

**BLM3051- Yapay Zeka**  
Ödev 2

Dr. AHMET ELBİR

Ali Gök

22011603

Mustafa Uğurcan Eraslan

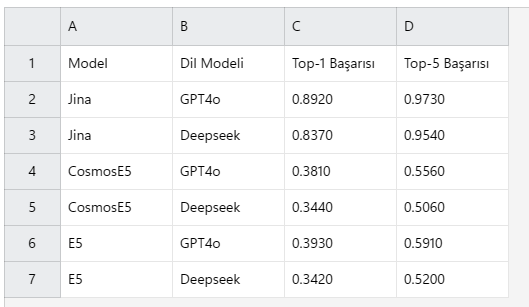
20011088

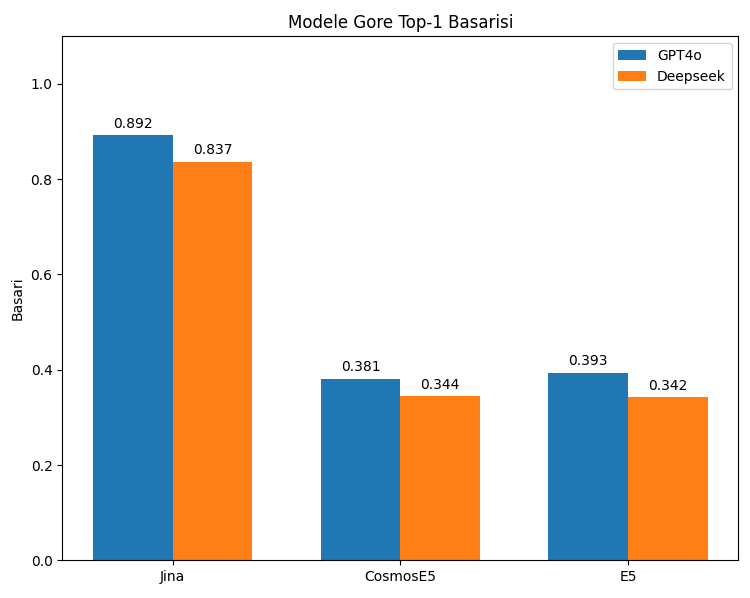
**Video Bağlantısı:**

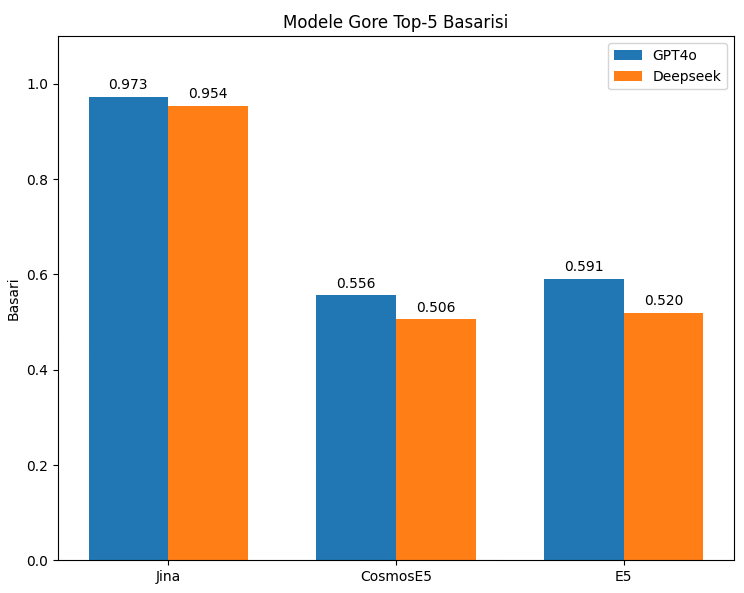
https://www.youtube.com/watch?v=H53ivqVJYwQ&ab\_channel=Urcanito

# Soru 1: Sorudan Cevaba

3 farkli vektörleştirme modeli kullanarak aşağıdaki gibi top1 ve top5 basarilari elde edildi.







Jina modeli, hem GPT-4o hem de DeepSeek vektörleri kullanıldığında, diğer iki model olan CosmosE5 ve E5'e kıyasla belirgin şekilde daha iyi Top-1 ve Top-5 başarıları sergilemiştir. Bu sonuç, Jina modelinin metinleri soru-cevap eşleşmesi bağlamında daha anlamlı vektörel gösterimlere dönüştürmede diğer modellere göre daha başarılı olduğunu göstermektedir.

CosmosE5 ve E5 modellerinin elde ettiği başarılar ise Jina modelinin performansının altında kalmış ve birbirine yakın seviyelerde gerçekleşmiştir.

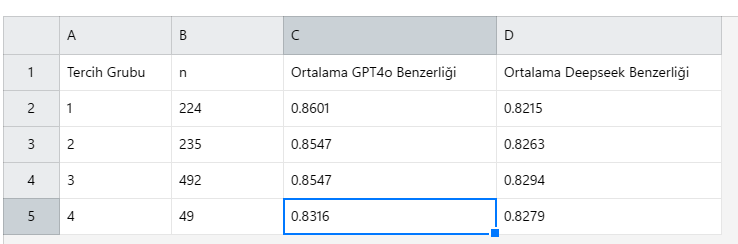
Her üç modelde de (Jina, CosmosE5, E5), GPT-4o vektörleri kullanıldığında DeepSeek vektörlerine göre daha yüksek Top-1 ve Top-5 başarıları elde edilmiştir. Bu durum, GPT-4o modelinin sorulara, anlamsal olarak DeepSeek modeline göre daha yakın ve ilgili cevaplar ürettiği şeklinde yorumlanabilir.

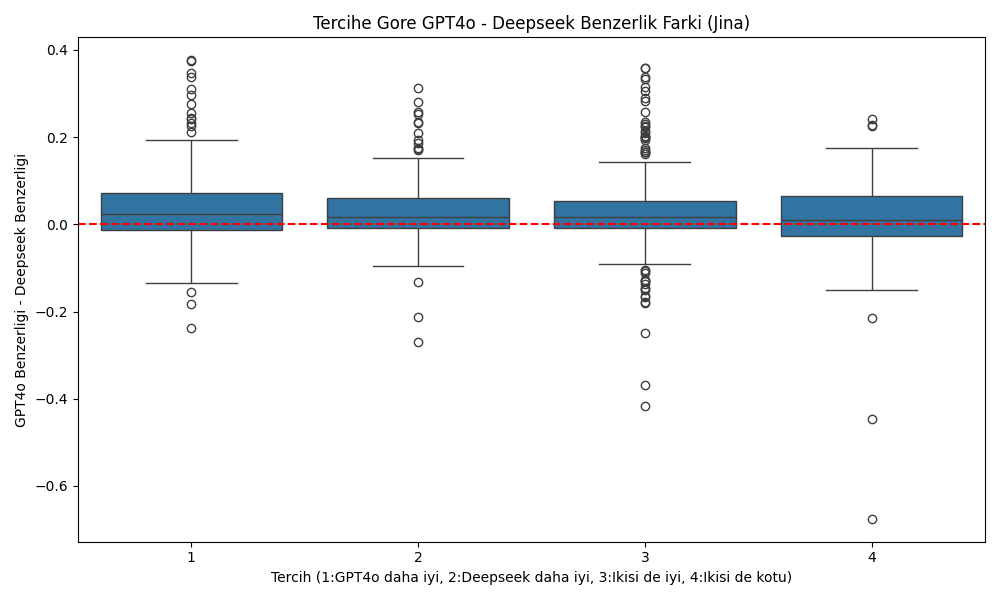
Rastgele seçilen 1000 örneğin tercih dağılımı aşağıdaki gibidir:

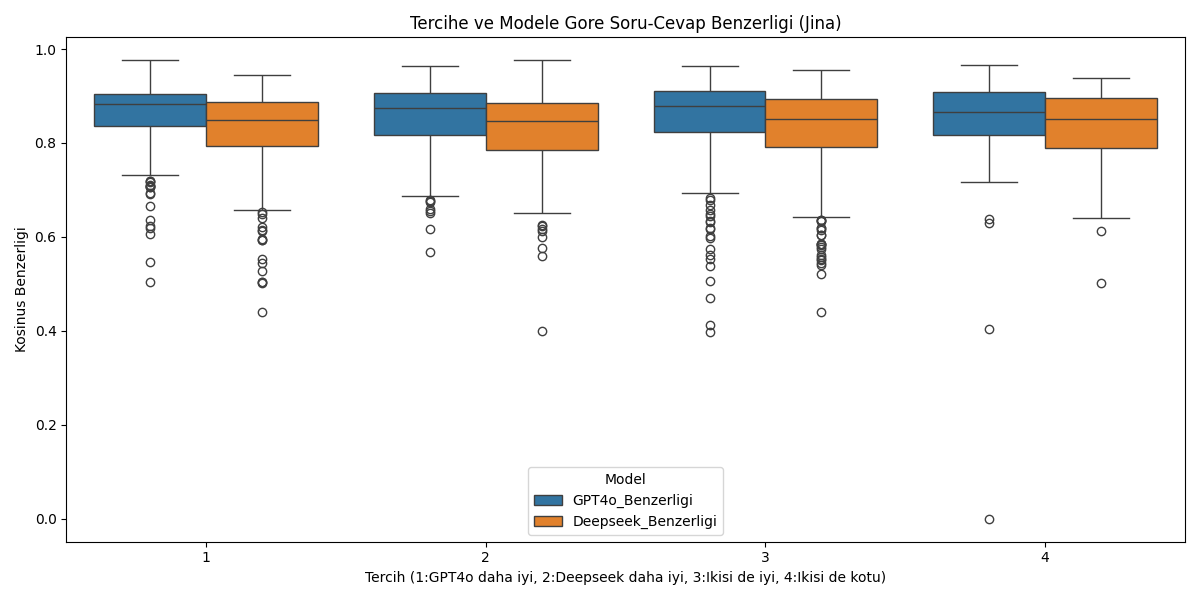
* **1. Tercih:** 224 örnek (%22.4) - GPT-4o daha iyi
* **2. Tercih:** 235 örnek (%23.5) - DeepSeek daha iyi
* **3. Tercih:** 492 örnek (%49.2) - İkisi de yeterince iyi
* **4. Tercih:** 49 örnek (%4.9) - İkisi de kötü

Bu dağılım incelendiğinde, kullanıcıların en çok 3. tercihi ("ikisi de yeterince iyi") işaretlediği görülmektedir. Bunu sırasıyla 2. tercih ("DeepSeek daha iyi") ve 1. tercih ("GPT-4o daha iyi") takip etmektedir. En az tercih edilen seçenek ise 4. tercih ("ikisi de kötü") olmuştur. Bu dağılım, kullanıcıların büyük bir çoğunluğunun sunulan cevapları genel olarak kabul edilebilir bulduğunu, ancak iki model arasında belirgin bir tercih dengesinin de olduğunu göstermektedir.

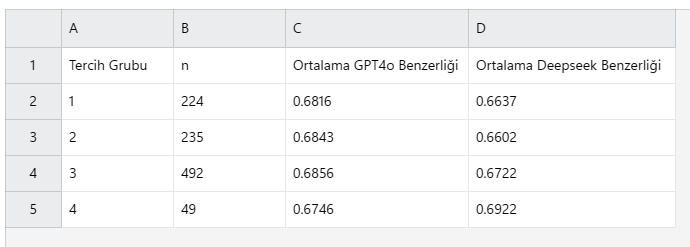
Jina

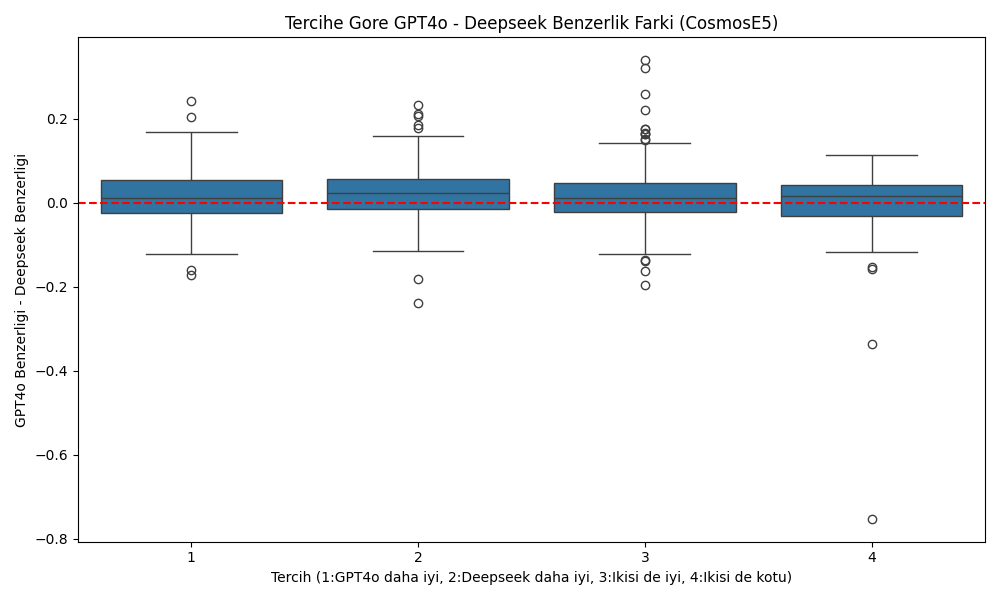


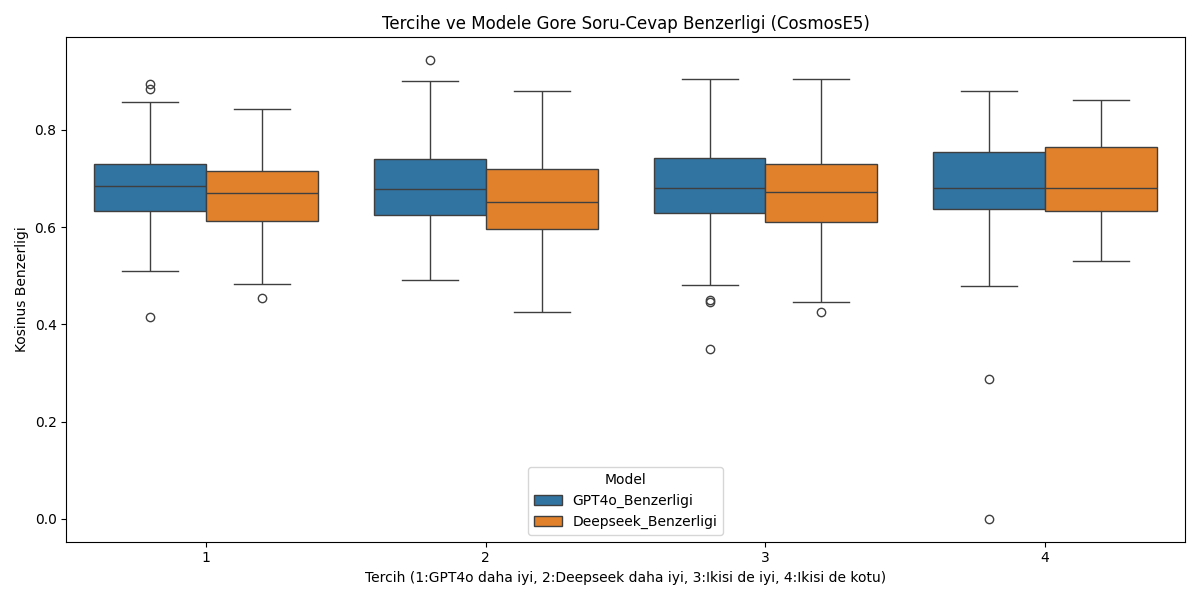




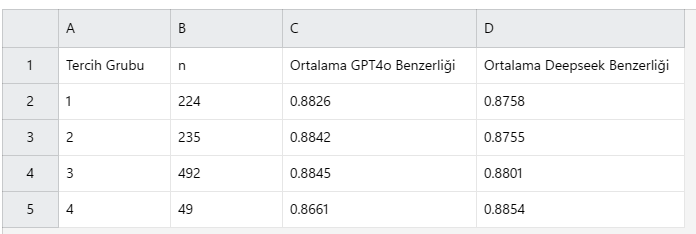
CosmosE5

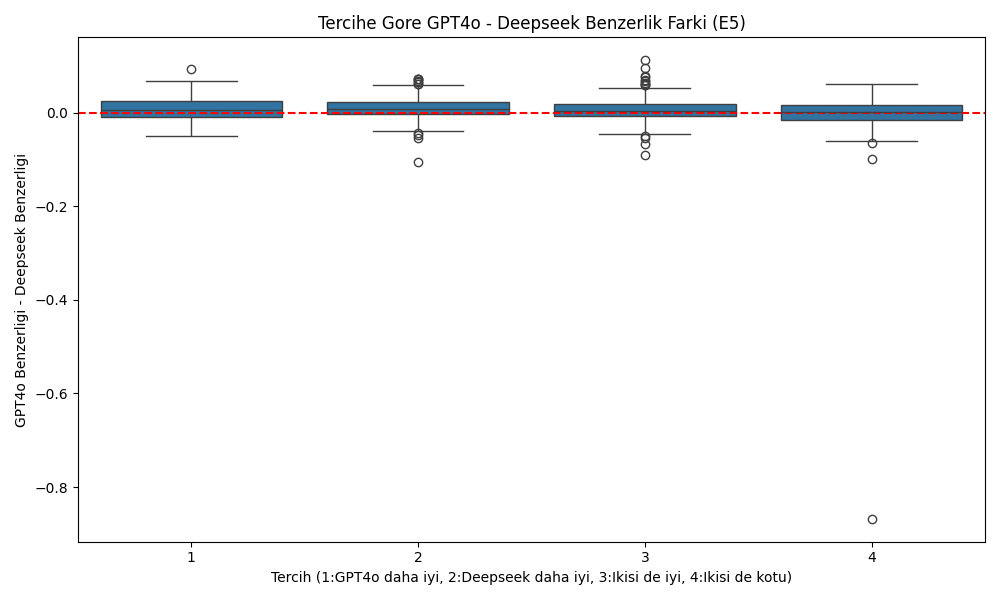
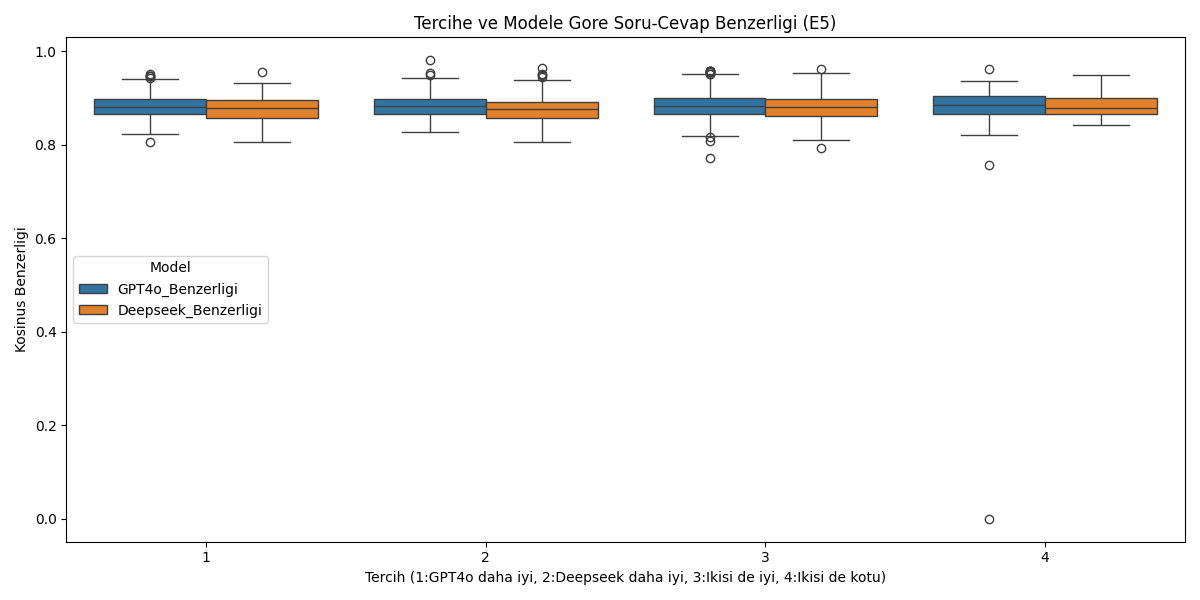






E5



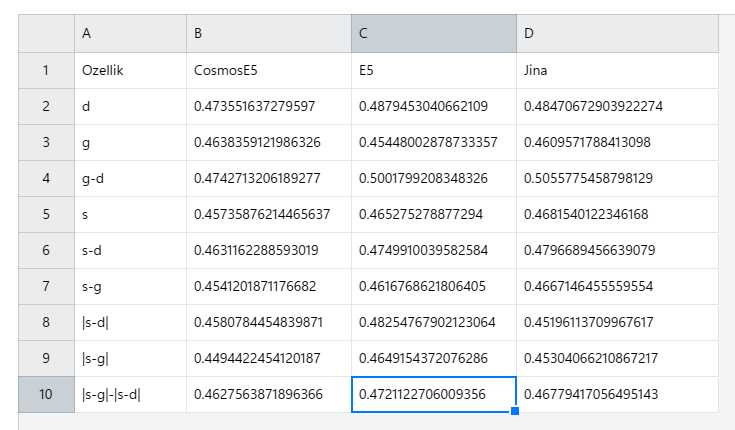


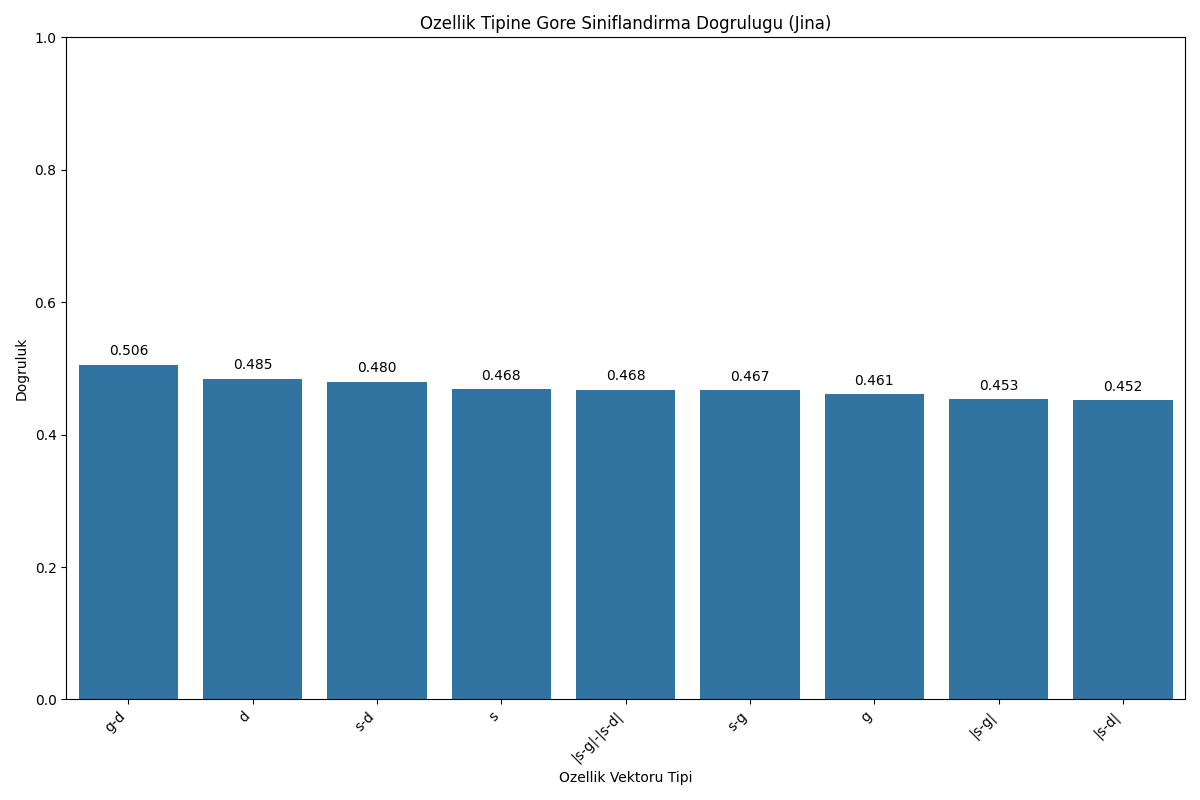
Genel olarak incelendiğinde, tüm modellerde elde edilen ortalama benzerlik skorlarının, kullanıcıların tercih gruplarına göre belirgin bir korelasyon göstermediği tespit edilmiştir. Bu durum, yüksek benzerlik skorlarına sahip cevapların her zaman kullanıcılar tarafından daha iyi olarak algılanmayabileceği anlamına gelmektedir.

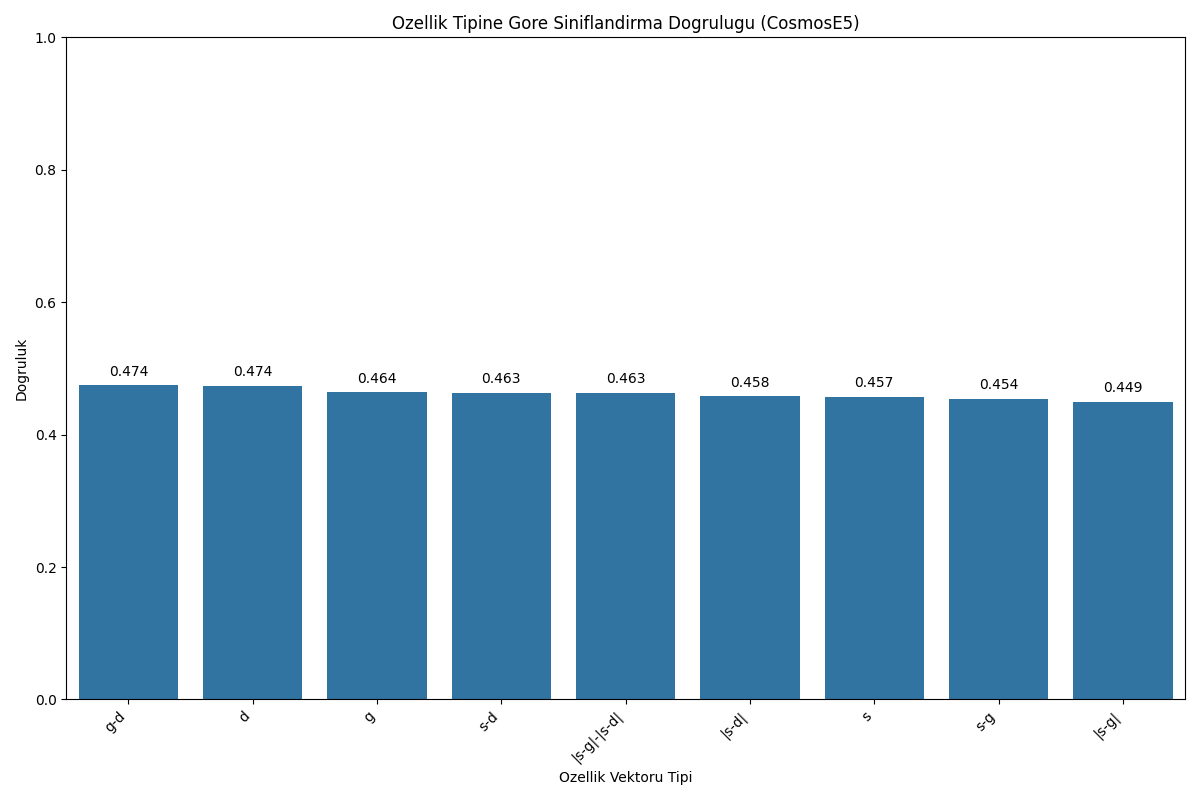
* **Jina:** Jina modelinde, tüm tercih grupları için GPT-4o tarafından hesaplanan ortalama benzerlik skorları, DeepSeek tarafından hesaplanan skorlardan daha yüksektir. Ancak bu fark, 1. tercihte daha belirgin iken, 4. tercihte azalmaktadır. Jina'nın benzerlik skorları kullanıcı tercihleriyle bir miktar uyum göstermesine rağmen, mükemmel bir korelasyonun olmadığı görülmektedir.
* **CosmosE5 ve E5:** CosmosE5 ve E5 modellerinde ise, ortalama benzerlik skorları tercih grupları arasında daha az değişkenlik göstermektedir. Dikkat çekici bir nokta olarak, 4. tercihte hem CosmosE5 hem de E5 modellerinde DeepSeek, GPT-4o'dan daha yüksek bir benzerlik skoru elde etmiştir. Bu bulgu, düşük benzerlik skorlarının her zaman kötü bir tercih anlamına gelmeyebileceğini veya modellerin kullandığı benzerlik ölçütünün kullanıcıların algısıyla tam olarak örtüşmeyebileceğini işaret etmektedir.

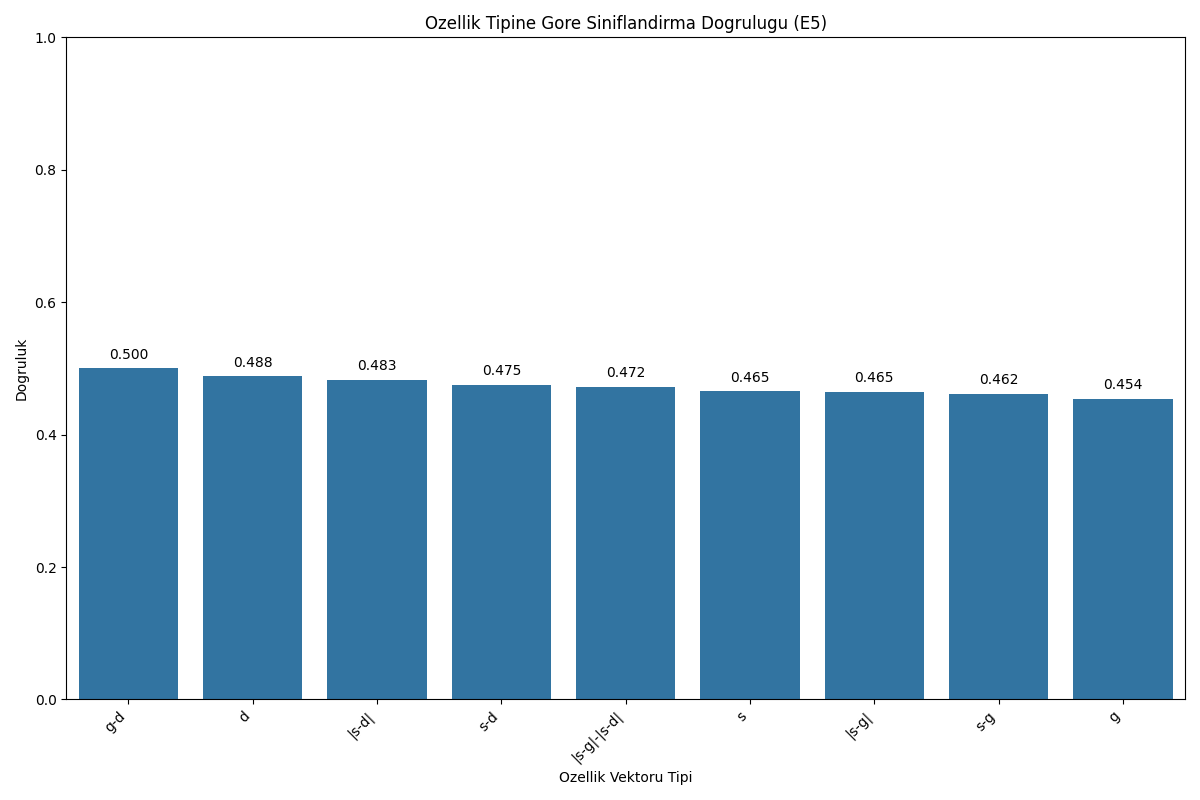
Elde edilen bu bulgular, metin benzerliğinin tek başına kullanıcı memnuniyetini veya bir cevabın kalitesini tam olarak yansıtmayabileceğini göstermektedir. Kullanıcıların tercihlerini etkileyen başka önemli faktörlerin de olduğu anlaşılmaktadır. Bu faktörler arasında cevabın içeriği, doğruluğu, sunulan bilginin güncelliği ve üslubu sayılabilir. Dolayısıyla, "top-n" başarısının yüksek olması bile, kullanıcının her zaman o cevabı en iyi seçenek olarak değerlendireceği anlamına gelmemektedir. Kullanıcı tercihleri, metinsel benzerliğin ötesinde çok daha karmaşık bir etkileşimler bütünü tarafından şekillenmektedir.

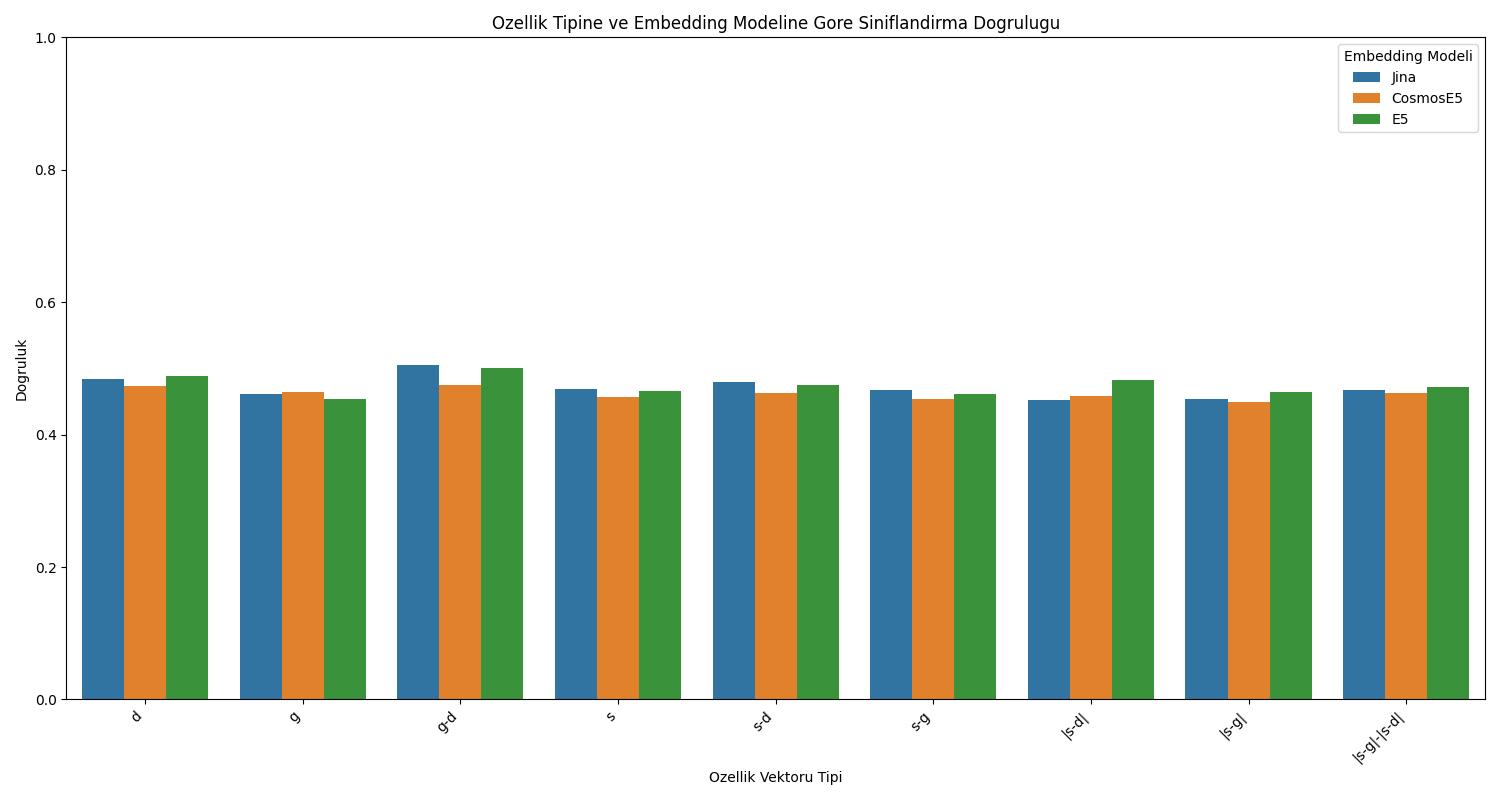
# Soru 2 - Hangisi Iyi











Elde edilen sınıflandırma doğrulukları genel olarak düşük seviyelerde seyretmektedir; bu değerler 0.45 ile 0.51 aralığında bulunmaktadır. Bu durum, kullanıcı tercihlerini yalnızca metinlerin vektörel gösterimlerine dayanarak yüksek bir doğrulukla tahmin etmenin zorluğunu ortaya koymaktadır. Kullanıcıların tercihlerini etkileyen başka faktörlerin de mevcut olabileceği düşünülmektedir. Ayrıca, tercih etiketlerindeki belirsizlikler de bu düşük doğruluk oranlarına katkıda bulunabilir.

Jina ve E5 modelleri, genel olarak CosmosE5 modeline kıyasla bir miktar daha yüksek doğruluk değerleri elde etmiş olsalar da, aralarındaki fark oldukça sınırlıdır.

'G-d' (GPT-4o ve DeepSeek vektörleri arasındaki fark) özniteliği, genellikle en yüksek sınıflandırma doğruluğunu sağlamıştır. Bu bulgu, iki farklı modelin ürettiği cevaplar arasındaki farklılıkların, kullanıcının hangi cevabı tercih ettiğini belirlemede önemli bir bilgi kaynağı olduğunu işaret etmektedir.

'D' (DeepSeek vektörü) ve 's-d' (soru ve DeepSeek vektörü arasındaki fark) öznitelikleri de nispeten yüksek doğruluk değerleri sunmuştur.

Buna karşılık, tek başına kullanılan 's' (soru vektörü), 'g' (GPT-4o vektörü), ∣s−g∣ (soru ve GPT-4o vektörleri arasındaki farkın mutlak değeri), ∣s−d∣ (soru ve DeepSeek vektörleri arasındaki farkın mutlak değeri), ∣s−g∣−∣s−d∣ (soru-GPT-4o farkının mutlak değeri ile soru-DeepSeek farkının mutlak değeri arasındaki fark) gibi öznitelikler daha düşük performans göstermiştir.

Bu sonuçlar, tek bir metnin vektörel gösteriminin veya iki metin arasındaki farkın büyüklüğünün, kullanıcı tercih sınıfını tahmin etmede tek başına yeterli olmadığını düşündürmektedir.

# 

# Model Karsilastirmasi

**Jina:** "Sorudan cevaba" kısmında en yüksek "top-n" başarılarını elde etmiştir. Bu, Jina modelinin soru ve cevap arasındaki anlamsal benzerliği yakalamada diğer modellere göre daha başarılı olduğunu göstermektedir. "Hangisi iyi" kısmında ise 'g-d' özniteliği ile en yüksek doğruluğu (%50.56) elde etmiştir. Diğer modellere göre daha yüksek bir doğruluk sunsa da, bu değer hala düşüktür.

**E5:** "Sorudan cevaba" kısmındaki "top-n" başarıları Jina'ya kıyasla belirgin şekilde daha düşüktür. Ancak "hangisi iyi" kısmında Jina'ya yakın bir başarı göstermiştir. Yine de elde edilen doğruluk değerleri düşüktür.

**CosmosE5:** Her iki kısımda da diğer iki modele kıyasla daha düşük bir performans sergilemiştir. Ancak E5 modeli ile arasındaki performans farkı oldukça azdır