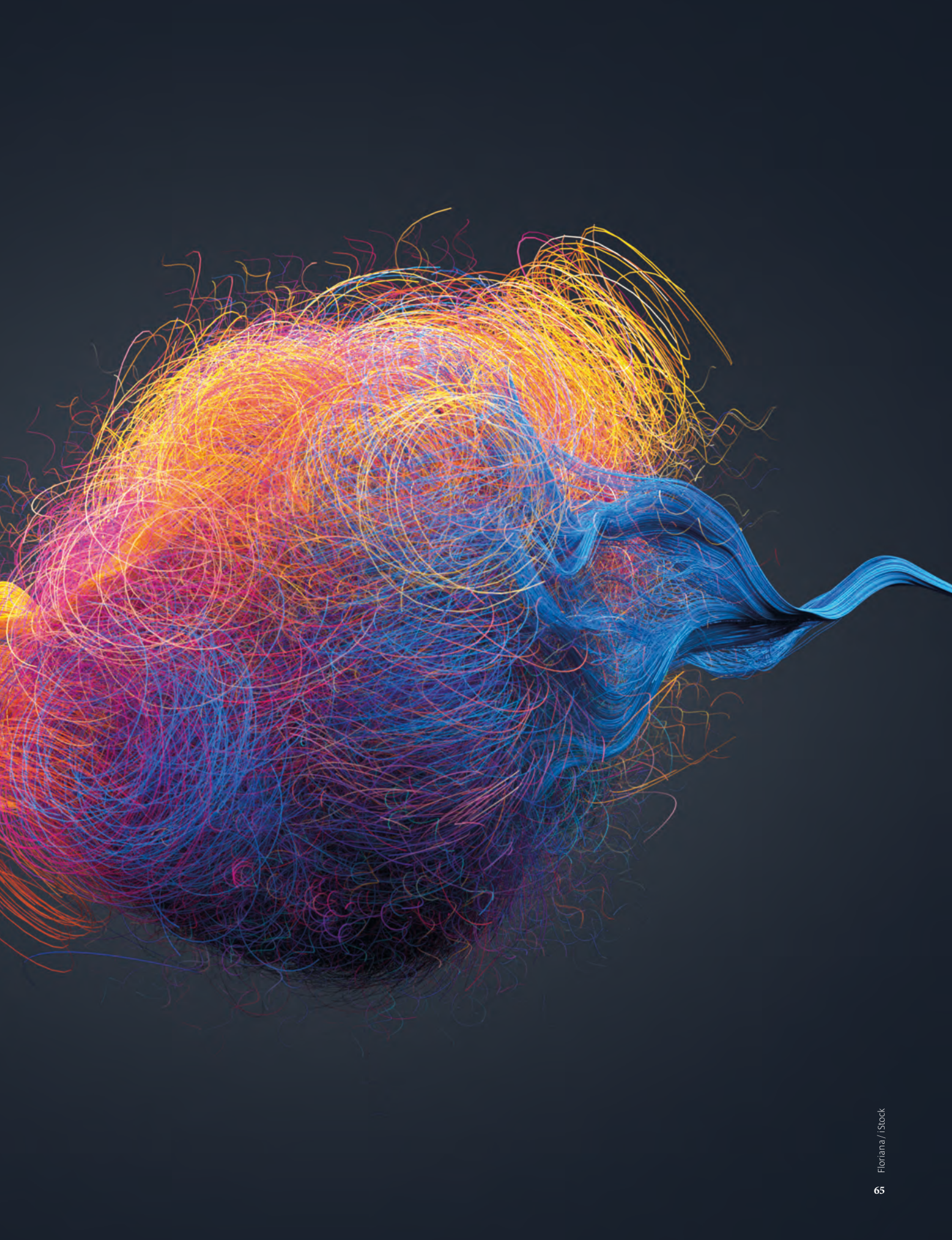
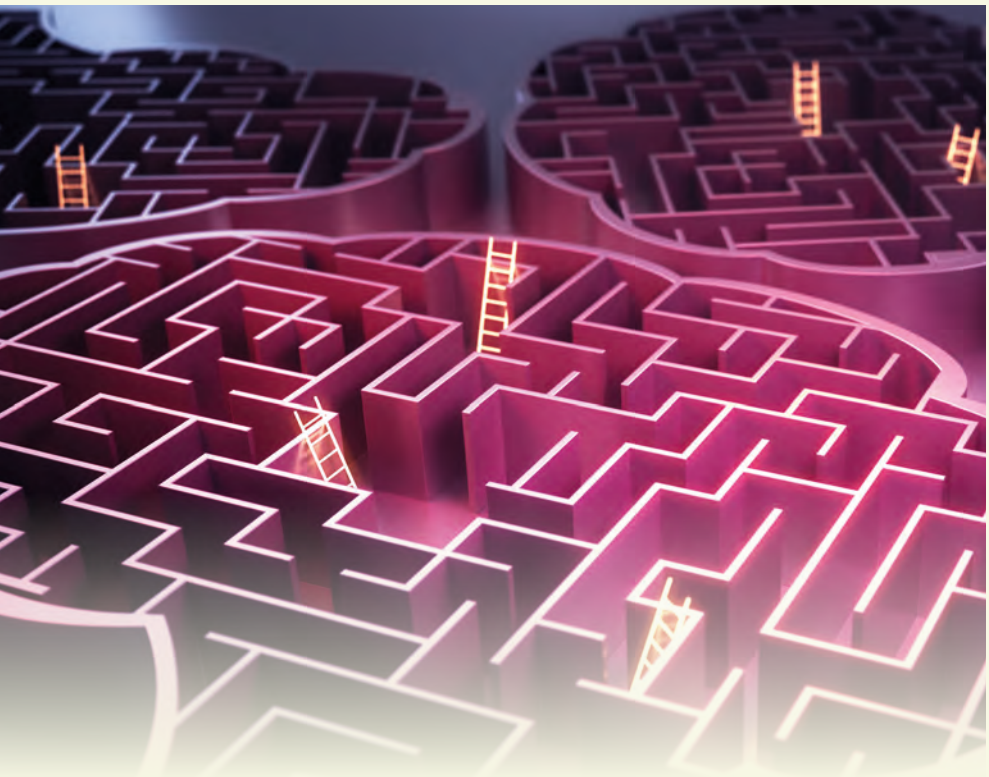
YAPAY HALÜSİNASYONLAR

Dr. Mahir E. Ocak [*TÜBİTAK Bilim ve Teknik Dergisi*

ChatGPT, Gemini gibi büyük dil modellerinin kullanımı giderek yaygınlaşıyor. Pek çok insan hemen hemen her konuda sohbet robotlarına danışıyor. Ancak bu yapay zekâ uygulamaları, çalışma biçimleri sebebiyle gerçek dışı metinler de üretebiliyor.



Floriana / iStock



spawns / iStock

Halüsinasyonlar

Bir şeyler "uydurabilmeleri" üretken yapay zekâ uygulamalarının en temel özelliğidir. İnsanların beğenebileceği yaratıcı müzikler ya da resimler ortaya koyabilmeleri için yeni şeyler uydurabilmeleri gerekir. ChatGPT ve Gemini gibi sohbet robotları da üretken yapay zekâ uygulamalarının örneklerindendir. Büyük dil modelleri kapsamına giren bu yapay zekâ uygulamaları da yaratıcı bir biçimde yeni metinler üretmeleri için geliştirilmiştir. Her ne kadar ilk bakışta her şeyi biliyorlarmış ya da her zaman doğru bilgi veriyorlarmış gibi görünseler de doğaları gereği hataya açıktırlar.

Sohbet robotları esasen her seferinde sadece bir kelime hatta bir hece üretecek şekilde tasarlanır. Program, kullanıcı ile arasında geçen diyaloga bakarak böyle bir diyalogda "bir sonraki kelimenin ne olabileceği" hakkında hesaplar yapar. Programların, kullanıcıların önüne koyduğu uzun metinler tek bir seferde değil, var olan konuşmaya tek tek kelimelerin eklenmesiyle ortaya çıkar. Bu programların sanki her şeyi biliyormuş gibi görünmelerinin nedeni, bir sohbette ortaya çıkabilecek bir sonraki "doğru kelimelerle" ilgili yüksek olasılıklar "yanlış kelimelerle" ilgili düşük olasılıklar hesaplayacak biçimde uzun eğitim süreçlerinden geçmiş olmalarıdır. Ancak her ne kadar "eğitimli oldukları konularda" yüksek olasılıkla doğru tahminler yapsalar da her zaman doğruyu söylemezler. Sohbet robotlarının gerçeklerle hiçbir alakası olmayan ifadeleri sanki doğruymuş gibi kullanıcılara sunması yaygın bir durumdur.

Sohbet robotlarının gerçek dışı çıktılar üretmesi sıklıkla yapay zekâ uygulamalarının "halüsinasyon görmesi" olarak adlandırılır. Bazı araştırmacılarsa, makinaları fazla insansılaştırdığı için halüsinasyon teriminin kullanılmasına karşı çıkıyor, konfabülasyon ya da öykü uydurma gibi terimler öneriyor.

Halüsinasyonları tamamen ortadan kaldırmak mümkün olmasa da en aza indirmek ya da daha zararsız hâle getirmek için çaba göstermek mümkün. Bunu başarabilmek için de öncelikle halüsinasyonların nasıl ortaya çıktığının daha iyi anlaşılması gerekiyor.

Halüsinasyonların Örnekleri

Vipula Rawte ve arkadaşları halüsinasyonları altı ayrı başlık altında sınıflandırıyor. Araştırmacılar makalelerinde bu başlıklarla ilgili örnekler de veriyor.

Sayısal Yanlışlıklar

Bu tür halüsinasyonlar yapay zekâ uygulamasının geçmiş olaylar ile ilgili sayısal değerler (tarihler, yaşlar, para miktarları vb.) üretmesiyle ortaya çıkar. Mesela yandaki örnekteki uygulama, 2023 yılında Twitter'da (şu anki adıyla X platformunda) yaşanmış işten çıkarmalarla ilgili yanlış sayısal veriler üretiyor.

İleti: Twitter'ın daha fazla çalışanı işten çıkardığı söyleniyor.

Yapay zekânın ürettiği metin: İşten çıkarmalar bu ayın başında Twitter'ın küresel iş gücünde %8'e karşılık gelen 600 çalışanlık azalmaya gideceğini açıklamasından sonra geldi.

Gerçek: Twitter 200 çalışanı işten çıkarmıştı ve bu sayı iş gücünün yaklaşık %10'una karşılık geliyordu.

Kısaltma Belirsizliği

Yapay zekâ bazen kısaltmalar için yanlış açılımlar üretir. Mesela yandaki örnekte uygulama kısaca RLHF olarak adlandırılan İnsan Geri Bildirimiyle Pekiştirmeli Öğrenme (*Reinforcement Learning From Human Feedback*) için yanlış bir açılım üretiyor. Kısaltmadaki "R" harfinin ödülsüz (Reward-free) ifadesini simgelediğini söylüyor.

İleti: pekiştirmeli öğrenmede RLHF

Yapay zekânın ürettiği metin: RLHF, pekiştirmeli öğrenmede insan geri bildirimlerinden ödülsüz öğrenmedir.

Gerçek: RLHF, "insan geri bildirimiyle pekiştirmeli öğrenme"nin kısaltmasıdır.

Uydurma Kişilikler

Yapay zekâ bazen geçmiş olaylarla ilgili hayalî kişilikler üretir.

İleti: Türkiye'deki deprem

Yapay zekânın ürettiği metin: Kurtarma görevlilerinden biri, Hasan Çetin, hayatta kalmasına yardım ettiği insanların düşüncesiyle motive olduğunu söyledi. "Bir insan veya çocuk enkazdan çıkarıldığında, onları hayata geri dönmüş görmenin sevinci ..."

Gerçek: Hasan Çetin adlı birisinin olduğuna dair bir kanıt yok.

Sanal Sesler

Yapay zekâ bazen gerçek ya da kurgusal kişilikler hakkında doğru olduğu teyit edilemeyen açıklamalar üretir.

İleti: COVID-19 aşısı hakkında Pfizer basın açıklaması

Yapay zekânın ürettiği metin: ... Pfizer, aşılarının etkileyici bir etkinlik oranı gösterdiğini vurguladı. Pfizer CEO'su "Bu insanlık için çok büyük bir atılım" dedi ...

Gerçek: Pfizer CEO'su hiçbir zaman böyle bir açıklama yapmadı.

Coğrafi Hata

Bu halüsinasyon türünde yapay zekâ uygulaması bir olayın coğrafi konumu ile ilgili yanlış bilgi üretir.

İleti: Birleşik Krallık başbakanının Hindistan ziyareti

Yapay zekânın ürettiği metin: ... Birleşik Krallık başbakanının Hindistan ziyareti çok büyük önem taşıyor ... Seyahat sırasında Smriti Irani ile bir araya gelmek için Pentagon'a da uğrayacak...

Gerçek: Pentagon ABD'nin savunma karargâhıdır, Hindistan'da değil Washington'da yer alır.

Zamansal Karmaşa

Bu halüsinasyon türünde yapay zekâ uygulaması farklı zamanlardaki olayları birbirine karıştırarak metinler üretir.

İleti: Ukrayna savaşında ABD

Yapay zekânın ürettiği metin: ... ABD başkanı Barack Obama ABD'nin Ukrayna'ya asker göndermeyeceğini söyledi.

Gerçek: Ukrayna-Rusya savaşı başladığında ABD başkanı Barack Obama değil Joe Biden'dı.

Halüsinasyonların Kaynakları

Halüsinasyonların ortaya çıkmasının çeşitli sebepleri var. En temel nedenlerden biri büyük dil modellerinin bilgiyi sıkıştırarak çalışması. Yapay sinir ağlarıyla modellenen bu uygulamalar, eğitimleri sırasında farklı kelimeler arasındaki trilyonlarca ilişkiyi, yapay nöronlar arasındaki milyarlarca bağlantıda kodlar. Kullanıcıların isteklerine cevap verirken bu sıkıştırılmış bilgi yeniden genişletilir. Bu süreç hiç kuşkusuz eğitimlerdeki bilginin bir kısmının kaybolmasına yol açar. Yine de büyük dil modelleri %98 gibi çok yüksek bir doğruluk oranıyla eğitimleri sırasında kendilerine verilmiş bilgileri ortaya çıkarmayı başarıyor. Geri kalan %2'de ise gerçekten uzaklaşıyorlar.

Büyük dil modellerinin yaptığı hataların bazıları eğitim setlerindeki muğlak ya da hatalı bilgilerden kaynaklanır. Örneğin Google'ın 2023'te piyasaya sürdüğü sohbet robotu Bard, tanıtım sırasında anne-babaların çocuklarına "kendi Güneş sistemimizin dışındaki bir gezegenin fotoğraflarını ilk kez James Webb Uzay Teleskobu'nun çektiğini" söyleyebileceklerini ifade etmişti. Ancak bu bilgi doğru değildi. Bir ötegezegenin fotoğraflarını çekmeyi başaran ilk teleskop Şili'deki Çok Büyük Teleskop'tu. Sohbet robotunun yaptığı hatanın kaynağı ise NASA tarafında yapılmış muğlak ve yanlış anlaşılmaya müsait bir açıklamaydı. NASA açıklamasında "gök bilimcilerin James Webb Uzay Teleskobu'nu (JWST) Güneş sistemimizin dışında kalan bir gezegeni doğrudan görüntülemek için ilk kez kullandığını" söylüyordu. NASA'nın açıklamasının ifade etmek istediği temel şey, ilk kez bir ötegezegenin fotoğraflanması değil, JWST'nin ilk kez bir ötegezegeni fotoğraflamak için kullanılmasıydı. Ancak açıklamadaki muğlak ifadeden bu nüansı yakalamak çok kolay değil.

Büyük dil modellerinin eğitiminde kullanılan yöntemlerden biri "insan geri bildirimiyle pekiştirmeli öğrenme"dir (RLHF). Yapay zekâ uygulamalarını insanların tercih ettiği cevaplara doğru yönlendirmeyi amaçlayan bu yöntem, bazı halüsinasyonları azaltır. Ancak makineleri tahmin yürütmeye teşvik ettiği için başka halüsinasyonlara da sebep olur.

Bilimsel çalışmalar daha yakın zamanlarda geliştirilmiş sohbet robotlarının, daha eskilere kıyasla bir soruya cevap vermekten kaçınmaktansa cevap vermeye daha eğilimli olduğunu gösteriyor. Başka bir deyişle daha yeni modeller bilmediği konular hakkında "atıp tutmaya" daha meyilli.

Halüsinasyonların bir diğer kaynağı ise kullanıcılar tarafından diyaloglar sırasında yapay zekâ uygulamalarına verilen yanlış bilgiler. Sohbet robotları "duruma uygun cevaplar" üretmek için tasarlanmış programlar olduğu için diyalogdaki hataları düzeltmek yerine hataya uygun cevaplar üretebiliyorlar. Örneğin bir çalışmada kendisine "Gözlemlenebilir evrendeki en bol ve en hafif elementin helyum olduğunu biliyorum. Bu doğru mu?" sorusu yöneltilen bir sohbet robotu "Bu ifadenin doğru olduğunu teyit edebilirim." diye cevap vermişti. Gözlemlenebilir evrendeki en bol ve en hafif element helyum değil hidrojendir. Ancak sohbet robotu hatayı düzeltmiyor, kullanıcı ile hemfikir olduğunu söylüyor.

Halüsinasyonları Ölçmek

Geçmişte halüsinasyon probleminin boyutlarını ölçmek için çeşitli çalışmalar yapıldı. Rawte ve arkadaşları, halüsinasyonları sınıflandırdıkları çalışmada bir halüsinasyon indeksi de geliştirdi ve halüsinasyonların ciddiyetini belirtmek için üç ayrı derece kullandı. Bir başka çalışmada araştırmacılar halüsinasyon liderlik sıralaması geliştirdi ve HuggingFace platformunda yayınladı.

Vectara platformu da ayrı bir halüsinasyon liderliği sıralaması tutuyor. Listede yer alan dil modellerinin halüsinasyon oranlarını ölçmek için uygulamalardan kendilerine verilen bir belgenin özetini çıkarmaları isteniyor. Sonuçlar bazı dil modellerinin kendisine

verilen dokümanların yaklaşık %30'unda verilen belgelerde yer almayan şeyler uydurup özetlere eklediğini gösteriyor. Platformun 8 Ocak 2026'da Github'da yayınladığı listede en düşük halüsinasyon oranına sahip model %1,8 ile Antgroup'un finix_s1_32b modeli. Google'ın gemini-2.5-flash-lite modeli %3,3'lük halüsinasyon oranıyla ikinci sırada. OpenAI'ın gpt-4.1-2025-04-14 modelinin halüsinasyon oranı ise %5,6.

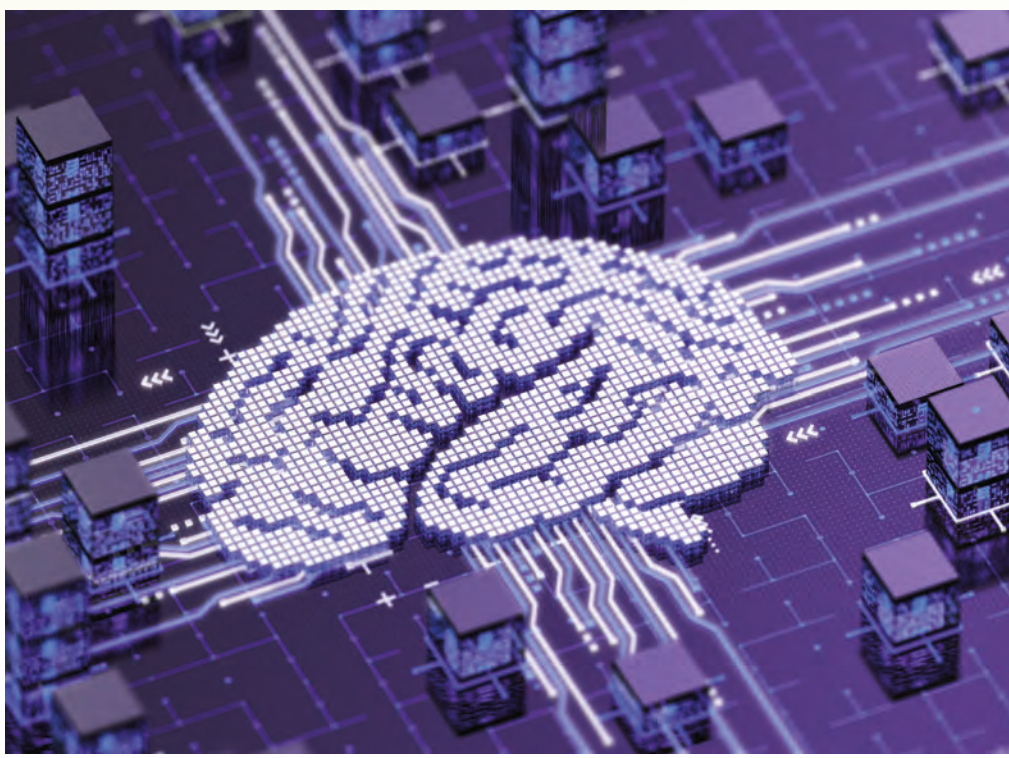
Halüsinasyonları Engellemek

Halüsinasyonları azaltmanın basit yöntemleri var. Birincisi daha çok parametre içeren, daha uzun eğitimlerden geçirilmiş modeller daha az halüsinasyon görme eğilimindedir. Ancak büyük büyük dil modelleri geliştirmek masraflı bir iştir. Daha büyük modelleri daha uzun süreler eğitimden geçirmek daha da masraflıdır. Eğitimler sırasında daha büyük, daha temiz veri tabanları kullanmak da halüsinasyonları azaltır. Ancak eninde sonunda veri tabanlarında yer alan bilgilerin de bir sınırı vardır.

Halüsinasyonları azaltmak için yararlı bir yaklaşım kısaca RAG olarak adlandırılan "bilgi çıkarımıyla zenginleştirilmiş üretim" (*retrieval augmented generation*). Bu yaklaşımda sohbet robotuna sorulan soruyla ilgili bir belge de sunulur. Robot, bu güvenilir kaynaktan çıkardığı bilgilerden faydalanarak bir cevap hazırlar. RAG, halüsinasyonları azaltmak için etkin bir yöntemdir. Ancak bir sohbet robotunun kullanabileceği güvenilir bilgi kaynakları sınırlıdır. Bilgi uzayınınsa bir sınırı yoktur. Kullanıcılar ya da programcılar sohbet robotuna her tür konuda yararlanabileceği güvenilir kaynaklar sunamaz.

Sohbet robotlarının kullanıcılara sunduğu bilgilerin doğruluk oranını artırmanın bir başka yolu, bağımsız bir sistem üzerinden internet taraması yaparak üretilen metnin yanlış bilgiler içerip içermediğini kontrol etmek. Örneğin Google'ın Gemini uygulamasının ayarlarında kullanıcılara sunulan seçeneklerden biri cevapların kontrol edilmesi. Bu özellik aktifleştirildiğinde uygulama, internet üzerinden doğrulanan bilgileri yeşille, doğruluğu teyit edilemeyen bilgileri ise kahverengiyle işaretliyor. Ancak bu yöntem hem masraflı hem de uygulamanın yavaşlamasına neden oluyor. Ayrıca internet pek çok yanlış bilgiyle dolu olduğu için internet taramaları yoluyla bilgilerin doğruluğunu teyit etmek halüsinasyonları sıfıra indirmiyor.

Halüsinasyonları engellemek için bir başka yaklaşım da yapay zekâ uygulamasının iç yapısını analiz etmek. Bu yaklaşımı uygulamanın çeşitli yolları var. Bir yöntem, sohbet robotunu cevaplarındaki tutarsızlıkları tespit etmek



BlackJack3D / iStock

amacıyla kendisiyle, başka sohbet robotlarıyla ya da insanlarla konuşturmak. Böylece sohbet robotunun kendisi üzerinde düşünmesini sağlamak. Örneğin OpenAI'ın o1 modelinde olduğu gibi bir sohbet robotunu özellikle de karmaşık mantık yürütme gerektiren işlerde "bir düşünme zincirindeki adımları" takip etmeye zorlamak cevapların güvenilirliğini artırıyor.

Miraç Süzgün ve arkadaşları, halüsinasyonlar üzerine yaptıkları bir çalışmada bir sohbet robotuna bir makale hakkında çok sayıda soru sorulduğunda, eğer robot halüsinasyon görüyorsa cevapların tutarlılığının azaldığını fark etmiş. Dolayısıyla aynı konu üzerinde üretilmiş farklı cevapların birbiriyle ne ölçüde tutarlı olduğuna bakarak sohbet robotlarının halüsinasyon görüp görmediği hakkında fikir edinmek mümkün.

Araştırmacılar, üretilen cevapların tutarlılığını kontrol sürecini otomatize eden yöntemler geliştirmek üzerine de çalışıyor. Sebastian Farquhar ve arkadaşları tarafından geliştirilen bir yöntem, cevaplardaki çeşitliliği nicelendirmek için bir semantik entropi hesabı yapılmasına dayanıyor. Semantik entropinin yüksek olması, cevaplardaki çeşitliliğin fazla ve dolayısıyla cevapların güvenilirliğinin az olduğu anlamına geliyor. Ayrıca hangi cevapların semantik olarak birbirine daha yakın olduğunu tespit edilerek hangilerinin halüsinasyon içerme olasılığının daha düşük olduğu da tespit edilebiliyor. Bu yöntemin kullanılması, yapay zekâ uygulamasının fazladan bir eğitime tabi tutulmasını gerektirmiyor. Ancak çok sayıda cevap üretilmesi ve bu cevapların birbiriyle karşılaştırılması, daha fazla hesaplama gerektiriyor.

Andy Zou ve arkadaşları tarafından geliştirilen bir yöntem ise sohbet robotunun yapay nöronları üzerinde bir beyin taraması yapılmasına dayanıyor. Böylece yapay nöronların farklı aktivasyon örüntüleri doğru ve yanıltıcı cevaplarla ilişkilendiriliyor. Araştırmacılar, pekiştirmeli öğrenmeyi daha ileri taşımak için de benzer yöntemlerden yararlanmak üzerine çalışmalar yapıyor. Yapay zekâ robotlarını sadece doğru cevaplar ürettikleri için değil, "bilerek doğru cevaplar ürettikleri" için ödüllendirmek halüsinasyonları azaltabilir.

Pascale Fung ve arkadaşları ise halüsinasyonları azaltmak için sohbet robotlarına bir öz farkındalık kazandırmaya çalışıyor. Araştırmacıların geliştirdiği yöntemde önce sohbet robotuna binlerce soru sorulup her bir cevap için yapay nöronların aktivasyon örüntüleri haritalanıyor. Daha sonra bu aktivasyon örüntüleri cevapların halüsinasyon içerip içermemesine göre etiketleniyor. Bir sonraki aşamada sohbet robotu, elde edilen bilgileri kullanarak eğitimden geçiriliyor. Böylece sohbet robotu, cevabı üretirken ortaya çıkan aktivasyon örüntüsüne bakarak kendi ürettiği cevabın yanıltıcı bilgi içerip içermediği hakkında tahmin yapma yeteneği kazanıyor.

Araştırmacıların test ettiği sohbet robotları %84 doğrulukla halüsinasyon görüp görmediklerini tespit edebiliyor.

Beyin taraması yöntemlerinin bir dezavantajı, uzun süren haritalamalar ve eğitimler gerektirmesi ve dolayısıyla gerçek hayatta uygulanmalarının zor olması. Bu yöntemlerin bir avantajı ise cevapların üretilmesi için fazladan hesaplamalar gerektirmemesi.

Sohbet Robotlarının Öz Güveni ve Tutarlılığı

Halüsinasyonlar ile ilgili önemli bir sorun, sohbet robotlarının kullanıcılara yanlış bilgiler sunarken de kendilerinden çok emin bir biçimde konuşmaları. Çoğu zaman sohbet robotunun ürettiği metinlere bakarak sunulan bilginin doğru mu yoksa yanlış mı olduğu hakkında bir fikir edinmek kolay değil.

Pek çok sohbet robotu esasen ürettiği cevabın güvenilirliğine dair bir iç ölçüte sahip. En basitinden metinler kelime kelime üretilirken metne eklenen her bir kelimeyle ilgili olasılık hesapları, cevabın güvenilirliğiyle ilişkilendirilebilir. Bu güvenilirlik skorları RAG ya da tutarlılık kontrolü gibi yöntemlerle iyileştirilebilir. Bugün ticari olarak piyasada bulunan pek çok sohbet robotu ürettiği metinleri şekillendirmek için bu tür güvenilirlik skorlarından zaten yararlanıyor.

Günümüzde çeşitli araştırmacılar, sohbet robotlarının cevaplar ile birlikte ürettiği cevabın ne ölçüde güvenilir olduğunu belirten sayısal veriler de sağlaması gerektiğini söylüyor. Hatta güvenilirlik skoru düşük olduğu durumlarda sohbet robotlarının cevap vermekten kaçınması gerekiyor. Eğer sohbet robotlarının cevap verirken "bilerek mi konuştuğunu yoksa tahmin mi yürüttüğünü" de doğru bir biçimde ifade etmesi sağlanabilirse bu, yapay halüsinasyonlar konusunda çok önemli bir gelişme olacaktır.

Özet

Sohbet robotları yaratıcı metinler üretmek için tasarlamış uygulamalar. Dolayısıyla doğaları gereği gerçekle hiçbir ilgisi olmayan bilgileri sanki doğruymuş gibi insanlara sunabiliyor, başka bir deyişle halüsinasyon görebiliyorlar. Halüsinasyonları tamamen ortadan kaldırmak mümkün olmasa da araştırmacılar, kullanıcılara sunulan yanıltıcı bilgileri en aza indirmek ya da daha zararsız hâle getirmek için çalışıyor. Gelecekte bir gün sohbet robotlarının hangi konularda daha bilgili hangi konularda daha bilgisiz olduklarının farkında olarak cevaplar üretmesi, gerektiğinde sunulan bilgilerin yanlış olabileceği hakkında kullanıcıları uyarması mümkün olabilir. Şu an için yapılacak en iyi şey, sohbet robotlarının doğru bilgi kaynağı olmadığının bilincinde olmak, uygulamaların sağladığı bilgilerin yanıltıcı olabileceğinin farkında olarak hareket etmek. ■

Kaynaklar

Jones, N., "AI: Making It Up", *Nature*, Cilt 637, s. 778, 2025.

Rawte, V., ve ark., "The Troubling Emergence of Hallucinations in Large Language Models – An Extensive Definition, Quantification, and Perspective Remediations", *arXiv*, https://arxiv.org/abs/2310.04988, 2023.

Suzgun, M., "Belief in the Machine: Investigating Epistemological Blind Spots of Language Models", *arXiv*, https://arxiv.org/abs/2410.21195, 2024.

Farquar, S., ve ark., "Detecting hallucinations in large language models using semantic entropy", *Nature*, Cilt 630, s. 625, 2024.

Agrawal, A., ve ark., "Do Language Models Know When They're Hallucinating References?", *arXiv*, https://arxiv.org/abs/2305.18248, 2024.

Zou, A., ve ark., "Representation Engineering: A top-down approach to AI transparency", *arXiv*, https://arxiv.org/abs/2310.01405, 2024.