Büyük Dil Modellerinde

"Birden Beliren Beceriler" Tartışması

İlay Çelik Sezer | *TÜBİTAK Bilim ve Teknik Dergisi*

İstenen bir konuda tutarlı ve bütünlük arz eden metinler üretme, sorulan sorulara cevap verme, bir metni özetleme gibi görevleri şaşırtıcı bir beceriyle yerine getirebilen, üstelik bunları günlük konuşma dilinde verilen komutlar ya da sorulan sorularla yapabilen sohbet robotları hızla hayatımıza girdi. Zamanla kanıksanmış olsa da birçok kullanıcı için sohbet robotlarıyla ilk deneyim bilim kurgu unsurlarını andıracak kadar şaşırtıcıydı. Ancak sohbet robotlarının son yıllardaki hızlı gelişimi sadece genel kullanıcılar için değil bilim insanları için de şaşırtıcı özellikler gösteriyordu.

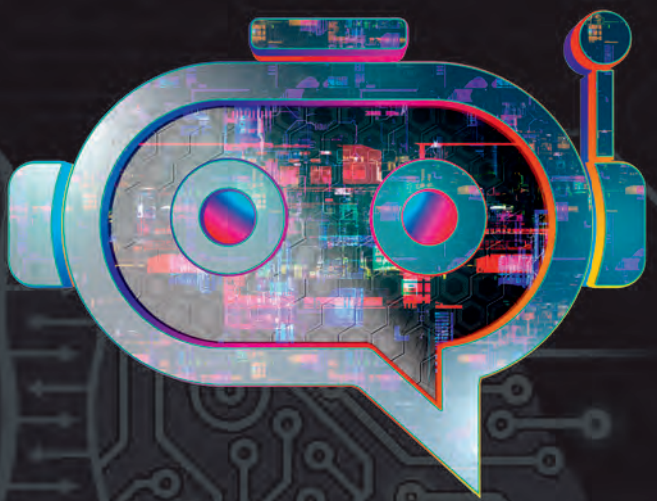
Kisimage / iStock

Kiyoshi Takahase Segundo / Alamy

Sohbet robotlarının temelini oluşturan "Büyük Dil Modelleri"nin (LLM) karmaşıklıkları artarken kimi becerilerin bir anda ortaya çıktığına ilişkin gözlemler, "birden beliren beceriler" (emergent abilities) kavramının doğmasına neden oldu. Yapay zekâ camiasında hayli popüler hale gelen bu kavram, büyük dil modellerinin gelişiminin öngörülemez bir yanı olduğu anlamına da geldiği için bunu kaygı verici bulan bir kesim bile oluştu. Stanford Üniversitesinden bir grup bilim insanının yeni bir araştırması ise çeşitli araştırmalarda büyük dil modellerine atfedilen birden beliren becerilerin en azından bir kısmının farklı kriterlerle değerlendirilmeleri durumunda aslında aşamalı olarak ve öngörülebilir şekilde geliştiğini gösteren sonuçlar ortaya koyarak konuya ilişkin önemli bir tartışma başlattı.

Yaklaşık yirmi yıldır yapılan yapay zekâ araştırmaları yapay sinir ağlarının, örneğin, bir resmin belirli bir nesneyi içerip içermediğini anlama, bir haber metnini özetleme ya da Türkçeden İngilizceye çeviri yapma gibi belirli görevleri yerine getirmek üzere eğitilmesine odaklanmıştı. Son yıllarda ise dil modelleri çevresinde yeni bir paradigma ortaya çıktı. Dil modelleri, bir cümlede önceki kelimelerin verilmesi durumunda bir sonraki kelimeyi tahmin etmeye yarayan yapay sinir ağlarıdır. Bugün ulaştıkları karmaşıklık düzeyinde artık büyük dil modelleri adını alan bu modeller, daha teknik bir anlatımla devasa veri tabanlarında bulunan kelimeler arasındaki istatistiksel ilişkileri haritalayarak bir komut cümlesinin devamında gelmesi en olası kelimeyi tahmin eden matematiksel modeller olarak tanımlanabilir.

Bu hedefe yönelik olarak çok büyük miktarlarda metinler kullanılarak eğitildiklerinde, büyük dil modellerine "sonraki kelimenin tahmini" biçiminde çerçevelendirilecek herhangi görevi yerine getirmeleri için komutlar verilebiliyor. Örneğin, Türkçe bir kelimeyi Almancaya çevirme görevi "sonraki kelimenin tahmini" çerçevesine şu şekilde dönüştürülebilir: "yapay zekânın Almancası…". Siz "Yapay zekânın Almancası…"yazdığınızda dil modeli bunu "Yapay zekânın Almancası nedir?" diye algılar ve cevap verir. Modelin, eğitim verilerindeki metinlerde Türkçe kelimelerle Almanca kelimeler arasındaki ilişkileri analiz etmesi ve bu ilişkilere dayanarak bir sonraki kelimeyi tahmin etmesi beklenir. Bu yeni paradigma tek bir görev için eğitilen göreve özgü modellerden pek çok görevi yerine getirebilen genelleyici modellere geçişi temsil ediyor. Genelleyici modellerin eğitim veri setinde özel olarak yer verilmeyen yeni görevleri bile yerine getirebildiği oluyor. Mesela az önceki örneğe dönersek, her ne kadar dil modeli yapay zekânın Almancasını söylemek üzere eğitilmese de büyük metin kümeleri üzerinde eğitilen dil modelleri bu soruyu doğru olarak yanıtlayabiliyor. Keza dil modellerinin bu iş için özel olarak eğitilmemiş olsalar bile iki basamaklı sayılarla çarpma işlemi yapabildiği GPT-3'te gösterilmişti. İşte birden beliren beceriler kavramı asıl büyük dil modellerinin bu tür başarılar göstermeye başlamasıyla birlikte ortaya çıktı.



da-kuk / iStock

Birden Beliren Beceriler İlgi Odağı

Dil modellerinin geçmişi 20-30 yıl öncesine uzanıyor. Yaklaşık altı yıl önce en güçlü dil modelleri devirli sinir ağı adı verilen bir yapıya dayanıyordu. Bunlar temelde, verilen bir sözcük dizisindeki bir sonraki kelimeyi tahmin ediyordu. Bir modeli devirli yapan

kendi çıktılarından öğrenebilmesi. Devirli bir modelin yaptığı tahminler, gelecekteki performansı iyileştirmek amacıyla sinir ağının eğitiminde kullanılıyor.

2017 yılında Google Brain'den araştırmacılar, dil modellerine transformer adı verilen yeni bir yapı kazandırdı. Devirli bir sinir ağı, bir cümleyi kelime kelime analiz ederken transformerler tüm kelimeleri aynı anda işlemliyor. Bu da transformerlerin büyük miktarlardaki metinleri paralel olarak işlemleyebilmesi anlamına geliyor. Transformerler başka faktörlerin yanı sıra modellerdeki parametre sayısı artırılarak dil modellerinin karmaşıklığının hızla artmasına olanak tanıyor.

Parametreler kelimeler arasındaki bağlantılar olarak düşünülebilir ve modeller eğitimleri sırasında metinlerin üzerinden geçerken bu bağlantılarda ayarlamalar yaparak kendilerini geliştirir. Başka faktörlere de bağlı olmakla birlikte genel olarak bir modelde ne kadar çok parametre varsa bağlantıları o kadar isabetli şekilde kurabiliyor ve insan dilini kabul edilebilir şekilde taklit etmeye o kadar yaklaşabiliyor. Ancak parametre sayısının artmasının bazı durumlarda performansı düşürebildiğini de belirtmek gerekir.

2020 yılında OpenAI araştırmacılarının yaptığı bir analiz, modellerin doğruluk ve beceriler açısından bir kaç yıl içinde ciddi bir gelişme kat ettiğini gösteriyordu. Ancak büyük dil modelleri, kullanıcılar tarafından kullanılmaya ve performansları araştırmacılar tarafından incelenmeye başladığında gerçekten beklenmedik şeyler de oldu. 175 milyar parametreye sahip GPT-3 ve parametre sayısı 540 milyara ulaşan PaLM gibi modellerin ortaya çıkmasıyla kullanıcılar giderek daha fazla birden beliren davranış bildirmeye başladı. Bu birden beliren davranışların pek çoğu büyük dil modellerinin ya önceden hiç karşılaşmadığı ya da nadiren

karşılaştığı problemleri çözmesi şeklindeydi. Büyük dil modellerinin, ilgili probleme özel bir eğitim veri seti olmaksızın bu tür problemleri çözebilmesi bir kısım araştırmacıya birden beliren davranışları daha yakından inceleme yönünde bir motivasyon verdi. Bunun ilk aşaması ise bu tür davranışların belirlenmesinden geçiyordu.

2020 yılında Google Research'te bilgisayar bilimci Ethan Dyer ve başka araştırmacılar büyük dil modellerinin dönüştürücü etkiler göstereceği öngörüsünde bulundu, ancak bu etkilerin ne olacağı ucu açık bir soru olarak duruyordu. Bu yüzden araştırma camiasına, büyük dil modellerinin yapabileceklerinin sınırlarını belirlemek için sınama amacıyla kullanılacak zor ve farklı görev örnekleri bulmaları için bir çağrı yaptılar. "Beyond the Imitation Game Benchmark", kısa adıyla BIG-bench adlı bu projede araştırmacılar özellikle büyük dil modellerinin bir anda, daha önce olmayan yeni beceriler kazandığı örneklerle ilgileniyordu. Proje kapsamında 450 araştırmacı büyük dil modellerinin kabiliyetlerini sınamak amacıyla tasarlanmış 204 görevden oluşan bir liste oluşturdu.

Tahmin edildiği gibi bazı görevlerde karmaşıklık arttıkça modelin performansı düzgün ve öngörülebilir şekilde gelişti. Bazı başka görevlerde parametre sayısının artması hiçbir gelişme sağlamadı. Ancak görevlerin yaklaşık %5'inde, modelin büyüklüğü (dolayısıyla karmaşıklığı) belirli bir eşiğin üzerine çıktığında model performansında hızlı ve çarpıcı sıçramalar görüldü. Bu eşik, göreve ve modele bağlı olarak çeşitlilik gösteriyordu. Bu tür davranışları bazı araştırmacılar fizikteki faz geçişlerine benzetirken, karmaşık sistemlerdeki genel birden beliren özellik kavramıyla ilişkilendirenler de oldu. Birden beliren özellikler fizik, biyoloji, ekonomi ve bilgisayar bilimi gibi çok çeşitli disiplinlerde karmaşık sistemlerde gözlemlenmiş ve uzun yıllar araştırmalara konu olmuş bir olgu. Nobel Fizik Ödülü sahibi Philip Anderson'ın 1972 tarihli "Fazla, Farklıdır" (More is Different) başlıklı tartışma yazısı bunun dikkat çekici bir örneği. Anderson bu yazıda bir sistemin karmaşıklığı arttıkça, sistemin mikroskobik ayrıntılarına ilişkin isabetli bir nicel anlayışa dayanılarak bile önceden kolayca ya da hiç öngörülemeyen yeni özelliklerin kendini gösterebildiğini savunur.

BIG-bench araştırmacıları 2022 Ağustos ayında yayımladıkları bir makalede birden beliren davranışların sadece şaşırtıcı ve öngörülemez olmadığını, aynı zamanda yapay zekânın güvenliği, potansiyeli ve riskleri etrafında dönen tartışmalar için de yönlendirici olması gerektiğini vurguluyordu. Örneğin, bir model giderek büyümesi durumunda tamamen öngörülemez ve kötü niyetli bir aktör olarak ortaya çıkabilir miydi? Aslında bu kaygı teknoloji endüstrisinde çok sayıda kişi tarafından paylaşılıyordu. Hatta bundan yaklaşık bir yıl önce teknoloji alanında lider 1000'den fazla profesyonelin imzaladığı bir açık mektupta birden beliren becerilere sahip giderek daha çok büyüyen, öngörülemez "kara-kutu" modeller geliştirmeye yönelik tehlikeli yarıştan geri adım atmak adına devasa yapay zekâ deneylerinin en az altı aylığına durdurulması isteniyordu.

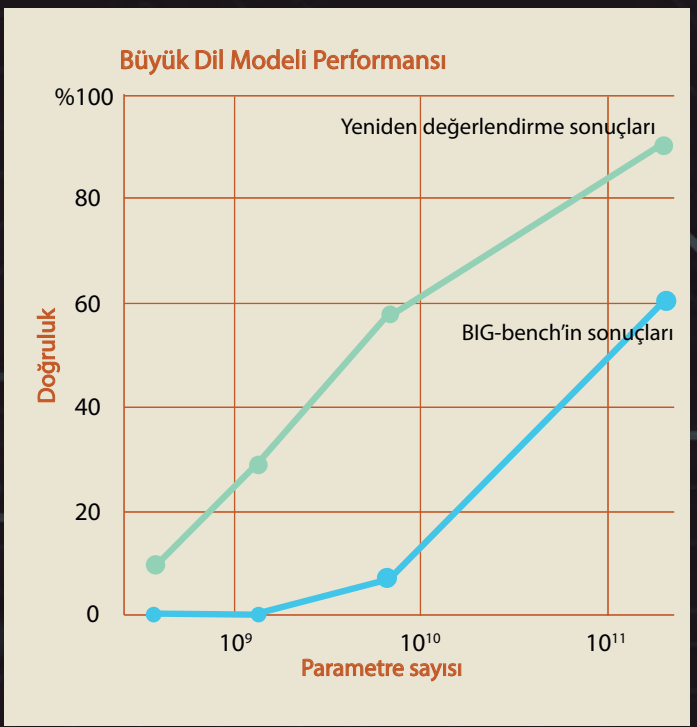

imaginima / iStock

Birden Beliren Davranışlar Yanılsama mı?

Birden beliren becerilere ilişkin ilgi ve araştırmalar süregitmekte iken Stanford Üniversitesinden bir grup araştırmacı büyük dil modellerinin birden beliren beceriler sergilediği yönündeki iddiaları mercek altına aldıkları bir çalışma yaptı. Sonuçlarını geçtiğimiz Aralık ayında New Orleans'ta düzenlenen Sinirsel Bilgi İşlemleme Sistemleri Yıllık Konferansı'nda (The NeurIPS 2023) sunan araştırmacılar birden beliren biçiminde nitelenen becerilerin bir anda ortaya çıkmış görünmesinin araştırmacıların büyük dil modellerinin performansını ölçme şeklinin bir sonucu olduğunu öne sürdü. Bu becerilerin aslında ne öngörülemez olduğunu ne de ani olarak ortaya çıktığını savunan araştırmacılardan konferans makalesinin başyazarı Sanmi Koyejo, modeller karmaşıklaştıkça bu becerilere geçişin insanların sandığından çok daha öngörülebilir olduğunu belirtiyor. Koyejo, birden beliren becerilere ilişkin güçlü iddiaların modellerin gerçekte ne yaptığına ek olarak bunları ölçme yollarına ilişkin seçimlerle de ilgili olduğunu belirtiyor.

Araştırmacılardan Rylan Schaeffer konferanstaki sunumunda, büyük dil modellerinin performansını değerlendirmede kullanılan ölçütlerin çok katı olduğunu ve kısmi puanlamaya imkân vermediğini belirtti. Araştırmacılar, BIG-Bench'in içerdiği test görevlerini incelediklerinde, belirlenen birden beliren becerilerin %90'dan fazlasının iki ölçüt altında incelendiğini tespit etti. Biri çoktan seçmeli notlandırma, diğeri birebir eşleşmeye dayalı bu iki ölçütün ise önemli kısıtları bulunuyor. Schaeffer bunu, "Tam doğru cevabı ya vermişsindir ya da vermemişsindir, ikisinin arası yok" şeklinde tarif ediyor. Bu ikili ölçüm yaklaşımıyla değerlendirme yapılması, bir problemin çözümünde, örneğin, beş basamaklı sayıları toplamada eğer bir model, büyüklüğü arttıkça aşamalı bir ilerleme kaydediyorsa bunun araştırmacılar tarafından fark edilemeyeceği anlamına geliyor. Bunun yerine belirli bir büyüklükte modelin ilerleyişi bir anda tavan yapmış görünüyor ve araştırmacılar bunun nasıl ortaya çıktığını merak etmeye başlıyor. Schaeffer, birden beliren becerilerin pek çok durumda dil modellerindeki temel değişimlerden kaynaklanmayıp araştırmacıların analizleri sonucunda ortaya çıkan birer yanılsama olabileceğini söylüyor.

Üç basamaklı sayıları toplama işlemi, söz konusu yanılsamanın bir örneği olarak verilebilir. BIG-bench çalışmasında araştırmacılar daha az sayıda parametreye sahip olduklarında hem GPT-3'ün hem de LAMDA adlı büyük dil modelinin toplama işlemlerini doğru yapmayı başaramadığını bildiriyordu. Ancak GPT-3, 13 milyar parametre kullanılarak eğitilmesi durumunda bu işlemleri bir anda yapabilmeye başlıyordu. Yine LAMDA, 68 milyar parametrede bunu başarıyordu. Bu bulgular toplama becerisinin belirli bir eşik değerinden sonra ortaya çıktığını düşündürüyordu. Ancak değerlendirme yöntemi sonuca "doğru ya da yanlış" etiketi verme biçimindeydi. Dolayısıyla bir büyük dil modeli



"BIG-bench" projesinde büyük dil modellerinin belirli bir büyüklüğe ulaştıklarında, daha küçük modellerin başaramadığı kimi görevleri bir anda ve öngörülemeyen bir şekilde yapmaya başladıkları öne sürülüyordu. Ancak bu görevlerdeki performans farklı ölçütlerle değerlendirildiğinde bu ani gelişim görünümü yok oluyor. Yukarıdaki grafik bu durumu temsili olarak örnekliyor. Koyejo ve ekibinin büyük dil modellerindeki birden beliren becerilerin aslında hem aşamalı olarak geliştiğini hem de öngörülebilir olduğunu öne sürdükleri makale, The NeurIPS 2023 Konferansı'nda üstün başarılı makale ödülü aldı.

basamaklardan çoğunu doğru olarak tahmin etse bile yanlış yapmış ya da başarısız sayılıyordu. Bu yaklaşım Koyejo ve ekibine pek de akla yatkın gelmedi. Çünkü örneğin, 100 ile 278'i toplamaya çalışıyorsanız "376", söz gelimi "-9" ya da "65"ten çok daha isabetli bir sonuçtur.

Sonuçta Koyejo ve ekibi, aynı görevi kısmi notlandırma yapılabilen bir ölçütle tekrar test etti. Bu yaklaşımla modelin, ilk basamağı, sonraki basamağı ve bir sonraki basamağı ne kadar iyi tahmin edebildiğini sorgulayabiliyorlardı. Koyejo yaptıkları araştırmanın dayandığı fikrin lisansüstü öğrencisi Rylan Schaeffer'den geldiğini, Schaeffer'in büyük dil modellerinin performansının, becerilerinin nasıl ölçüldüğüne göre değiştiğine ilişkin gözleminden yola çıktıklarını söylüyor. Bunun üzerine başka bir

lisansüstü öğrencisi olan Brando Miranda ile birlikte yeni ölçütler belirlediler. Bu ölçütler parametre sayısı arttıkça büyük dil modellerinin toplama problemlerine, basamakları giderek daha iyi tahmin ederek yanıt verdiğini gösterdi. Bu da toplama işlemi yapma becerisinin birden beliren özellikte değil aşamalı ve öngörülebilir olduğunu düşündürüyor. Farklı bir ölçüt kullanılınca birden belirme görünümü yok oluyor. Koyejo ve ekibi yaptıkları çalışmada bir adım daha öteye giderek, verili bir görüntü kümesine dayalı olarak yeniden görüntü oluşturan basit bir yapay sinir ağının doğruluğunu değerlendirmek için kullanılan ölçütleri değiştirerek bir becerinin "birden belirmiş" gibi görünmesini sağladı. Schaeffer kullandıkları modelin makine öğrenmesine giriş derslerinde kullanılanlar kadar basit olduğunu söylüyor. Bu derslerde öğrenciler genellikle, modelin ürettiği görüntünün orijinal görüntüyle benzerliğinin istatistiksel olarak karşılaştırılması sonucunda, modelin büyüdükçe görüntüleri kopyalamakta giderek ustalaştığını gözlemliyor. Ancak bu klasik ölçüm şekli değiştirilerek belirli bir kalite eşiğinin altındaki görüntü çıktılarına sıfır, üstündekilere bir puan verilmesi durumunda sanki model görevi hatasız şekilde gerçekleştirmeyi bir anda öğrenmiş gibi bir manzara oluşuyor. Schaeffer görüntüyle ilgili görevlerde birden beliren becerilerin bildirildiği bilinen hiçbir çalışma olmadığı için bunu kasıtlı olarak oluşturmanın oldukça yeni bir yaklaşım olduğunu belirtiyor.

Birden Beliren Beceriler Tartışılıyor...

Koyejo ve ekibinin araştırması, alandaki profesyoneller arasında hararetli bir tartışma başlattı. Bazı bilim insanları araştırmanın büyük dil modelleri bağlamındaki birden belirme kavramını tamamen yok etmediğini düşünüyor. Örneğin, Northeastern Üniversitesinden Tianshi Li'ye göre makalede, büyük dil modellerinde hangi ölçütlerin ne zaman ani gelişime işaret edeceğinin nasıl öngörüleceği açıklanmıyor.

JIRAROJ PRADITCHAROENKUL / Alamy

Li, bu anlamda bu becerilerin yine de öngörülemez olduğunu düşünüyor. BIG-bench'in 2022'deki makalesinin yazarlarından Jason Wei ise birden belirmeye ilişkin önceki raporları makul bulduğunu çünkü aritmetik gibi becerilerde zaten asıl meselenin doğru cevabı bulmak olduğunu belirtiyor.

Koyejo ve ekibinin araştırmasını, özellikle de iddiasının teknik açıdan basitliğinden dolayı etkileyici ve şaşırtıcı bulan ve bilimsel çalışmalardaki bu tür yaklaşım hatalarına karşı uyanık olma konusunda bir hatırlatıcı olarak niteleyen araştırmacılar da oldu. Bir derin öğrenme firmasında yönetici olan Ofer Shai, çalışmanın doğru ölçütler kullanmanın önemini vurguladığını belirtiyor.

Öte yandan Koyejo ve ekibi, makalelerinde yazılan hiçbir şeyin büyük dil modellerinin birden beliren beceriler gösteremeyeceği biçiminde yorumlanmaması gerektiğini, sadece daha önceki birden beliren beceri iddialarının büyük ihtimalle araştırmacıların analiz yöntemlerinden kaynaklanan yanılsamalar olduğu mesajını vermek istediklerini vurguluyor. Yapay zekâ modellerindeki becerilerin daha yetkin ve potansiyel olarak tehlikeli olması için birden beliren nitelikte sayılması gerekmediğini düşünen Schaeffer, bu yüzden de yapay zekânın nasıl geliştiğini izleyebilmek adına hassas değerlendirme araçları geliştirilmesinin zaruri olduğunu söylüyor. Birçok meslektaşı gibi Schaeffer da yapay zekâ araştırmalarının ne kadar büyük bir hızla ilerlediği konusunda endişeli çünkü bu durum zaman zaman bilimsel yöntemin temel dayanağı olan kontrol mekanizmalarının hiçe sayılmasına neden olabiliyor.

Schaeffer'e göre, büyük yapay zekâ modelleriyle ilgili sorun, bu modellere erişimin mümkün olmaması. Bu modeller özel şirketlerin kontrolünde olduğu için bunlara girdi bile veremediklerini, bağımsız araştırmacıların çoğu zaman veri setleri hazırlayıp bunları modelde koşturmaları için şirketlere göndermek zorunda kaldıklarını anlatıyor. Şirketler de daha sonra modelin çıktılarını araştırmacılara gönderiyor. Öte yandan Schaeffer bu şirketlerin ticari kaygılar yüzünden yapay zekânın kabiliyetlerini abartma, buna karşılık olası zararlı yan etkilerini küçümseme eğiliminde olduklarını; bu modellerin özel şirketlere ait olmasının ve bunlara ilişkin bilgilerin de onların kontrolleri altında bulunmasının bu alandaki bilimsel çalışmaları zorlaştırdığını da ekliyor.

Bugünkü büyük dil modellerindeki birden belirme iddiaları farklı ölçme araçlarıyla açıklanabilse bile bunun, yarının çok daha büyük ve karmaşık büyük dil modelleri için muhtemelen geçerli olmayacağını düşünenler de var. Rice Üniversitesinden bilgisayar bilimci Xia Ben Hu büyük dil modelleri büyüyerek bir sonraki evreye geçtiklerinde ister istemez başka görevlerden ve başka modellerden bilgi alacaklarını düşünüyor.

Öte yandan birden belirme olgusuna ilişkin tüm bu tartışmalar sadece araştırmacıları ilgilendiren soyut bir sorundan ibaret değil. Yapay zekâ alanında bir girişimci de olan araştırmacı Alex Tamkin bu olgunun, büyük dil modellerinin nasıl davranacağını öngörme çabalarıyla doğrudan ilişkili olduğunu belirtiyor. Tamkin bu tartışmanın tüm yapay zekâ camiasına, modellerin geleceğini öngörmeye yönelik araştırmaların da ne kadar önemli olduğunu hatırlatmasını umuyor ve şu soruyu soruyor: "Sonraki nesil modellere şaşırmamayı nasıl başarabiliriz?" ■

Kaynaklar

https://hai.stanford.edu/news/examining-emergent-abilities-large-language-models
https://hai.stanford.edu/news/ais-ostensible-emergent-abilities-are-mirage
https://www.quantamagazine.org/the-unpredictable-abilities-emerging-from-large-ai-models-20230316/
https://www.quantamagazine.org/how-quickly-do-large-language-models-learn-unexpected-skills-20240213/
https://www.marktechpost.com/2023/05/04/a-new-ai-research-from-stanford-presents-an-alternative-explanation-for-seemingly-sharp-and-unpredictable-emergent-abilities-of-large-language-models/
https://www.americanscientist.org/article/is-there-an-ai-metrics-mirage
https://www.forbes.com/sites/andreamorris/2023/05/09/ai-emergent-abilities-are-a-mirage-says-ai-researcher/?sh=540dfa3e283f