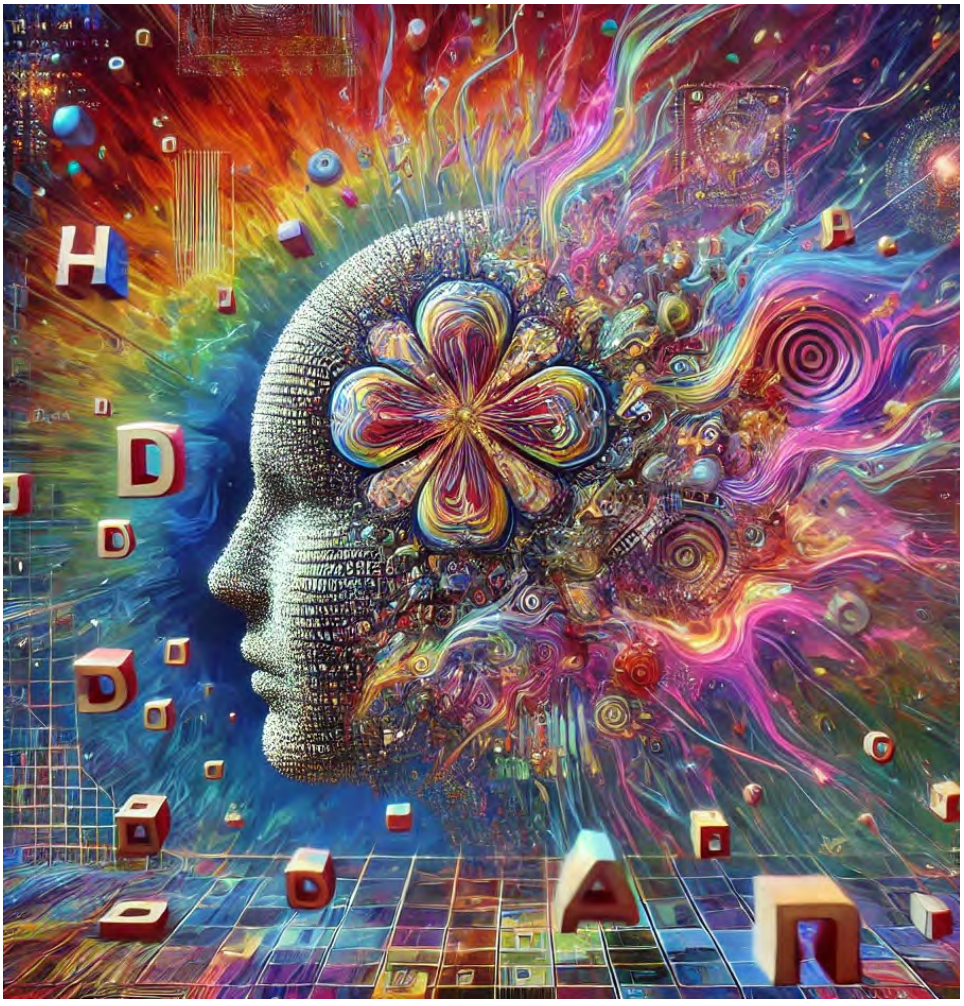
Tekno-Yaşam

Gürkan Caner Birer [*teknoyasam@tubitak.gov.tr*

Yapay Zekâ Ne Kadar “Uyduruyor”?



ChatGPT gibi dil modellerinin sundukları bilgilerin zaman zaman uydurma yani gerçek olmayan bilgilerle dolu olduğunu görmüşsünüzdür. Teknik olarak buna “halüsinasyon” deniyor. Son derece profesyonel dille sunulan bir metnin içerisinde yazan bazı bilgilerin uydurma olması önemli bir problem. Okuyucular bunun farkına varmayarak gerçek zannedebiliyor. Bu durumu çözmek için dil modeli geliştiricileri ciddi bir çaba sarf ediyor. Galileo adlı bir araştırma şirketi çeşitli uzunluktaki metinleri kullanarak 22 farklı dil modelinin ne sıklıkla halüsinasyon gördüğünü ölçtü.

Anthropic’in Claude 3.5 Sonne’si, yaklaşık 3.000 karakterden oluşan kısa metinlerde %97, 3.000-15.000 karakterden oluşan orta ve 25.000-60.000 karakterden oluşan uzun metinlerde %100 doğrulukla ilk sırada yer aldı. Açık kaynak modeller içinde Qwen2-72b Instruct, kısa metinlerde %95 ve orta uzunluktaki metinlerde %100 doğrulukla en yüksek puanı aldı.

Çoğu model orta uzunluklardaki metinlerde en iyi performansı gösterdi. Anlaşılan, ChatGPT gibi dil modellerine elimizdeki bir metinle ilgili sorular sorarken orta uzunlukta metinler sunmak en doğru sonucu almamızı sağlıyor. Geçen seneki testlerde en başarılı modelin %73 doğrulukta olduğu dikkate alındığında dil modellerinin ciddi gelişme kaydettiği görülüyor. Yine de araya serpiştirilmiş ufak tefek yanlış bilgilerin son derece tehlikeli olduğunu unutmamak gerekiyor.

Kaynak: rungalileo.io/hallucinationindex