LLM’lerin Sınırları, Zorlukları ve Etik Boyutu

Eğitmen
Kubilay Tuna

Ders İçeriği

1. LLM’lerin Yetenekleri ve Sınırları
2. LLM’lerin Uygulama Alanları
3. Güvenlik ve Etik Kaygıları
4. Responsible AI: Etik ve Güvenilir Yapay Zeka
5. Güncel Çözümler ve İyileştirme Yaklaşımları

LLM'lerin Yetenekleri ve Sınırları

Yapabilecekleri

- **Metin Oluşturma:** Kullanıcının verdiği bir konuda makale, hikaye ya da şiir yazabilir.
- **Soru Yanıtlama:** Verilen bir bilgi kümesine dayanarak sorulara yanıt verebilir.
- **Özetleme:** Uzun metinleri daha kısa ve öz bir şekilde özetleyebilir.
- **Çeviri:** Bir dilden diğerine metin çevirisi yapabilir.
- **Çok aşamalı muhakeme (Multi-hop reasoning):** Birden fazla bilgi kaynağını ilişkilendirerek mantıksal çıkarımlarda bulunabilirler.

Yapamayacakları

- **Gerçek Zamanlı Bilgiye Erişim:** Eğitim verileri sabit olduğundan, güncel olaylara dair bilgi sunamaz.
- **Duygusal Anlayış ve Empati:** Metin üzerinden duyguları anlama kapasitesi sınırlıdır.
- **Kapsamlı Bilgi Analizi:** Derinlemesine analizler yapma yeteneği yoktur, çünkü bağlam ve ön bilgi eksikliği bulunabilir.
- **Model yanlılıkları (Bias):** Eğitildikleri veri setlerinde mevcut olan yanlılıkları öğrenirler ve bu, çıktılara yansıyabilir.

LLM’lerin Sınırları: Üretken Yapay Zekanın Yetersizlikleri

Veri Bağımlılığı ve Genelleme Problemi

Modeller, eğitim sırasında kullanılan verilere bağımlıdır. Yeterince çeşitlilik içermeyen veya taraflı verilerle eğitilen modeller, yeni ve beklenmedik durumlarla karşılaştıklarında genelleme yapamazlar.

Modellerdeki Taraflılık (Bias)

Büyük dil modelleri ve üretken yapay zeka modelleri, eğitim verisinde mevcut olan taraflılıkları öğrenir. Irk, cinsiyet, kültürel önyargılar gibi sosyal yanlılıklar, modellerin çıktısına doğrudan yansıyabilir.

Kapasite Sınırları

Modellerin parametre sayısı ve kapasitesi ne kadar büyük olursa olsun, gerçek dünyadaki her durumu kapsayacak şekilde genelleme yapamaz. Örneğin, 175 milyar parametreye sahip bir model olan GPT-3 bile, belirli görevlerde yanıt verme konusunda sınırlamalar gösterebilir.

Gerçek Zamanlı Performans Zorlukları

Modellerin gerçek zamanlı veya etkileşimli kullanımında, gecikmeler, hesaplama süresi ve enerji tüketimi gibi faktörler ciddi bir performans sorunu olabilir.

Güvenilirlik ve Hatalı Çıktılar

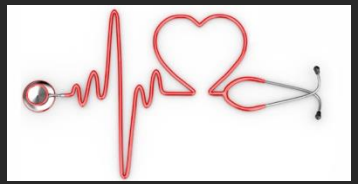
Üretken yapay zeka modelleri bazen yanlış veya eksik bilgi üretebilir. Özellikle dil modelleri, tutarlı görünebilecek şekilde yanlış bilgiyi güvenle sunabilir.

LLM'lerin Uygulama Alanları

Eğitim: Kişiselleştirilmiş öğrenme deneyimi: Öğrencinin seviyesine göre eğitim içeriklerini ayarlayabilir.
Otomatik test hazırlama ve değerlendirme: Sınavlar ve ödevler için içerik üretimi ve değerlendirme.

Pazarlama: İçerik oluşturma: Blog yazıları, sosyal medya paylaşımları, reklam metinleri gibi pazarlama içeriklerinin üretilmesi. Müşteri ilişkileri: Otomatik müşteri destek botları ve kişiselleştirilmiş müşteri iletişimleri.

Sağlık: Tıbbi danışmanlık: Hastaların sorularını yanıtlayabilir, genel sağlık önerileri verebilir. Tıbbi verilerin analizi: Tıbbi raporların özetlenmesi ve doktorlara karar destek sistemleri sağlanması.



Yaratıcılık: Görsel ve müzik üretimi: Görsel ve ses verileri üzerinde çalışarak yeni sanat eserleri, müzikler veya tasarımlar oluşturabilir. Senaryo yazma: Sinema ve televizyon için yaratıcı hikaye ve senaryolar oluşturabilir.

Güvenlik ve Etik Kaygıları

Yanlış Bilgi ve Manipülasyon: Deepfake metinler, dezenformasyon yayma potansiyeli

Kötüye Kullanım Senaryoları: Spam, kimlik avı, kötü amaçlı kod üretimi

Veri Gizliliği ve Kişisel Bilgiler: Eğitim verilerinden özel bilgilerin sızma riski

Modelin Saldırıya Açıklığı: Adversarial attack'lar ve prompt injection saldırıları

Kritik Sistemlerde Güvenlik Riski: Otomasyon sistemlerinde hata yapma riski ve sonuçları

Önyargı ve Ayrımcılık: Irk, cinsiyet, kültür gibi alanlarda zararlı önyargıların pekiştirilmesi

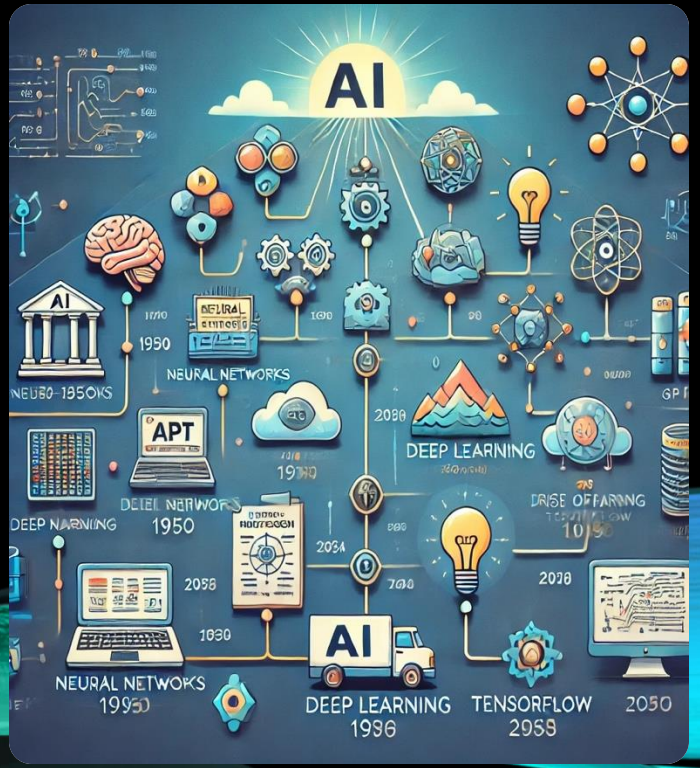
Şeffaflık ve Hesap Verebilirlik: Modellerin nasıl çalıştığının anlaşılması ve sorumluluğun belirlenmesi

İstihdam ve Sosyal Etkiler: Otomasyon nedeniyle iş kayıpları ve ekonomik etkiler

Sahte İçerik ve Telif Hakları: Üretilen içeriklerin orijinalliği ve yasal sorunlar

Erişim ve Adalet: Teknolojiye erişim eşitsizliği ve dijital uçurum

İnsan Denetimi ve Karar Mekanizmaları: İnsanların kritik kararların dışında bırakılması



Responsible AI: Etik ve Güvenilir Yapay Zeka

Adalet (Fairness)
Yapay zeka (AI) modelleri, tüm kullanıcılar ve gruplar için adil olmalıdır. Bu, modellerin farklı demografik gruplara karşı tarafsız kararlar almasını ve her bireyi eşit şekilde değerlendirmesini gerektirir.

Sorumluluk ve Hesap Verilebilirlik (Accountability)
Yapay zeka sistemlerinin nasıl ve neden belirli bir sonuca ulaştığını açıklayabilir olmalıyız. Sistemlerin şeffaf ve izlenebilir olması önemlidir.

Gizlilik ve Güvenlik (Privacy & Security)
Yapay zeka sistemlerinin kullandığı veriler kullanıcıların gizliliğini ihlal etmemeli ve güvenlik açıklarına karşı dayanıklı olmalıdır. Kişisel verilerin korunması en öncelikli etik sorumluluklardan biridir.

Responsible AI: Etik ve Güvenilir Yapay Zeka

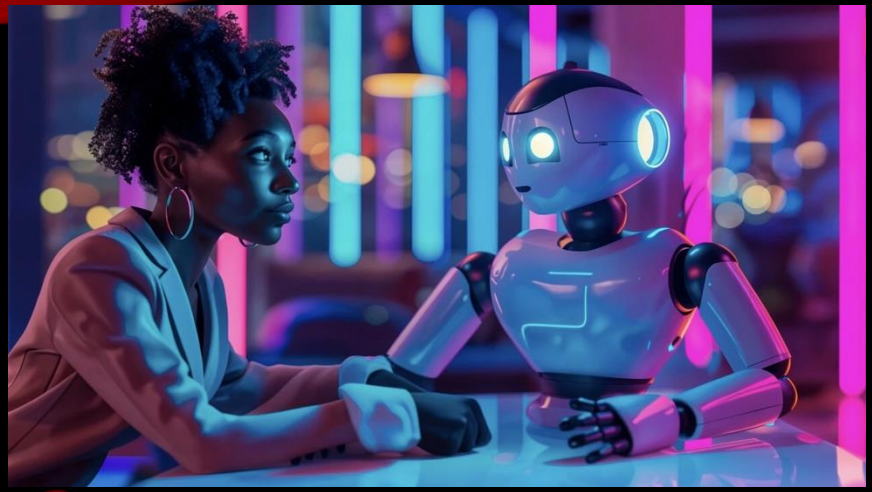
Çevresel Etki (Environmental Impact)

Yapay zeka modellerinin eğitimi ve çalıştırılması büyük miktarda enerji ve kaynak gerektirebilir. Modellerin çevresel etkisi de sorumlu yapay zeka uygulamalarında göz önünde bulundurulmalıdır.

Yanıtlanabilirlik (Responsiveness)

Yapay zeka sistemleri, toplumsal ihtiyaçlara göre güncellenmeli ve iyileştirilmelidir. Değişen şartlara karşı hızlı ve etkili bir yanıt verebilmelidir.

Güncel Çözümler ve İyileştirme Yaklaşımları



Explainability (Açıklanabilirlik) Çalışmaları
Model kararlarının yorumlanabilir hale getirilmesi

Regülasyonlar ve Standartlar
Yasal ve etik çerçeveler (örneğin Avrupa AI Act)

İnsan ve Makine İşbirliği Modelleri
İnsan denetimli sistemler ve sorumluluk paylaşımı

Model Değerlendirme ve Güvenlik Testleri
Toxicity ve bias testleri, adversarial testler

Filtreleme ve Moderasyon Sistemleri
Zararlı içeriklerin önlenmesi için katmanlı yaklaşımlar

Eğitim Veri Seti Düzenlemeleri
Dengeli, etik veri setleri oluşturma


Q&A

TEŞEKKÜRLER!

Kubilay Tuna

Senior Data Scientist

kubilaytuna26@hotmail.com