştirilen bu bilgi transferi, daha küçük ölçekli modellerin de gelişmiş akıl yürütme yetenekleri kazanmasını sağladı. Her ne kadar bu modeller R1'in performans seviyesine ulaşamasa da R1-Zero'nun ötesine geçmeyi başardılar.

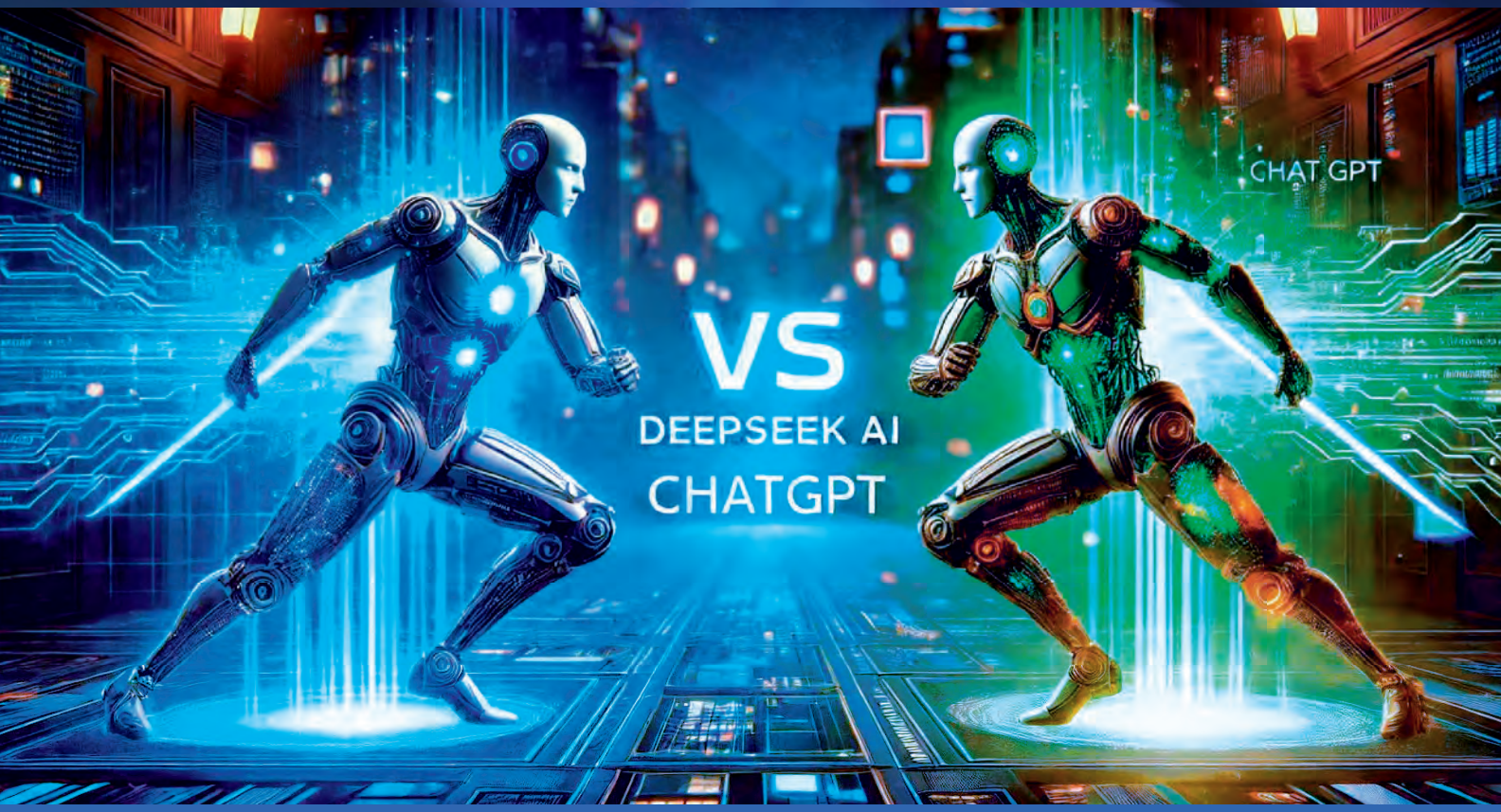
DeepSeek'in bu başarısı sektörde geniş yankı uyandırdı ve benzer yaklaşımları benimseyen diğer araştırma ekipleri de kayda değer sonuçlar elde etti. Özellikle Berkeley'deki California Üniversitesinden NovaSky ekibinin sadece 450 dolarlık bir yatırımla Qwen 2.5 modelini o1 seviyesine çıkarması, bu teknolojik yaklaşımın potansiyelini gözler önüne serdi. Geliştirilen bu model hem açık kaynak hem de yüksek performanslı.

## Uzman Karışımı Kullanımı

DeepSeek-V3, yapay zekâ dünyasında çığır açan uzman karışımı (MoE) yaklaşımıyla 671 milyar parametreli devasa bir model mimarisini yönetilebilir bir yapıya dönüştürmeyi başardı. Bu teknolojik atılım, yapay zekâ sistemlerinin verimliliğini ve erişilebilirliğini kökten değiştirme potansiyeli taşıyor.

Uzman karışımı yaklaşımı, tek bir büyük modelin içinde farklı uzmanlık alanlarına sahip altyapıların bulunması prensibine dayanıyor. Bu yaklaşım, ChatGPT'nin ilk sürümü olan GPT-3.5 gibi tüm parametreleri sürekli aktif tutan geleneksel modellerden önemli ölçüde ayrılıyor. GPT-4'ün benimsediği MoE mimarisi, her biri yaklaşık 110 milyar parametreye sahip 16 uzman yapıyı barındırarak bu alanda öncü bir adım atmıştı.

DeepSeek'in getirdiği yenilik, "kayıpsız yardımcı" yük dengeleme stratejisiyle performans düşüşünü engellerken verimli uzman kullanımını mümkün kılması. Sistem, gelen sorguları en uygun uzmanlara




yönlendirerek kaynakları optimize ediyor. Bu yapı, adeta farklı alanlarda uzmanlaşmış profesyonellerden oluşan bir danışma kurulu gibi çalışıyor. Teknik açıdan bakıldığında DeepSeek'in sistemi, 256 uzmandan oluşuyor ve her sorgu için bunların arasından sekizini seçiyor. Ayrıca tüm girdileri işleyen ve paylaşan bir uzman da sistemin önemli bir parçasını oluşturuyor. Bu mimarinin en önemli avantajı, modelin toplam parametre sayısı yüksek olmasına rağmen herhangi bir anda yalnızca küçük bir alt kümenin aktif olması. Böylece 671 milyar parametrelik devasa modelin yalnızca 37 milyar parametresi aynı anda işlem görüyor.

Bu verimlilik artışının pratik sonuçları oldukça etkileyici. Model, her biri 40 bin dolar değerindeki H100 GPU'lar yerine fiyatı bin dolar civarında olan iki Nvidia 4090 GPU ile çalışabiliyor. Bu maliyet optimizasyonu, yapay zekâ teknolojilerinin büyük şirketlerin teklinden çıkarak daha erişilebilir olması açısından kritik bir gelişme olarak değerlendiriliyor.

DeepSeek'in sistemi, ayrıca özelleşmiş ve genel amaçlı uzmanlar arasında dengeli bir dağılım sağlayarak eğitim sürecindeki verimliliği de artırıyor. Geleneksel MoE modellerinde eğitim sırasında karşılaşılan iletişim yükü problemi, DeepSeek'in yenilikçi yaklaşımıyla büyük ölçüde çözülmüş durumda. Bu başarıya uzman karışımı modellerinin en büyük sınırlamalarından biri olan bellek gereksinimi optimize edilerek ulaşıldı.

# Eğitim Stratejisi

DeepSeek'in geliştirdiği yapay zekâ modeli, etkileyici bir veri havuzunda gerçekleştirilen yoğun eğitim süreciyle dikkat çekiyor. Model, yaklaşık 15 trilyon belirteçli (25 trilyon karakter) devasa bir veri setiyle 2,79 milyon GPU saati boyunca eğitildi. Bu süre, ilk bakışta oldukça uzun görünse de sektördeki benzer çalışmalarla karşılaştırıldığında önemli bir verimlilik artışına işaret ediyor. Örneğin Meta'nın açık kaynak modellerinden Llama 3.1 405B'nin eğitimi için harcanan sürenin onda biri kadar bir zamana denk geliyor.

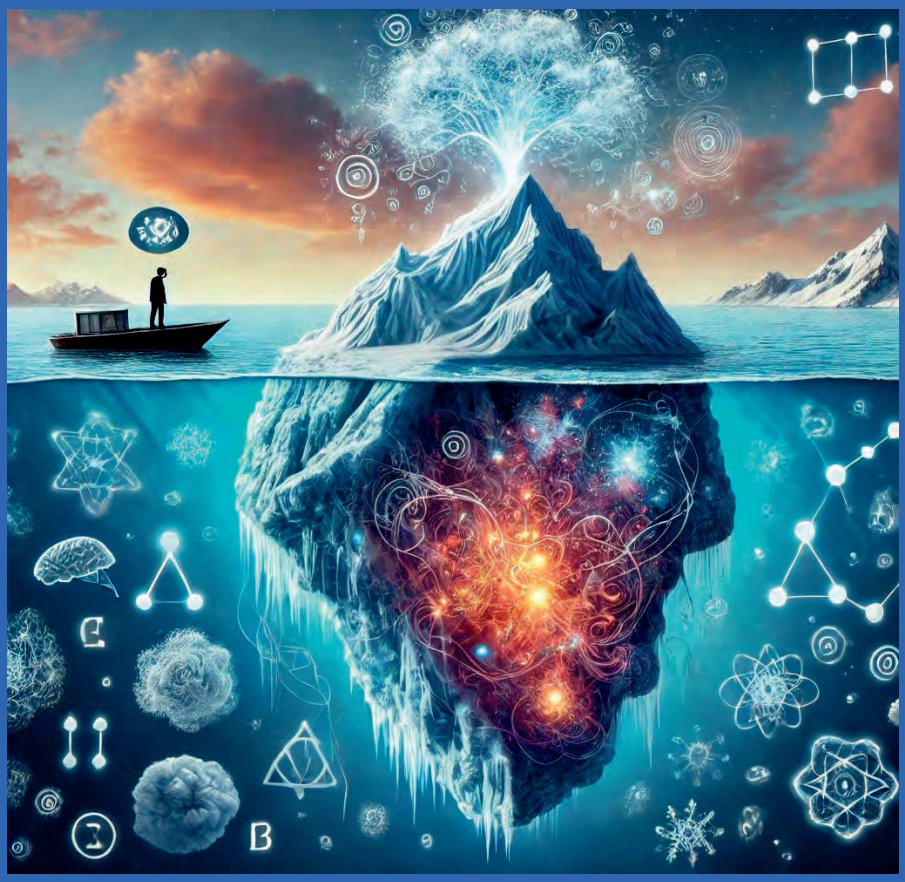
Modelin eğitim sürecinde kodlama ve matematik alanlarındaki verileri özel bir önem taşıyor. DeepSeek ekibi, göreceli grup politika iyileştirmesi (GRPO) olarak adlandırılan yenilikçi ve pekiştirmeli bir öğrenme yaklaşımıyla özellikle mantık yürütme yeteneklerinde çarpıcı gelişmeler elde etti.

Pekiştirmeli öğrenme teknolojisi, yapay zekâ sistemlerinin belirli hedeflere ulaşmasını teşvik eden ödül-ceza mekanizmasına dayanıyor. Bu yaklaşımda modeller, çevreleriyle etkileşime girerek aldıkları geri bildirimlere göre karar mekanizmalarını optimize ediyor. GRPO'nun getirdiği yenilik, her tekil etkileşim yerine benzer özellikteki etkileşim gruplarının ortalamasını baz alarak karar mekanizmalarını güncellemesi.

Geleneksel denetimli öğrenme yöntemlerinden farklı olarak pekiştirmeli öğrenme, modelin kendi başına optimal davranışları keşfetmesini sağlıyor. Bu teknoloji, özellikle oyun yapay zekâsı ve robotik kontrol sistemlerinde yaygın olarak kullanılıyor. Büyük dil modellerinde ise düşünce zinciri (CoT) süreçlerinin geliştirilmesinde kritik rol oynuyor.

DeepSeek'in eğitim stratejisinde dikkat çeken bir diğer nokta, ön eğitim sonrasında uygulanan çift aşamalı pekiştirmeli öğrenme yaklaşımı. İlk aşamada model, mantık yürüterek doğru cevapları kendiliğinden bulma yeteneği kazanıyor. Ancak bu süreçte karşılaşılan tutarsızlıkları ve dil karışıklıklarını gidermek için önceden hazırlanmış mantık yürütme metinleriyle desteklenen ikinci bir pekiştirmeli öğrenme aşaması uygulanıyor. Bu kapsamlı yaklaşım, R1 modelinin ortaya çıkmasıyla sonuçlandı.

DeepSeek R1'in bir başka önemli iyileştirmesi de metnin devamını tahmin etmesine dayalı sistemde oldu. Normalde dil modelleri, elindeki metnin devamına



eklenecek bir sonraki belirtecin ne olduğunu tahmin etmeye çalışıyor. Bu şekilde sözcükler oluşuyor. Ancak literatürdeki bilimsel çalışmalar, tek seferde bir sonraki belirteci değil de hem bir sonraki hem de ondan sonraki belirteci tahmin etmeye çalışarak yapılan model eğitiminin daha başarılı olabileceğini gösteriyor. DeepSeek ekibi, model eğitiminde bu yöntemi uyguladı. İlk belirteci tahmin etmek için kullanılan katmana ek olarak ikinci belirteci de tahmin etmeyi sağlayan başka katmanlar eklediler. Eğitim bu şekilde gerçekleştikten sonra model kullanıma sunulduğunda ilk belirteci tahmin eden katmanın dışında kullanılan diğer katmanları devre dışı bıraktılar. Buna rağmen ilk katman çok daha başarılı çalışmaya devam etti.

DeepSeek-V3'te uygulanan çok başlı gizli dikkat (Multi-head Latent Attention - MLA) teknolojisi de bellek optimizasyonu konusunda önemli bir atılım sağladı. Bu sistem, yapay zekâ modellerinin geleneksel olarak karşılaştığı gereksiz veri depolama sorununa yenilikçi bir çözüm getiriyor. MLA, verileri eğitim sürecinin başından itibaren akıllı bir sıkıştırma algoritmasıyla işleyerek yalnızca kritik bilgilerin saklanmasını sağlıyor. Bu yaklaşım hem eğitim maliyetlerini düşürüyor hem de modelin odaklanma kabiliyetini artırıyor. Bu sıkıştırma işlemi, modelin öğrenme sürecinin içine entegre edilmiş. Yani önce veriyi saklayıp sonra gereksiz kısımları atmak yerine en baştan sadece en önemli bilgileri saklamayı öğreniyor. Böylece model eğitimi için çok daha az işlemci (GPU) gerekiyor, bu da maliyeti düşürüyor ve süreci hızlandırıyor. Ayrıca model, gereksiz detaylara takılmak yerine gerçekten önemli bilgilere odaklanıyor.

DualPipe algoritmaları ve özel iletişim çekirdekleri, GPU kaynak kullanımında önemli bir optimizasyon sağladı. Sistem, GPU'ların akışlı çoklu işlemcilerinden (SM) yalnızca 20'sini iletişim için kullanırken geri kalan kaynakları, hesaplama işlemlerine ayırıyor. Bu akıllı kaynak yönetimi, geleneksel eğitim sistemlerine kıyasla çok daha yüksek GPU verimliliği sunuyor.

DeepSeek'in teknik makalelerde açıkladığı diğer optimizasyonlar arasında

- Tensör paralelliğinden kaçınan yeni yaklaşımlar
- Geri yayılım sürecinde seçici yeniden hesaplama stratejileri
- Ana model ve yardımcı tahmin modülleri arasında parametre paylaşımı

yer alıyor. Tüm bu teknolojik yeniliklerin birleşimi, eğitim sürecinde yaklaşık 45 kat verimlilik artışı sağladı.

## Maliyet Optimizasyonu

DeepSeek'in yapay zekâ alanındaki başarı hikâyesi, finans teknolojileri sektöründeki deneyimlerinden beslenen özgün bir inovasyon stratejisine dayanıyor. Firma, risk yönetimi, portföy optimizasyonu ve algoritmik ticaret alanlarındaki uzmanlığını, 2021'de hayata geçirdiği 10.000 GPU'luk devasa altyapı yatırımıyla yapay zekâ alanına taşıdı.

ABD'nin Çin'e yönelik teknoloji ihracat kısıtlamaları, DeepSeek'i zorunlu bir inovasyon sürecine yönlendirdi. Firma, donanım kısıtlamalarını aşmak için mevcut altyapısını maksimum verimlilikle kullanmaya odaklandı. Bu süreçte geliştirilen düşük seviyeli sistem optimizasyonları ve özgün iletişim mimarisi, çoklu ajan sistemlerinin performansında önemli iyileştirmeler sağladı.

DeepSeek'in maliyet yapısı, sektör ortalamalarıyla karşılaştırıldığında çarpıcı bir verimlilik tablosu ortaya koyuyor. Ana model V1'in eğitim maliyeti 5,6 milyon dolar seviyesinde gerçekleşti. Sektör deneyimleri, ön hazırlık ve optimizasyon süreçlerinin bu maliyeti 2-4 kat arturabileceğini gösteriyor. Buna rağmen toplam maliyet, OpenAI'ın benzer modeller için harcadığı 100-200 milyon dolarlık bütçenin çok altında kaldı.

DeepSeek, teknik bölümde açıkladığımız çoklu uzman kullanımı, ileri düzey bellek yönetimi, yenilikçi eğitim metodolojisi gibi yöntemler sayesinde rakiplerinin onda birinden daha az bir maliyetle rekabetçi bir ürün ortaya çıkardı.

Bu maliyet avantajı, hizmet fiyatlarına da doğrudan yansıdı. DeepSeek, 1 milyon belirteç işlemi için 0,55 dolar gibi rekabetçi bir fiyat sunabilirken sektör lideri o1 modelinin fiyatları, bunun 20-30 katı seviyesinde seyrediyor. Bu rekabetçi fiyatlandırma stratejisi, sektördeki diğer aktörleri de fiyat politikalarını gözden geçirmeye ve daha ekonomik alternatifler sunmaya yönlendirdi.

## Performans Analizi

DeepSeek-R1'in sektördeki önde gelen diğer yapay zekâ modelleriyle karşılaştırıldığı testler, teknoloji dünyasında dikkat çekici sonuçlar ortaya koydu. Model, OpenAI'nin güçlü yapay zekâ sistemi o1 ile gerçekleştirilen 11 farklı performans testinin 5'inde üstün başarı gösterdi.

Özellikle matematik ve yazılım alanlarındaki testlerde etkileyici bir performans sergileyen DeepSeek-R1; AIME 2024, MATH-500 ve SWE-Bench Verified gibi kritik değerlendirmelerde o1'i geride bıraktı. LiveCodeBench testlerinde elde edilen sonuçlar, bu başarıyı sayısal verilerle destekliyor: DeepSeek-R1 %65,9'luk kod çözüm oranıyla o1'in %63,4'lük performansının üzerine çıkmayı başardı.

Model ayrıca düşünce zinciri teknolojisini kullanmayan, sektörün önde gelen iki modeli Anthropic Claude 3.5 Sonnet ve OpenAI GPT-4o'yu geride bırakarak önemli bir teknolojik üstünlük gösterdi.

DeepSeek ekibinin açık kaynak modeller üzerindeki çalışmaları da dikkat çekici sonuçlar üretti. Qwen modelinin geliştirilmiş versiyonu olan DeepSeek-R1-Distill-Qwen-32B, AIME 2024 ve GPQA Diamond dahil tüm test platformlarında OpenAI-o1-mini'den daha iyi performans gösterdi. Benzer şekilde Meta'nın Llama-70B modelinin optimize edilmiş versiyonu DeepSeek-R1-Distill-Llama-70B, Codeforces dışındaki tüm karşılaştırma testlerinde o1-mini'yi geride bıraktı.

Teknik kapasite açısından değerlendirildiğinde DeepSeek-R1'in 128.000 belirteçli maksimum girdi kapasitesi, endüstri standartlarıyla uyumlu bir performans sunuyor. Google'ın 1 milyon belirteç işleyebilen modellerine göre düşük kalsa da tek seferde 300 sayfayı aşan metin işleme kapasitesine sahip olan modelin performans seviyesi pratik kullanımlar için yeterli.

Bu sonuçlar, DeepSeek'in geliştirdiği teknolojik inovasyonların sadece maliyet optimizasyonu değil, performans açısından da sektörde yeni bir rekabet düzeyi belirlediğini gösteriyor. Özellikle açık kaynak modellerin iyileştirilmesindeki başarı, yapay zekâ teknolojilerinin yaygınlaşması açısından umut verici bir gelişme olarak değerlendiriliyor.

## DeepSeek'in Sektörel Etkisi

DeepSeek'in yüksek performanslı yapay zekâ modellerini düşük maliyetle geliştirme ve ticari kısıtlamalardan uzak bir şekilde sunma başarısı, teknoloji sektöründeki güç dengelerini yeniden şekillendiriyor. Bu gelişmenin sektördeki büyük oyuncular üzerindeki etkileri, farklı iş modellerine sahip şirketler için çeşitli fırsatlar ve zorluklar yaratıyor.

Microsoft açısından bakıldığında yüksek performanslı modellerin düşük maliyetle erişilebilir hâle gelmesi, veri merkezi yatırımlarında önemli bir optimizasyon potansiyeli sunuyor. Şirketin GPU altyapısına yapacağı yatırımları azaltabilmesi, kullanıcılara sunulan hizmetlerin maliyet yapısını iyileştirmesine imkân sağlıyor.

Amazon Web Services (AWS) için durum daha da ilginç bir boyut kazanıyor. Kendi üst düzey modelini geliştirme konusunda yaşadığı zorluklar, yüksek kaliteli açık kaynak modellerin varlığıyla dengelenebilir. Bu durum, AWS'nin rekabetçi fiyatlarla kaliteli yapay zekâ hizmetleri sunmasına olanak tanıyor.

Apple, birleşik bellek mimarisi sayesinde bu dönüşümden en fazla fayda sağlayabilecek şirketlerden biri olarak öne çıkıyor. Apple Silicon çiplerin 192 GB'a kadar çıkabilen birleşik bellek kapasitesi, Nvidia'nın oyun GPU'larındaki 32 GB VRAM sınırının çok ötesinde bir potansiyel sunuyor. CPU, GPU ve NPU arasındaki entegre bellek paylaşımı, düşük bellek gereksinimli modellerin optimum performansla çalışmasına imkân tanıyor. Meta için bu gelişmeler, sosyal platformlarda yapay zekâ entegrasyonunun maliyetlerini düşürme potansiyeli taşıyor. Etkileşim maliyetlerindeki azalma, yapay zekâ destekli özelliklerin daha yaygın kullanımını mümkün kılıyor.

Google ise bu dönüşümden farklı şekilde etkilenebilir. Şirketin TPU çipleri üzerinden elde ettiği rekabet avantajı, donanım gereksinimlerinin düşmesiyle birlikte göreceli olarak azalabilir. Daha da önemlisi, düşük maliyetli yapay zekâ etkileşimlerinin yaygınlaşması, geleneksel arama hizmetlerine alternatif çözümlerin geliştirilmesini hızlandırabilir.

Bu teknolojik dönüşüm, sektördeki rekabet dinamiklerini derinden etkilerken özellikle mobil cihazlar ve kişisel bilgisayarlarda yapay zekâ modellerinin yerel olarak çalıştırılabilmesi, kullanıcı deneyiminde yeni bir dönemin kapılarını aralıyor. ■

**Kaynaklar**

https://arxiv.org/abs/2501.12948
https://novasky-ai.github.io/posts/sky-t1/