Büyük Dil Modellerinde (LLM) Değerlendirme Metodolojileri ve Süreç Denetimi

Bu rapor, modern yapay zeka mühendisliğinde modellerin performansını ölçmek için kullanılan stratejik yaklaşımları ve akıl yürütme (reasoning) kapasitelerini artıran denetim modellerini teknik bir perspektifle ele almaktadır.

1. LLM Değerlendirme Stratejileri (Evaluation Approaches)

Büyük Dil Modellerinin başarısını ölçmek, klasik regresyon veya sınıflandırma metriklerinden (Accuracy, F1-Score vb.) çok daha karmaşıktır. Sebastian Raschka'nın vurguladığı üzere, bu süreç dört ana sütun üzerinden yürütülmektedir:

A. Standart Kıyaslamalar (Benchmarks)

Modellerin genel yeteneklerini ölçmek için kullanılan statik veri setleridir.

- **MMLU (Massive Multitask Language Understanding):** Sosyal bilimlerden matematiğe kadar geniş bir bilgi yelpazesini test eder.
- **HumanEval:** Modellerin Python kodlama kabiliyetini ölçer.
- **Kısıtlar:** "Veri Sızıntısı" (Data Contamination) riski nedeniyle, modellerin bu test sorularını eğitim aşamasında görmüş olma ihtimali sonuçları yanıltabilmektedir.

B. LLM-as-a-Judge (Hakem Olarak LLM)

Bu yöntemde, GPT-4o veya Claude 3.5 Sonnet gibi yüksek kapasiteli modeller, diğer modellerin çıktılarını değerlendirmek için kullanılır.

- **Süreç:** Hakem modele belirli bir rubrik (puanlama kriteri) verilir. Model, hedef çıktıyı doğruluk, akıcılık ve güvenilirlik açısından puanlar.
- **Önemi:** İnsan değerlendirmesine en yakın ve en ölçeklenebilir otomasyon yöntemidir.

C. İnsan Değerlendirmesi (Human Evaluation)

Yapay zekanın nihai kullanıcısı insan olduğu için "altın standart" kabul edilir. **LMSYS Chatbot Arena**, bu yöntemin en popüler uygulama alanıdır. Elo puanlama sistemi kullanılarak modellerin kör testler aracılığıyla bir liderlik tablosu oluşturulur.

D. Model-Bazlı Metrikler (Perplexity)

Modelin bir metni ne kadar "tahmin edilebilir" bulduğunu ölçen matematiksel bir metriktir. Düşük **Perplexity** skoru, modelin veri setindeki dile olan aşinalığını gösterir; ancak bu her zaman yanıtın doğruluğu ile korele değildir.

2. Ödül Modellerinde Yeni Nesil: PRM vs. ORM

Modellerin akıl yürütme (reasoning) süreçlerini iyileştirmek için kullanılan ödül mekanizmaları (Reward Models), modellerin "nasıl" düşündüğünü belirleyen kritik unsurlardır.

Outcome-Supervision Reward Models (ORM)

Geleneksel yaklaşımdır. Model bir çözüm üretir ve sistem sadece sonucun doğruluğuna bakar.

- **Risk:** Modelin hatalı bir mantıkla (logical fallacy) doğru cevaba ulaşmasını (tesadüfi başarı) ödüllendirebilir.

Process-Supervision Reward Models (PRM)

Özellikle OpenAI'ın **o1** serisi gibi gelişmiş akıl yürütme modellerinde kullanılan yöntemdir.

- **Mekanizma:** Çözüm sürecindeki her bir ara adım (step-by-step) ayrı ayrı puanlanır.
- **Avantaj:** Yanlış mantık silsilelerini erkenden durdurarak halüsinasyonları minimize eder ve karmaşık matematik/kodlama problemlerinde başarı oranını dramatik şekilde artırır.

3. Endüstriyel Uygulama: RAG ve Üretim Ortamı

Değerlendirme metodolojileri, sadece akademik bir süreç değil, üretim (production) aşamasındaki sistemlerin güvenilirliği için bir zorunluluktur.

- **RAG Sistemleri:** Retrieval-Augmented Generation projelerinde **RAGAS** gibi çerçeveler kullanılarak "Faithfulness" (Sadakat) ve "Context Precision" (Bağlam Hassasiyeti) ölçülür. Bu, sistemin döküman dışına çıkıp uydurma bilgiler vermesini (hallucination) engellemek için kritiktir.

- **CI/CD Entegrasyonu:** Modern AI mühendisliğinde her model güncellemesi, otomatik bir değerlendirme hattından (evaluation pipeline) geçirilerek önceki versiyonlarla kıyaslanır.

Sonuç

LLM dünyası, sadece daha büyük modeller inşa etmekten, modelleri **daha iyi ölçebilen ve her düşünce adımını denetleyebilen** sistemler inşa etmeye doğru evrilmektedir. PRM ve LLM-as-a-judge yaklaşımları, bu yeni mühendislik standartlarının temel taşlarını oluşturmaktadır.