Yapay Zekada Faillik ve Eşitlik Bozma (Tie-breaking)

1. Failliğin (Agency) Tanımı ve Karar Verme

- Faillik, bir varlığın kasıtlı ve özerk eylemler gerçekleştirme yeteneğidir.
- Eylem (action) ile basit bir hareket (act) arasındaki fark, eylemin bir seçime dayanmasıdır.
- Karar verme süreci, rasyonel failliğin temel taşıdır ve ajanın hedeflerine ulaşmak için en iyi araçları belirlemesini sağlar.

2. "Eşitlik" (Tie) Problemi ve Gönüllü Nedenler

- Bazen bir fail, iki seçenek arasında kaldığında her iki seçeneği seçmek için de eşit nedenlere sahip olabilir.
- **Gönüllü Nedenler (Voluntarist Reasons):** Dışsal veya verili nedenlerin tükendiği noktada, failin kendi iradesiyle yarattığı nedenlerdir.
- Bu yetenek, failliğin ayırt edici (hallmark) bir özelliği olarak kabul edilir.
- İnsanlar bu yetenek sayesinde "karar felci" yaşamaktan kurtulurlar.

3. Yapay Zekada Mevcut Durum (Teknik Kısıtlar)

- AI sistemleri (özellikle topluluk ağları - ensemble networks), oylama süreçlerinde sık sık eşitlik durumlarıyla karşılaşır.
- **Mevcut Çözüm Stratejileri:**
    - **Mantık:** Eğitim verisindeki frekanslara dayalı varsayımlar.
    - **Matematik:** Güven değerlerinin ortalaması veya k-NN algoritmaları.
    - **Keyfiyet:** Programlama kodundaki indeks numarasına göre seçim (örneğin en düşük ID'li ağın seçilmesi).
    - **Rastgelevari:** Predictable olmayan rastgele sayı üretimi.

4. Etik ve Teknik Eleştiri

- AI'nın kullandığı bu yöntemler, yüksek riskli (high-stakes) kararlarda (işe alım, otonom araçlar vb.) "gönüllü nedenler"in yerini tutmaz.
- AI, bir nedenin "neden" önemli olduğunu veya kararın bağlamını kavrayamaz.
- Gerçek bir faillik için AI'nın şu özelliklere ihtiyacı vardır:
    - **Öz-Kavrayış:** "Ben" duygusu ve öz-farkındalık.
    - **Anlaşılabilirlik (Intelligibility):** Dünyayı ve kendi kararlarını anlamlandırma dürtüsü.

https://www.researchgate.net/publication/379410238_Artificial_Intelligence_and_Agency_Tie-breaking_in_AI_Decision-Making/link/660781d1b839e05a20a9a894/download

1. Eşitlik Durumları Nerede Oluşur?

Metin, LLM süreçlerinde eşitliğin dört ana noktada ortaya çıktığını belirtiyor:

- **Token Seviyesinde:** Bir sonraki kelime tahmini yapılırken birden fazla kelimenin (token) aynı olasılık değerine sahip olması.
- **Sekans (Dizi) Seviyesinde:** "Beam search" gibi arama algoritmalarında, farklı cümle yapılarının toplam skorlarının birbirine eşit çıkması.
- **Kopya Hipotezler:** Modelden birden fazla örnekleme yapıldığında, birbirinin tıpatıp aynısı olan çıktıların üretilmesi.
- **Değerlendirme/Sıralama:** Metrik skorları (BLEU, ROUGE vb.) birbirine çok yakın veya eşit olan modellerin kıyaslanması.

2. Teknik Çözüm Stratejileri

Metinde bahsedilen somut çözüm yaklaşımları şunlardır:

- **Deterministik (Belirlenimci) Kurallar:** Eğer iki token eşitse, genellikle sözlükteki (vocabulary) ID numarası küçük olan seçilir. Bu, üretimin her seferinde aynı olmasını sağlar ancak modelin çıktılarını sözlük sıralamasına duyarlı hale getirebilir.
- **Stokastik (Rastlantısal) Yaklaşımlar:** Çeşitlilik istenen durumlarda rastgele seçim yapılması. Metin, kapalı uçlu (doğruluk odaklı) görevlerde deterministik, açık uçlu (yaratıcılık odaklı) görevlerde ise stokastik yöntemlerin tercih edildiğini vurguluyor.
- **Aritmetik Örnekleme (Arithmetic Sampling):** Çıktı çeşitliliğini kontrol etmek ve tekrarları önlemek için kullanılan daha teorik bir yöntem.

3. Uygulama İçin Pratik Tavsiyeler

Eğer bir LLM sistemi geliştiriyorsan metin şu önerilerde bulunuyor:

- **Tekrarlanabilirlik İçin:** Sabit bir sıralama kuralı ve ikincil bir sıralama kriteri belirle.
- **Çeşitlilik İçin:** Sadece rastgeleliğe güvenmek yerine, kopyaları aktif olarak azaltan algoritmalar kullan.
- **Değerlendirme İçin:** Eşitlik durumlarını "hata" olarak değil, bir veri kategorisi olarak ele al ve metriklerini buna göre kalibre et.

Emergent mind

lmarena