Veri Duvarı Krizi

Yapay zekâ teknolojisinin hızlı gelişimi, beklenmedik bir sorunla karşı karşıya: Veri kıtlığı. MIT öncülüğündeki Data Provenance Initiative'in yeni araştırması, yapay zekâ modellerini eğitmek için kullanılan içeriklerde önemli bir düşüş olduğunu ortaya koydu.

Araştırmacılar, yaygın olarak kullanılan üç yapay zekâ eğitim veri setinde yer alan 14.000 web alanını inceledi. Sonuçlar, yayıncıların ve çevrimiçi platformların verilerinin yapay zekâ firmaları tarafından izinsiz toplanmasını önlemek için adımlar attığını gösteriyor. En kaliteli kaynaklardan gelen verilerin %25'i artık kısıtlanmış durumda. Bu kısıtlamalar genellikle "Robots Exclusion Protocol" adı verilen, web sitesi sahiplerinin otomatik botların sayfalarını taramasını engellemek için kullandıkları eski bir yöntemle gerçekleştiriliyor. Ayrıca, bazı veri setlerinde kullanım şartları nedeniyle verilerin %45'e varan oranlarda kısıtlandığı görülüyor. Bu durum, yapay zekâ geliştiricileri için ciddi bir sorun teşkil ediyor. ChatGPT, Google'ın Gemini'si ve Anthropic'in Claude'u gibi popüler yapay zekâ araçları, milyarlarca metin, görüntü ve video örneğiyle besleniyor. Daha fazla kaliteli veri, genellikle daha iyi çıktılar anlamına geliyor. Geçmişte veri toplamak nispeten kolaydı. Ancak son yıllardaki yapay zekâ patlaması, veri sahipleriyle gerilimlere yol açtı. Bazı yayıncılar ücretli dijital bariyerler kurdu veya verilerinin yapay zekâ eğitiminde kullanılmasını sınırlamak için kullanım şartlarını değiştirdi. Reddit ve StackOverflow gibi siteler, yapay zekâ şirketlerinden veri erişimi için ücret almaya başladı.

Bu kısıtlamalar, özellikle küçük yapay zekâ şirketleri ve akademik araştırmacılar için sorun yaratabilir. Büyük teknoloji şirketleri zaten kaliteli verilerin kullanım hakkını almışken sonradan gelen veya bağımsız aktörler aşılması zor bir "veri duvarı" ile karşılaşıyor. Yani herkesin işlemesine açık olarak sunulan internetteki tüm eğitim verilerinin tükendiği ve geri kalanının ücretli duvarlar arkasında saklandığı veya özel anlaşmalarla kilitlendiği bir durumla karşı karşıyayız. Bu durumun yapay zekâ gelişimini nasıl etkileyeceği henüz belirsiz. Ancak kesin olan bir şey var: Veri artık yapay zekâ dünyasında en değerli emtia haline geldi ve bu kaynağın kontrolü için mücadele yeni başlıyor.

Kaynak: https://nyti.ms/3zUlqcs