AIR LAB

LLMOps & Model Evaluation
Denetim ve Değerlendirme

Özge GÜNAY

AIR LAB

Bugün Nelerden Bahsedeceğim?

- Model Evaluation Nedir
- OpenAI Evals
- Intrinsic Metrics
- LLM-Based Evaluation
- Geleneksel Değerlendirme Yöntemleri
- Karşılaştırmalı Bakış

AIR LAB

TERİMLER SÖZLÜĞÜ

Benchmark

Modelleri karşılaştırmak için kullanılan, önceden hazırlanmış standart test setleridir. Herkes aynı testleri kullanır, böylece sonuçlar karşılaştırılabilir olur.

Model Evaluation (Model Değerlendirme)

Bir yapay zekâ modelinin ne kadar iyi çalıştığını ölçme sürecidir. Ürettiği cevapların doğruluğu, kalitesi ve tutarlılığı bu süreçte incelenir.

Human Evaluation

Model çıktılarının insanlar tarafından okunup yorumlanarak değerlendirilmesidir. Doğallık, bağlama uygunluk ve kullanıcı deneyimi gibi nitel özellikler burada öne çıkar.

Large Language Model (LLM)

İnsan diliyle yazılmış metinleri anlayabilen ve yeni metinler üretebilen büyük yapay zekâ modelleridir. Sohbet eden, özet çıkaran ya da soru cevaplayan yapay zekâların temelini oluşturur.

LLM-Based Evaluation

Bir yapay zekâ modelinin çıktılarının, başka bir LLM kullanılarak değerlendirilmesidir. Yani yapay zekâyı yine yapay zekâya yorumlatma yaklaşımıdır.

Intrinsic Metrics

Model çıktısını, önceden belirlenmiş doğru cevaplarla karşılaştırarak ölçüm yapan sayısal metriklerdir. Genellikle otomatik değerlendirmede kullanılır.

AIR LAB

TERİMLER SÖZLÜĞÜ

SuperGLUE

GLUE'un daha zor ve gelişmiş versiyonudur. Daha karmaşık dil anlama ve muhakeme yeteneklerini ölçer.

GLUE

Doğal dil işleme modellerini değerlendirmek için kullanılan popüler bir benchmark setidir. Birden fazla dil görevini içerir.

OpenAI Evals

LLM'lerin performansını ölçmek için geliştirilen açık kaynaklı bir değerlendirme framework'üdür. Otomatik test üretme, karşılaştırma ve analiz yapmayı kolaylaştırır.

BLEU

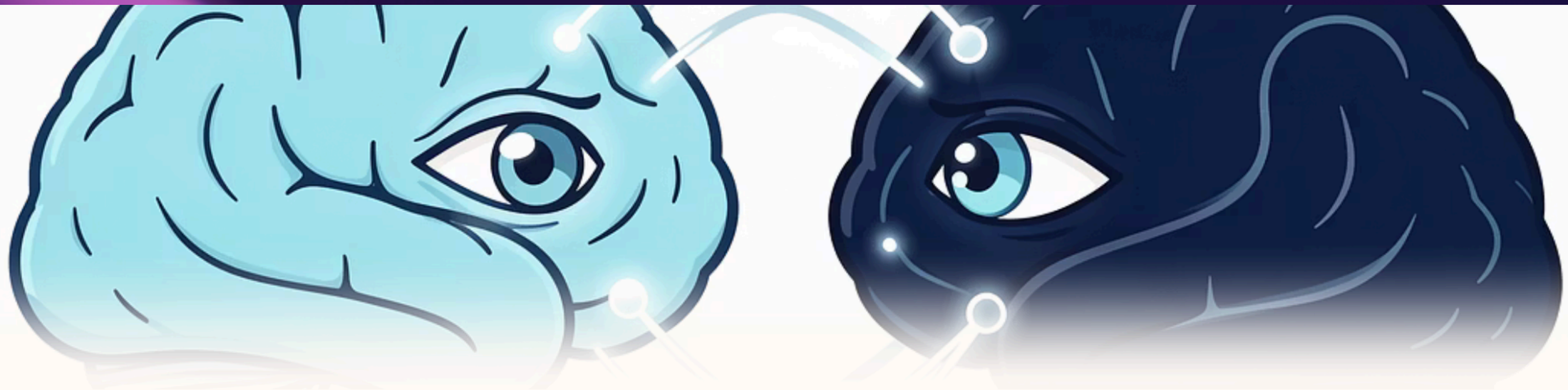
Özellikle makine çevirisinde kullanılan bir metriktir. Model çıktısındaki kelimelerin, referans çeviriyle ne kadar örtüştüğüne bakar.

ROUGE

Özetleme görevlerinde sık kullanılan bir metriktir. Modelin ürettiği özet ile gerçek özet arasındaki kelime ve ifade benzerliğini ölçer.

BERTScore

Kelime benzerliği yerine anlam benzerliğine odaklanan bir metriktir. İki metnin anlamsal olarak ne kadar yakın olduğunu ölçmeye çalışır.



LLM'lerin LLM'leri Değerlendirmesi: Yeni Bir Yaklaşım

Bu noktada devreye yeni bir fikir giriyor: LLM’leri yine başka LLM’lerle değerlendirmek. Amaç, insan gibi yorum yapabilen bir değerlendirme mekanizmasını, insanlara tek tek kontrol ettirmeden ve ölçeklenebilir bir şekilde sisteme dahil etmek. Kısacası, yapay zekânın ürettiği cevabı yine yapay zekâya “Bu ne kadar iyi?” diye sorduruyoruz.

Peki güzel... LLM’ler başka LLM’lerle değerlendiriliyor dedik. Ama bu tam olarak nasıl oluyor? Adım adım ve karmaşıklaştırmadan bakalım.

AIR LAB

LLM’ler LLM’leri Nasıl Değerlendiriyor?

01

Test Soruları Üretimi

İlk adımda, bir LLM’i kullanarak test soruları üretiyoruz. Ama rastgele değil. Farklı bağlamlarda, farklı giriş türlerinde ve farklı zorluk seviyelerinde olacak şekilde birçok senaryo oluşturuluyor. Böylece model sadece birkaç basit örnekle değil, gerçek hayata daha yakın ve geniş bir senaryo havuzunda test ediliyor.

02

Cevapların Toplanması

Sonra bu testler, değerlendirmek istediğimiz asıl LLM’e veriliyor ve modelin ürettiği cevaplar toplanıyor.

03

Cevapların Değerlendirilmesi

Asıl kritik nokta burada başlıyor. Bu cevapları kontrol eden bir başka LLM, önceden belirlediğimiz kriterlere göre yanıtları inceliyor. Yani “cevap doğru mu”, “akıcı mı”, “kendi içinde tutarlı mı” gibi sorulara insan gibi bakmaya çalışıyor.

04

Sonuçların Karşılaştırılması

En son aşamada ise bu değerlendirme sonuçları; ya bir referans modele ya da diğer LLM’lerin sonuçlarıyla karşılaştırılıyor. Böylece hangi modelin hangi konularda güçlü olduğu, nerelerde zorlandığı çok daha net bir şekilde ortaya çıkıyor.

AIR LAB

Neden Değerlendirmeyi LLM'lere Bırakıyoruz?

Ölçeklenebilirlik

LLM tabanlı değerlendirmenin en büyük artısı ölçeklenebilir olması. Yani çok sayıda test senaryosu kısa sürede ve otomatik olarak üretilebiliyor. Bu da insan gücüne olan ihtiyacı ciddi şekilde azaltıyor. Sonuç olarak modelleri, farklı görevler ve alanlar için çok daha hızlı test edebiliyoruz.

Esneklik

Bir diğer önemli avantajı esnek olması. Değerlendirme kriterleri, kullanım alanına göre rahatça değiştirilebiliyor. Mesela bir müşteri destek chatbotunda hız ve netlik önemliyken, bir eğitim asistanında açıklayıcılık ve tutarlılık ön planda olabiliyor. Aynı yöntem, farklı ihtiyaçlara göre uyarlanabiliyor.

Standardizasyon

Ayrıca tek bir LLM'in hem test üretmesi hem de değerlendirme yapması, insanlardan kaynaklanan öznel yorumları ve tutarsızlıkları azaltıyor. Yani herkesin "bence iyi" ya da "bence kötü" demesi yerine daha standart ve karşılaştırılabilir sonuçlar elde ediliyor.

Sürekli Gelişim

Son olarak, bu değerlendirmeler düzenli ve sürekli yapıldığında modellerin zaman içindeki performans değişimleri rahatça izlenebiliyor. Böylece hangi güncellemenin modeli gerçekten iyileştirdiği netleşiyor ve LLM'lerin sürekli gelişmesine doğrudan katkı sağlanıyor.

Tek Çözüm mü, Tamamlayıcı Bir Yöntem mi?

LLM Tabanlı Değerlendirme Diğer Yöntemlerin Yerini Alıyor mu?

Buraya kadar LLM tabanlı değerlendirmenin ne kadar güçlü olduğundan bahsettik. Ama doğal olarak şu soru geliyor: Peki bu yöntem diğer değerlendirme tekniklerinin yerini mi alıyor?

LLM tabanlı değerlendirme, sunduğu tüm avantajlara rağmen tek başına düşünülmemeli. Yani bu yöntem, mevcut değerlendirme tekniklerini tamamen ortadan kaldırmak için değil, onları tamamlamak için var.

Klasik metrikler bazı durumlarda hâlâ çok güçlü; özellikle net doğruluk gerektiren, sayısal ya da kesin cevaplı görevlerde. LLM tabanlı değerlendirme ise daha çok akıcılık, bağlama uygunluk ve tutarlılık gibi insan benzeri özelliklerin önemli olduğu durumlarda öne çıkıyor.

Bu yüzden en sağlıklı yaklaşım, tek bir yönteme körü körüne bağlı kalmak değil. Farklı değerlendirme tekniklerini birlikte kullanarak daha geniş ve dengeli bir bakış açısı elde etmek gerekiyor. Böylece hangi modelin hangi koşullarda gerçekten iyi performans gösterdiği çok daha net anlaşılabiliyor.

AIR LAB

İnsan Değerlendirmesi: Hâlâ Vazgeçilmez mi?

Şimdi biraz durup şunu düşünelim: Yapay zekâyı en iyi kim değerlendirebilir? Aslında hâlâ en güçlü cevaplardan biri şu: insan.

İnsan değerlendirmesi, LLM'lerin ürettiği cevapların uzmanlar ya da gerçek kullanıcılar tarafından incelenmesine dayanır. Burada amaç sayılar üretmekten çok, nitel geri bildirim almaktır. Yani cevap doğal mı, kullanıcıyı gerçekten tatmin ediyor mu, bağlama ya da kültüre aykırı bir durum var mı gibi daha ince detaylara bakılır.

Bu noktada insanlar çok güçlüdür. Çünkü bir cevabın "kulağa doğru gelip gelmediğini" ya da kullanıcıda nasıl bir etki bıraktığını makinelerden çok daha iyi anlayabilirler.

Ama bu yöntemin ciddi dezavantajları da var. İnsan değerlendirmesi hem zaman alıcı hem de maliyetlidir. Büyük ve sürekli çalışan sistemlerde bunu her çıktıda yapmak neredeyse imkânsızdır. Ayrıca farklı insanların aynı cevabı farklı şekillerde yorumlaması, sonuçların tutarsız ve öznel olmasına yol açabilir.

Bu yüzden insan değerlendirmesi genellikle tek başına kullanılmaz. Daha çok otomatik yöntemlerle birlikte kullanılarak, sistemin gerçekten doğru ve güvenilir çalışıp çalışmadığını kontrol eden bir denge unsuru olarak tercih edilir.

AIR LAB

Herkes İçin Aynı Test: Benchmark Yaklaşımı

Göreve Özel Benchmark'lar

Göreve özel benchmark’lar, LLM’leri önceden hazırlanmış standart görev setleri üzerinden değerlendirmeye yarar. GLUE ve SuperGLUE gibi benchmark’lar buna güzel örneklerdir. Her görev için belirlenmiş net metrikler vardır ve bu sayede farklı modeller arasında doğrudan, sayısal karşılaştırmalar yapılabilir.

Bu yaklaşımın en büyük avantajı, değerlendirme sürecinin tutarlı ve tekrarlanabilir olmasıdır. Ama burada da önemli bir sınırlama var. Bu benchmark’lar sadece kendi kapsadıkları görevlerle sınırlıdır. Gerçek hayattaki birçok kullanım senaryosu bu testlerin dışında kalabilir. Kısacası benchmark’lar bize faydalı ama dar bir pencere sunar.

Klasik Metrikler

Şimdi biraz da işin daha teknik ama sıkıcı olmayan tarafına bakalım. Yani modelleri sayılarla ölçmeye çalıştığımız klasik metriklere.

Intrinsic metrikler, modelin ürettiği cevapları önceden belirlenmiş doğru cevaplarla karşılaştırır. BLEU, ROUGE ve BERTScore gibi metrikler buna örnektir. Temelde kelime ya da cümle düzeyinde ne kadar benzerlik olduğuna bakarlar.

Bu metrikler, sözdizimsel yani yapısal benzerlikleri yakalamada oldukça iyidir. Ama sorun şu ki, iki cümle aynı anlama sahip olsa bile kelimeler farklıysa bu metrikler bunu her zaman doğru değerlendiremez. Yine de bu metrikler tamamen işe yaramaz değil. Modelin belirli görevlerde nasıl davrandığına dair ipuçları verir.

AIR LAB

Her Şey Bu Kadar Kolay mı?

LLM Tabanlı Değerlendirmenin Zorlukları ve Etik Riskleri

Buraya kadar LLM tabanlı değerlendirmenin ne kadar güçlü olduğundan bahsettik. Ama her güçlü yöntem gibi bunun da dikkat edilmesi gereken ciddi noktaları var.

Test Çeşitliliği Eksikliği

Eğer test senaryoları dar kalırsa, modelin kör noktaları ve önyargıları fark edilmeden geçebilir. Bu yüzden farklı bağlamları ve senaryoları kapsayan zengin test setleri oluşturmak çok kritik.

Yanlış Metrik Seçimi

Bir diğer önemli konu, doğru metrikleri seçmek. Yanlış veya eksik metrikler kullanıldığında, modelin performansı olduğundan iyi ya da kötü görünebilir. Yani değerlendirme hatalıysa, çıkan sonuçlar da yanıltıcı olur.

Etik Önyargılar

Etik açıdan bakıldığında ise daha hassas bir noktaya geliyoruz. Değerlendirici olarak kullanılan LLM'ler de önyargılar taşıyabilir. Eğer bu modeller yanlıysa, değerlendirme sonuçları da adaletsiz olabilir. Bu nedenle değerlendirme sürecinde şeffaflık, hesap verebilirlik ve adalet gibi etik ilkelerin merkezde olması gerekir.

İnsan Mantığıyla Uyumsuzluk

Microsoft Research'ün yaptığı çalışmalar da bu riskleri destekliyor. Araştırmalar, LLM'lerin değerlendirme sırasında kullandığı gerekçelendirme süreçlerinin her zaman insan mantığıyla örtüşmediğini gösteriyor. Bu tür sorunları azaltmak için değerlendirici LLM'lerin, insan değerlendirme mantığına daha yakın olacak şekilde dikkatli prompt engineering ile yönlendirilmesi gerekiyor.

OpenAI Evals ile Pratik Bir Çözüm

Şimdiye kadar hep yöntemlerden konuştuk. Peki tüm bunları gerçek hayatta uygulamak için elimizde ne var? İşte burada OpenAI Evals devreye giriyor.

OpenAI Evals, OpenAI tarafından geliştirilen ve LLM'leri değerlendirmek için kullanılan açık kaynaklı bir framework. Temel amacı, model performansını sistematik ve tekrarlanabilir bir şekilde ölçmeyi kolaylaştırmak.

Bu araç sayesinde geliştiriciler; veri setlerinden otomatik olarak prompt'lar üretebiliyor, model çıktılarının kalitesini ölçebiliyor ve farklı modelleri birbiriyle karşılaştırabiliyor. Hatta OpenAI'nin kendi modellerini değerlendirirken kullandığı hazır şablonlar da bu framework'ün içinde yer alıyor.

Evals'in en güzel yanlarından biri, kod yazmadan evaluation yapılabilmesine imkân tanıması. Bu da değerlendirme sürecini çok daha erişilebilir hale getiriyor.

Özetle şunu söyleyebiliriz: Yapay zekâyı anlamanın en iyi yollarından biri, ona kendini değerlendirmeyi öğretmektir.

AIR LAB

En İyi yapay zeka, sadece iyi cevap veren değil; doğru şekilde değerlendirilen yapay zekadır.

Teşekkürler!

ozgegunay64@gmail.com