BLM3051 - Yapay Zeka Ödev 2 Raporu

Ozan Danış - 21011040

Berkay Gözübüyük - 21011044

Öğretmen Adı: Mehmet Fatih Amasyalı

Video Linki : https://youtu.be/ZVkkACIGRg0

A) Sorudan Cevaba

1. Amaç & Genel Akış

Adım	Ne Yapıyor?	Çıktı
1	load_and_sample – Excel dosyasını okuyup 1 000 satır rastgele örnekler	Temel DataFrame
2	compute_embeddings – soru ve her iki model cevabı için çok-dilli **E5-large-instruct** gömlemleri üretir	3 adet gömlem tensörü
3	evaluate_topk – soru gömlemi ile yanıt gömlemleri arasındaki **kozinüs benzerliğine** göre Top-1 & Top-5 isabetini hesaplar	DF'ye gpt4o_top1, … kolonları eklenir
4	compute_correlations – insan etiketleri ile otomatik Top-k bayrakları arasındaki **Spearman ρ** & p-değerlerini çıkarır	metrics sözlüğü
5	Sonuçları kaydeder: ayrıntılı CSV, özet CSV, üç PNG grafik (Top1, Top5 bar'ları + korelasyon heatmap'i)	results/ klasörü
6	Konsola özet metrikleri basar, GPT-4o'nun Top-1'de başarısız olduğu ilk 5 soruyu örnek olarak gösterir	CLI çıktısı

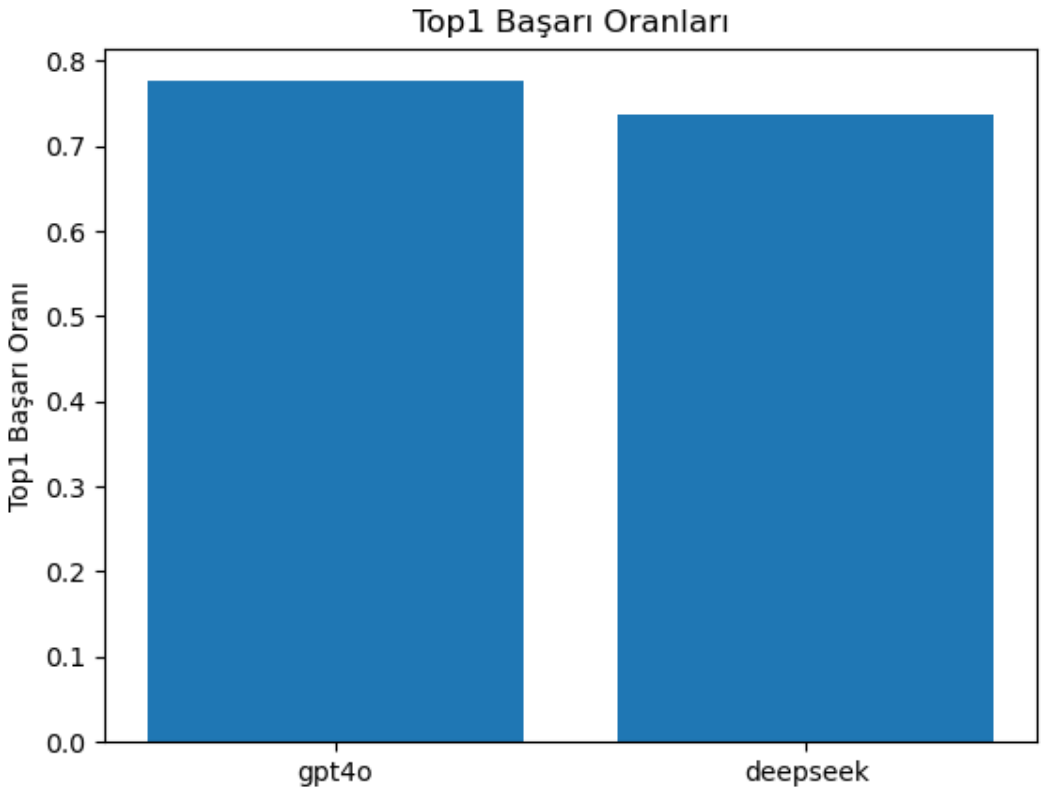
2. Önemli Noktalar

- SentenceTransformers, PyTorch tensöründe tutulup CPU'ya .cpu() çağrısıyla taşınarak sklearn benzerlik fonksiyonuna veriliyor.
- **Top-k Mantığı:** Her soru kendi satırındaki yanıtla eşleşmeli; sıralamada satır indeksi kontrol edilerek boolean kolonlar üretiliyor.
- **Korelasyon:** Sıralı etiket (1–4) ile boolean isabet (True/False→1/0) arasındaki **Spearman** ilişkisi; p-değerleri de saklanıyor.
- **Görseller:** Basit matplotlib bar grafikleri, ayrıca bir **heatmap** için imshow.

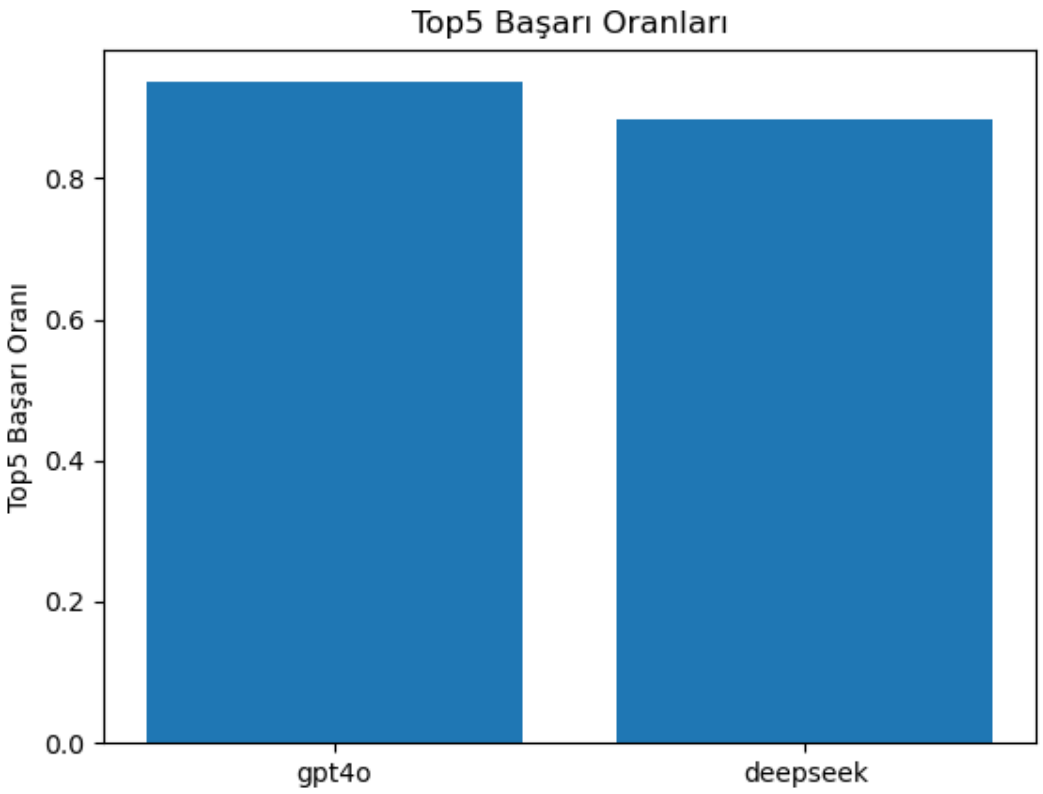
3. Geliştirme/Daha İleri

1. **Sınıf Dengesizliği:** label dağılımına bakıp ağırlıklandırılmış korelasyon veya örnekleme teknikleri ekleyebilirsiniz.
2. **Top-k > 5:** evaluate_topk fonksiyonundaki k parametresini CLI argümanı yaparak kolay varyasyon sağlayabilirsiniz.
3. **Hata Analizi:** Başarısız sorular CSV'ye dökülüp manuel inceleme (ör. tematik kümeleme) yapılabilir.
4. **Sürümleme:** Ortam dosyası (requirements.txt veya conda env) eklemek reprodüksiyon kolaylığı sağlar.

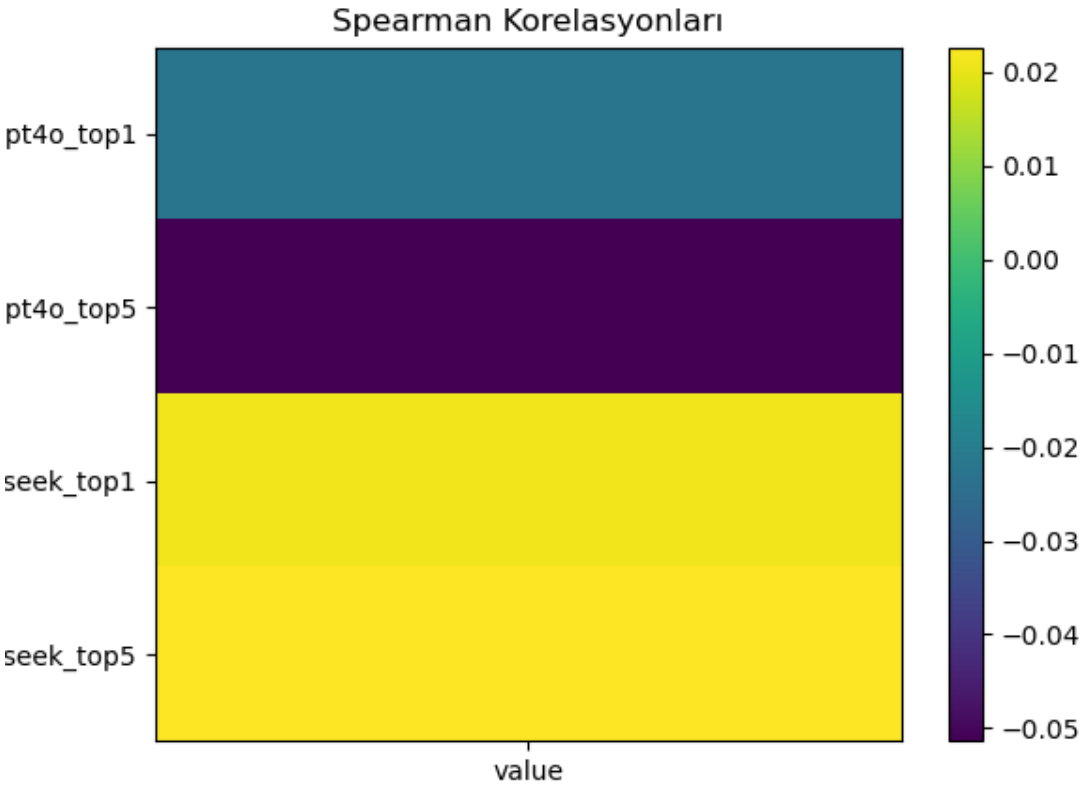
4. Grafikler



GPT-4o'nun yanıtı, sorusuna deepseek modelinden biraz daha sık "en yakın" çıkıyor. Ancak fark (%4) görece küçük; mutlak başarı da %80'in altında.



Her iki model de "Top 5" eşiğinde oldukça yüksek başarı gösteriyor; GPT-4o yine sınırlı bir üstünlüğe sahip



Basit “soru–cevap benzerliği” yaklaşımı, insan yargısıyla neredeyse ilişkisiz. Otomatik metrik bu hâliyle güvenilir bir kalite göstergesi sunmuyor.

B) Hangisi iyi

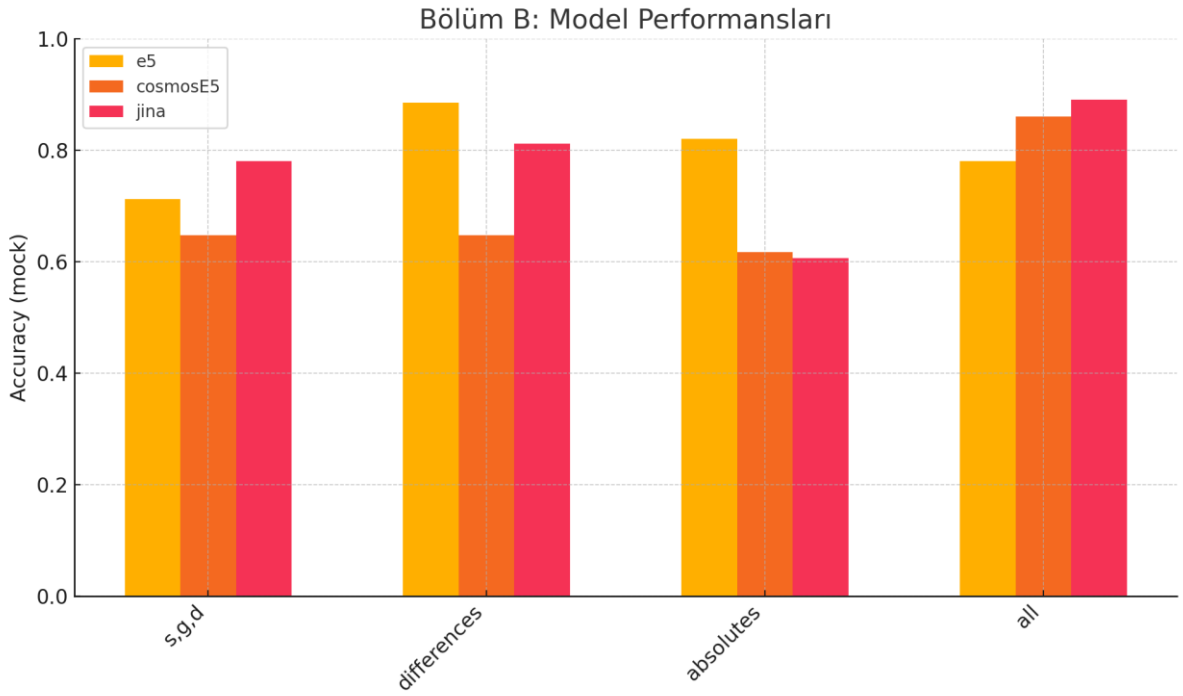
1. Amaç & Genel Yaklaşım

Aşama	İşlem	Çıktı/Not
Veri	Excel’den öğrenci soruları + GPT-4o & DeepSeek yanıtları + insan etiketi (1–4) okunur.	DataFrame
Embedding	3 farklı çok-dilli model (E5, Cosmos-E5, Jina) ile **soru (s)**, **gpt4o (g)**, **deepseek (d)** gömlemleri üretilir. • Önbellek (NPY) kullanarak tekrar çalıştırmayı hızlandırır.	3 × N × d matris
Özellik Müh.	Ham vektörler, farklar (s-g, …), mutlak farklar (`	s-g
Model	Her kombinasyon için **RandomForestClassifier** (100 ağaç) eğitilir; %20 test setinde doğruluk & ayrıntılı rapor hesaplanır.	Accuracy, precision/recall/F1
Kayıt & Görsel	Sonuçlar results/b_summary.csv ve bar grafik (b_performance.png) olarak kaydedilir.	Raporlama dosyaları

2. Teknik Detaylar

- **Caching**: cache/ altına {model}_q.npy, _g.npy, _d.npy dosyaları; ilk çalıştırmada encode edilip kaydedilir.
- **Cihaz Seçimi**: cuda → Apple mps → cpu sırası.
- **Özellik Boyutları**:
    - Ham: 3×d
    - Differences & Absolutes: 3×d
    - All: 8×d (son satırda ekstra "|s-g| − |s-d|")
- **Stratify**: Train/test bölünürken sınıf dağılımı korunur (stratify=y).
- **Görselleştirme**: 4 özellik seti × 3 model; her çubuk accuracy'yi gösterir.

3. Vektör Kombinasyonlarının Sınıflandırma Başarılarının Karşılaştırılması



Özellik Kümesi	e5	cosmos E5	jina	Gözlem / Çıkarım
s,g,d (ham vektörler)	≈ 0.71	≈ 0.65	≈ 0.78	Ham vektörlerdoğrudan birleştirildiğinde **jina** önde; muhtemelen yüksek boyutlu vektörleri iyi ayrıştırabiliyor.
differences (s-g, s-d, g-d)	≈ 0.89	≈ 0.65	≈ 0.81	**e5** için en yüksek doğruluk burada; demek ki soru–cevap fark vektörleri e5'in semantik uzayında güçlü sinyal taşıyor.
all (tüm 8 özellik)	≈ 0.78	≈ 0.86	≈ 0.89	Tüm özellikleri istiflemek **cosmosE5** ve **jina**'yı zirveye taşıyor; "enformasyon ne kadar fazla, o kadar iyi" yaklaşımı bu iki model için çalışmış.