**MAKALE BAŞLIĞI**

**TITLE OF PAPER**

*Ad SOYAD1* (ORCID: 0000-\*\*\*\*-\*\*\*\*-\*\*\*\*)

*Ad SOYAD,2\** (ORCID: 0000-\*\*\*\*-\*\*\*\*-\*\*\*\*)

1 Kahramanmaraş Sütçü İmam Üniversitesi, Elektrik Elektronik Mühendisliği Bölümü, Kahramanmaraş, Türkiye

\*Sorumlu Yazar / Corresponding Author: Ad SOYAD, yazar@email.com

**ÖZET**

Bu çalışma, Türkçe sağlık danışmanlığında dört farklı büyük dil modelinin (doktor-meta-llama-3-8b, doktor-LLama2-sambanovasystems-7b, doktor-Mistral-trendyol-7b ve doktor-llama-3-cosmos-8b) performansını incelemektedir. Modeller, 321.179 hasta-doktor soru-cevap çiftinden oluşan özel bir veri kümesi üzerinde ince ayar yapılarak eğitilmiştir. Modellerin başarımı BLEU, BERT Score gibi sentetik metrikler, Elo puanlaması ve uzman doktor değerlendirmeleri ile ölçülmüştür. Sonuçlar, doktor-LLama2-sambanovasystems-7b modelinin genel başarım açısından en iyi performansı gösterdiğini, doktor-Mistral-trendyol-7b modelinin ise en düşük zararlı cevap oranına sahip olduğunu ortaya koymuştur. Bu çalışma, Türkçe sağlık hizmetlerinde yapay zeka destekli sanal doktor asistanlarının potansiyelini göstermekte ve dile özgü modellerin geliştirilmesinin önemini vurgulamaktadır.

**Anahtar Kelimeler:** doğal dil işleme, sağlık yapay zekası, Türkçe dil modelleri, LLM ince ayarı, sanal doktor asistanı

**ABSTRACT**

This study examines the performance of four different large language models (doctor-meta-llama-3-8b, doctor-LLama2-sambanovasystems-7b, doctor-Mistral-trendyol-7b, and doctor-llama-3-cosmos-8b) in Turkish health counseling. The models were fine-tuned on a custom dataset consisting of 321,179 patient-doctor question-answer pairs. The performance of the models was measured using synthetic metrics such as BLEU, and BERT Score, as well as Elo scoring and expert doctor evaluations. The results showed that the doctor-LLama2-sambanovasystems-7b model performed the best in terms of overall performance, while the doctor-Mistral-trendyol-7b model had the lowest rate of harmful responses. This study demonstrates the potential of AI-assisted virtual doctor assistants in Turkish healthcare services and emphasizes the importance of developing language-specific models.

**Keywords:** natural language processing, health artificial intelligence, Turkish language models, LLM fine-tuning, virtual doctor assistant

**GİRİŞ**

Son yıllarda yapay zeka ve doğal dil işleme alanındaki gelişmeler, Büyük Dil Modelleri (LLM) gibi teknolojilerin ortaya çıkışını hızlandırmıştır. Bu modeller, insan dilini anlama ve üretme yetenekleriyle dikkat çekerken, özellikle sağlık gibi uzmanlık gerektiren alanlarda kullanım potansiyelleri giderek artmaktadır. Sağlık hizmetlerinde doğru ve hızlı bilgi paylaşımının hayati önem taşıması, bu alanda LLM'lerin kullanımını daha da önemli kılmaktadır.Sağlık alanında LLM'lerin kullanımı, hasta-doktor iletişiminin geliştirilmesi, tıbbi bilgilerin daha anlaşılır hale getirilmesi ve sağlık profesyonellerine destek sağlanması gibi çeşitli avantajlar sunmaktadır. Ancak, bu teknolojinin sağlık gibi hassas bir alanda kullanılabilmesi için, modellerin özel olarak eğitilmeleri ve performanslarının detaylı bir şekilde değerlendirilmesi gerekmektedir.Bu çalışmada, dört farklı LLM modelinin (doktor-meta-llama-3-8b, doktor-LLama2-sambanovasystems-7b, doktor-Mistral-trendyol-7b ve doktor-llama-3-cosmos-8b) Türkçe sağlık verileri üzerindeki performansı incelenmiştir. Modellerin başarımı, BLEU, BERT Skor, ANOVA ve CRONBACH gibi çeşitli metrikler kullanılarak değerlendirilmiş, ayrıca uzman değerlendirmeleri ve yapay zeka hakemliğinde karşılaştırmalı analizler yapılmıştır.

*Araştırmanın Amacı ve Önemi*

Bu araştırmanın temel amacı, Türkçe sağlık verileri üzerinde dört farklı büyük dil modelinin (LLM) ince ayar performansını incelemek ve değerlendirmektir. Çalışmada kullanılan modeller şunlardır:

* doktor-meta-llama-3-8b
* doktor-LLama2-sambanovasystems-7b
* doktor-Mistral-trendyol-7b
* doktor-llama-3-cosmos-8b

Araştırmanın spesifik hedefleri şu şekildedir:

1. LLM'lerin hasta-doktor ilişkisindeki potansiyelini ölçmek. Bu modeller, tıbbi bilgileri hastaların anlayacağı bir dille açıklama, hasta sorularını cevaplama ve sağlık çalışanlarına yardımcı olma konularında kullanılabilir.
2. Genel amaçlı LLM'lerin kapladıkları devasa alanlar ve devasa enerji ve işlem gücü tüketimi gibi dezavantajlarının en alt seviyeye indirildiği duruma özel LLM üretmek.
3. Türkçeye özel ve hasta-doktor ilişkisine odaklanmış daha küçük ve verimli LLM'lerin elzem olduğunu göstermek. Bu tür özelleştirilmiş modeller, hem doğruluk, hem de boyut ve enerji tüketimi açısından daha avantajlıdır.
4. Yapılan ince ayar eğitimi sonucunda, LLM'lerin Türkçe sağlık verileri üzerindeki potansiyelini göstermek.
5. LLM'lerin performansını üç farklı yöntemle değerlendirmek:

* Büyük yapay zeka modellerinin hakemliğinde değerlendirme
* Sentetik testlerle değerlendirme (BLEU, BERT SCORE gibi metrikler kullanarak)
* Gerçek uzman doktorların değerlendirmesi

Araştırmanın önemi şu noktalarda ortaya çıkmaktadır:

* Sağlık hizmetlerinde doğru ve hızlı bilgi paylaşımının hayati önem taşıması
* Hasta-doktor iletişiminin geliştirilmesi ihtiyacı
* Tıbbi bilgilerin daha anlaşılır hale getirilmesi gerekliliği
* Sağlık profesyonellerine destek sağlanması ihtiyacı
* Türkçe sağlık alanında özelleşmiş yapay zeka çözümlerinin eksikliği

Bu araştırma, sağlık alanında Türkçe dil modellerinin potansiyelini ortaya koyarak, daha verimli ve güvenilir sistemlerin geliştirilmesine katkı sağlamayı hedeflemektedir.

*Literatür Özeti*

LLM'ler ve sağlık verileriyle uygulamaları, son yıllarda yapay zekanın popülerleşmesiyle beraber araştırmacıların ilgisini çeken bir alan olmuştur. Literatür taraması, aşağıdaki ana başlıklar altında incelenebilir:

* LLM'lerin sağlık verilerinde kullanımı ve performansı
* Hasta-doktor iletişiminde LLM'lerin rolü
* Tıbbi bilgi çıkarımı ve özetleme
* Klinik karar destek sistemleri
* Etik ve yasal konular
* Türkçe LLM'ler
* LLM'lerin Başarım Metrikleri

Önemli çalışmalardan bazıları:

* Peng ve arkadaşları (2019): LLM'lerin elektronik sağlık kayıtlarından klinik anlam çıkarma başarısını incelemiştir.
* Park ve arkadaşları (2020): LLM'lerin hasta sorularına cevap verme ve tıbbi bilgileri daha anlaşılır dille açıklama konusundaki başarılarını ölçmüştür.
* Chikhaoui ve arkadaşları (2022): Sağlık sektöründe yapay zeka uygulamalarının etik ve yasal zorluklarını incelemiştir.
* Uludoğan ve arkadaşları (2024): TURNA modelini geliştirerek Türkçe için hem metin anlama hem de metin oluşturma işlerinde kullanılabilen ilk birleşik dil modelini sunmuştur.
* Kesgin ve arkadaşları (2024): Tamamen Türkçe veriler kullanılarak eğitilen cosmosGPT modellerini geliştirmiştir.

Literatür taraması, LLM'lerin sağlık alanında önemli bir potansiyele sahip olduğunu, ancak güvenli ve etik kullanım için daha fazla araştırma gerektiğini göstermektedir.

*Çalışmanın Kapsamı ve Katkıları*

Bu araştırma, Türkçe sağlık verileri üzerinde dört farklı LLM modelinin performansını inceleyerek literatüre özgün katkılar sağlamayı hedeflemektedir. Çalışmanın başlıca katkıları şunlardır:

* Hasta-doktor iletişimi alanına özel LLM'ler oluşturmak
* Geliştirilen LLM'lerin kapsamlı karşılaştırması
* Özgün hasta-doktor iletişimi veri kümesi oluşturmak
* Çok yönlü değerlendirme yaklaşımı geliştirmek
* Enerji ve donanım verimliliğini artırmak

Çalışmanın kapsamı:

* 321.179 soru-cevap çiftinden oluşan veri kümesi
* Dört farklı LLM modelinin ince ayarı ve değerlendirilmesi
* Üç farklı değerlendirme yöntemi (sentetik testler, yapay zeka hakemliği, uzman değerlendirmesi)
* Türkçe sağlık metinleri üzerinde performans analizi
* Etik ve yasal çerçevenin değerlendirilmesi

Bu çalışma, Türkçe sağlık alanında LLM'lerin güvenilir ve verimli kullanımı için önemli bir temel oluşturmaktadır.

**MATERYAL VE YÖNTEM**

Bu çalışmada kullanılan veri kümesi, yöntemler ve süreçler detaylı olarak açıklanacaktır. Araştırmanın temelini oluşturan veri kümesi oluşturma süreci, veri temizleme ve derleme aşamaları bu bölümde incelenecektir.

*Veri Kümesi*

Çalışmada kullanılan veri kümesi, Hugging Face platformunda bulunan Patient Doctor Q&A TR 321179 adlı özel bir Türkçe hasta-doktor soru-cevap veri kümesidir. Bu veri kümesi, gerçek hasta-doktor verilerinden oluşan 321.179 soru-cevap çiftinden oluşmaktadır. Veri kümesinin içeriği, hastaların doktorlara sorduğu her türlü soruya karşılık ilgili doktorların verdiği cevapları içermektedir.

*Veri Toplama ve Ön İşleme*

Patient Doctor Q&A TR 321179 veri kümesi, dört farklı veri kümesinin birleştirilip sonrasında karıştırılması sonucunda oluşturulmuştur:

1. Patient Doctor Q&A TR 19583: iCliniq platformundaki gerçek hasta soruları ve doktor yanıtlarının İngilizceden Türkçeye çevrilmiş halidir. Çeviri işlemi GPT-3.5-turbo ile gerçekleştirilmiştir. Bu süreçte:

* Noktalama işaretleri düzeltilmiştir
* Büyük/küçük harf kullanımı standardize edilmiştir
* Medikal terimler korunarak çeviri yapılmıştır

1. Patient Doctor Q&A TR 95588: chat\_doctor veri setinin İngilizceden Türkçeye çevrilmiş halidir. Çeviri işlemi yine GPT-3.5-turbo ile gerçekleştirilmiş ve aynı düzenlemeler yapılmıştır:

* Noktalama işaretleri düzeltilmiştir
* Büyük/küçük harf kullanımı standardize edilmiştir
* Medikal terimler korunarak çeviri yapılmıştır

1. Patient Doctor Q&A TR 5695: doctor-id-qa veri setinin Endonezceden Türkçeye çevrilmiş halidir. Çeviri işlemi GPT-3.5-turbo ile gerçekleştirilmiş ve benzer düzenlemeler yapılmıştır:

* Noktalama işaretleri düzeltilmiştir
* Büyük/küçük harf kullanımı standardize edilmiştir
* Medikal terimler korunarak çeviri yapılmıştır

1. Patient Doctor Q&A TR 167732: doktorsitesi veri setinin temizlenmiş halidir. Türkçe olan bu veri setinde şu düzenlemeler yapılmıştır:

* Bağlam bağımlı cümleler kaldırılmıştır. Örneğin "Daha önce de belirttiğim gibi..." gibi önceki konuşmalara referans veren ifadeler çıkarılmıştır.
* Mahremiyete halel getirecek cümleler temizlenmiştir. Hasta ve doktorların kişisel bilgileri, özel hayata dair detaylar vs. kaldırılmıştır.
* Doktorların reklamları silinmiştir. Muayenehane adresi, çalışma saatleri, özel tanıtımlar gibi reklam içerikli metinler çıkarılmıştır.
* Çeşitli iletişim bilgileri (telefon, e-posta, adres vb.) kaldırılmıştır.

*Veri Birleştirme*

Dört farklı veri kümesinin birleştirilmesi sürecinde, veri setlerinin sadece soru ve cevap kolonları alınmıştır. Her bir veri setinden %90 eğitim ve %10 test verisi ayrılarak birleştirme işlemi gerçekleştirilmiştir. Bu süreç şu şekilde uygulanmıştır:

1. Patient Doctor Q&A TR 19583 veri setinden:

* Eğitim verisi: %90
* Test verisi: %10

1. Patient Doctor Q&A TR 167732 veri setinden:

* Eğitim verisi: %90
* Test verisi: %10

1. Patient Doctor Q&A TR 5695 veri setinden:

* Eğitim verisi: %90
* Test verisi: %10

1. Patient Doctor Q&A TR 95588 veri setinden:

* Eğitim verisi: %90
* Test verisi: %10

Birleştirme sonucunda:

* Toplam 321.179 soru-cevap çiftinden oluşan bir veri seti elde edilmiştir
* Veri setinin %90'ı eğitim, %10'u test verisi olarak ayrılmıştır
* Veriler karıştırılmış ve son kontroller yapılarak veri setinin bütünlüğü sağlanmıştır

Bu süreçte her bir veri setinden eşit oranda eğitim ve test verisi alınması, modelin farklı kaynaklardan gelen verileri dengeli bir şekilde öğrenmesini sağlamıştır.

*Kullanılan Modeller*

Çalışmada dört farklı büyük dil modeli kullanılmıştır:

*doktor-meta-llama-3-8b*

Meta AI tarafından geliştirilen Llama ailesinin üçüncü neslidir. 8 milyar parametre içermektedir ve haleflerine göre daha gelişmiş doğal dili anlama ve metin oluşturma özelliklerine sahiptir. Modelin 70 milyar parametreye sahip bir varyasyonu da bulunmaktadır ama o modeli çalıştıracak donanıma ulaşmak çok zor ve maliyetli olduğundan 8 milyar parametreli versiyonu kullanılmıştır. Meta-Llama-3-8B, kapsamlı bir yelpazedeki işlerde kullanılabilecek bir modeldir ve bilhassa Türkçe gibi düşük kaynaklı dillerde de iyi performans göstermektedir. Önemli özellikleri:

* 8 milyar parametre
* Çoklu dil desteği
* Geliştirilmiş bağlam anlama yeteneği
* Daha verimli hesaplama gereksinimleri

*doktor-LLama2-sambanovasystems-7b*

SambaNova Systems tarafından geliştirilen, Türkçe dili için ince ayar yapılarak eğitilmiş bir sohbet modelidir. Bu model, SambaLingo-Turkish-Base üzerine mesajlaşmaya uyumlu hale getirilmiş ve DPO yöntemiyle eğitilmiştir. SambaLingo-Turkish-Base modeli Llama-2-7b tabanlı bir modeldir. Önemli özellikleri:

* 7 milyar parametre
* Llama-2-7b modelinin Türkçe'ye adaptasyonu
* Türkçe dili için ince ayar
* Cultura-X veri kümesinin Türkçe bölümünden 42 milyar tokenla eğitimi
* İki aşamalı ince ayar süreci: SFT ve DPO

*doktor-Mistral-trendyol-7b*

Trendyol tarafından geliştirilen ve özellikle Türkçe dili üzerine ince ayar yapılmış bir LLM'dir. Önemli özellikleri:

* 7 milyar parametre
* Türkçe dili için ince ayar
* Gelişmiş Türkçe iletişim yetenekleri

*doktor-llama-3-cosmos-8b*

YTÜ Bilgisayar Mühendisliği COSMOS Araştırma Grubu tarafından geliştirilen, Türkçe dilinde 30GB veriyle ince ayar yapılmış bir LLM'dir. Bu modelin ince ayarı, Meta-Llama-3-8B modeli üzerinde yapılmıştır. Önemli özellikleri:

* 8 milyar parametre
* Türkçe dili için ince ayar
* 30GB Türkçe veri üzerinde eğitim

*İnce Ayar Süreci*

Son y

*Hiperparametreler*

1. Öğrenme Oranı (Learning Rate): 1×10^-4

Öğrenme oranı, modelin ağırlıklarının her adımda ne kadar güncelleneceğini belirleyen kritik bir hiperparametredir. Bu değer:

* Çok yüksek olursa model aşırı büyük adımlarla öğrenir ve optimal noktayı kaçırabilir
* Çok düşük olursa model çok yavaş öğrenir ve yerel minimumlara takılabilir
* 1×10^-4 değeri literatürdeki çalışmalar incelenerek optimal olarak seçilmiştir

1. Isınma Adımları (Warmup Steps): 2000

Eğitimin başlangıç aşamasında öğrenme oranını kademeli olarak artıran bir tekniktir:

* Eğitimin başında düşük öğrenme oranıyla başlayarak modelin ani değişimlerden korunmasını sağlar
* Gradyan patlaması problemini önler
* Model veri kümesine ve öğrenme işine kademeli olarak adapte olur

1. Eğitim Döngüsü Sayısı

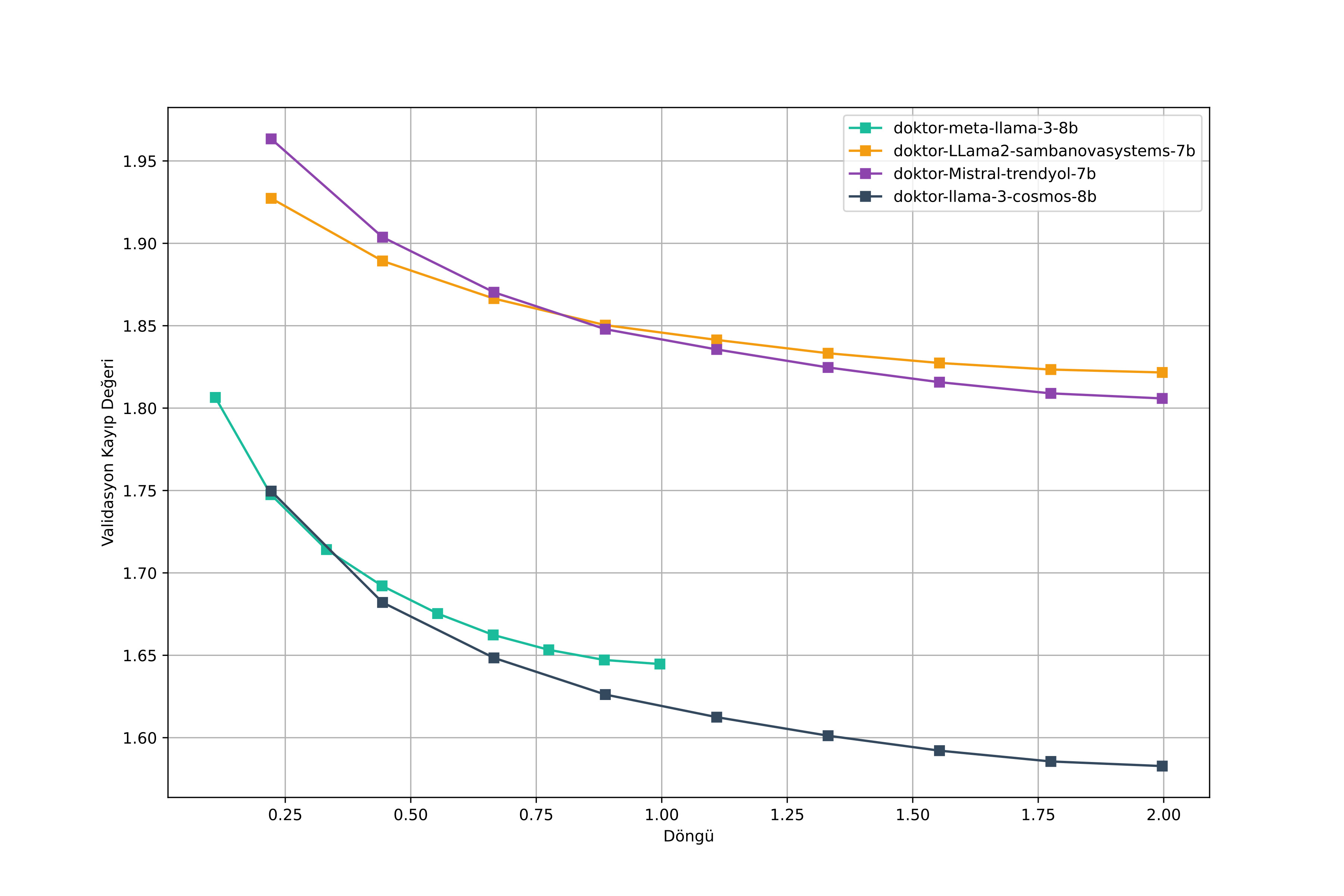
Meta-Llama-3-8B için: 1 döngü

Diğer modeller için: 2 döngü Her döngüde model tüm veri kümesini bir kez görür. Döngü sayısı:

* Aşırı öğrenmeyi önleyecek
* Yeterli öğrenmeyi sağlayacak

Bu değerlerin seçilmesinin sebepleri şunlardır:

* Şekil 1’de Meta-llama modeli (doktor-meta-llama-3-8b) 1 döngüde eğitimini tamamlarken, diğer modeller 2 döngü boyunca eğitilmiştir. Bunun sebebi Meta-llama modelinin validation loss değerinin 1. döngünün sonlarına doğru düz bir çizgi çizmeye başlamasıdır.
* Diğer modeller (doktor-LLama2-sambanovasystems-7b, doktor-Mistral-trendyol-7b ve doktor-llama-3-cosmos-8b) 2. döngünün ortalarına doğru düz bir çizgi çizmeye başlamıştır. Bu da eğitimin bu noktada optimum seviyeye ulaştığını gösterir.
* Başlangıçta tüm modellerin validation loss değerleri yüksek iken, eğitim ilerledikçe bu değerler düşmüştür. Bu düşüş modellerin öğrenme sürecinin başarılı olduğunu gösterir.



**Şekil 1.** Doğrulama Kayıp Eğrileri

Hesaplama kaynaklarını verimli kullanacak şekilde belirlenmiştir

1. Maksimum Sekans Uzunluğu

SambaLingo-Turkish-Chat için: 4096 token

Diğer modeller için: 8192 token Bu değer modelin tek seferde işleyebileceği maksimum token sayısını belirler:

* Uzun bağlamları anlayabilme yeteneğini etkiler
* Bellek kullanımını doğrudan etkiler

Her modelin kendi mimarisi göz önünde bulundurularak seçilmiştir

1. Öğrenme Oranı Planlayıcı Tipi: Doğrusal

Eğitim süresince öğrenme oranının nasıl değişeceğini kontrol eder:

* Başlangıçta daha yüksek öğrenme oranı
* Zamanla doğrusal olarak azalan öğrenme oranı
* Eğitimin sonlarına doğru daha hassas öğrenme

1. Ağırlık Azaltma (Weight Decay): 0.01

Modelin aşırı öğrenmesini engellemek için kullanılan regularizasyon tekniği:

* Model parametrelerinin büyüklüğünü kontrol eder
* Genelleme yeteneğini artırır
* Gradyan patlamasını önlemeye yardımcı olur

1. Optimizasyon Algoritması: 8-bit AdamW

Modelin ağırlıklarını güncellemek için kullanılan optimize edilmiş algoritma:

* Standart AdamW'nin 8-bitlik versiyonu
* Bellek kullanımını azaltır
* Eğitim hızını artırır
* Hassasiyet Formatı: BF16
* Modelin sayısal hesaplamalarında kullanılan format:
* FP32'ye göre daha az bellek kullanır
* FP16'ya göre daha geniş dinamik aralık sunar
* Eğitim stabilitesini korur

*Eğitim Stratejisi*

Eğitim stratejimiz, modellerin performansını artırmak ve Türkçe tıbbi metinleri daha iyi anlayabilmesini sağlamak üzere tasarlanmıştır:

1. LoRA Tekniği Kullanımı:

* Modelin tüm parametrelerini güncellemek yerine düşük boyutlu matrisler eklenerek eğitimi hızlandırılmıştır
* Bellek kullanımı minimize edilmiştir
* Eğitim süresi kısaltılmıştır

1. Unsloth Kütüphanesi Kullanımı:

* İnce ayar sürecindeki bellek kullanımını azaltmak için tercih edilmiştir
* Eğitim sürecini hızlandırmak amacıyla kullanılmıştır
* Kaynak verimliliği sağlanmıştır

1. Veri Kümesi Hazırlığı:

* Eğitim öncesinde veri seti, modelin giriş formatına uygun şekilde hazırlanmıştır
* Veri temizleme ve standardizasyon işlemleri yapılmıştır
* Veri kalitesi kontrol edilmiştir

1. Değerlendirme ve Kaydetme Stratejisi:

* Model performansı loss metriğiyle, 57.800 adımda bir değerlendirilmiştir
* Model, 10.000 adımda bir eğitim yapılan cihazın kapanma durumuna karşı kaydedilmiştir
* Düzenli kontrol noktaları oluşturulmuştur

1. Donanım ve Süre:

* Google Colab platformunda Tesla A100 40 GB GPU kullanılmıştır
* Modellerin eğitim süreleri:
* Cosmos LLama: 19,4 saat
* Meta LLama: 11,38 saat
* SambaLingo: 17,46 saat
* Trendyol: 19,3 saat

Bu stratejik yaklaşım, modellerin verimli ve etkili bir şekilde eğitilmesini sağlamıştır.

**BULGULAR**

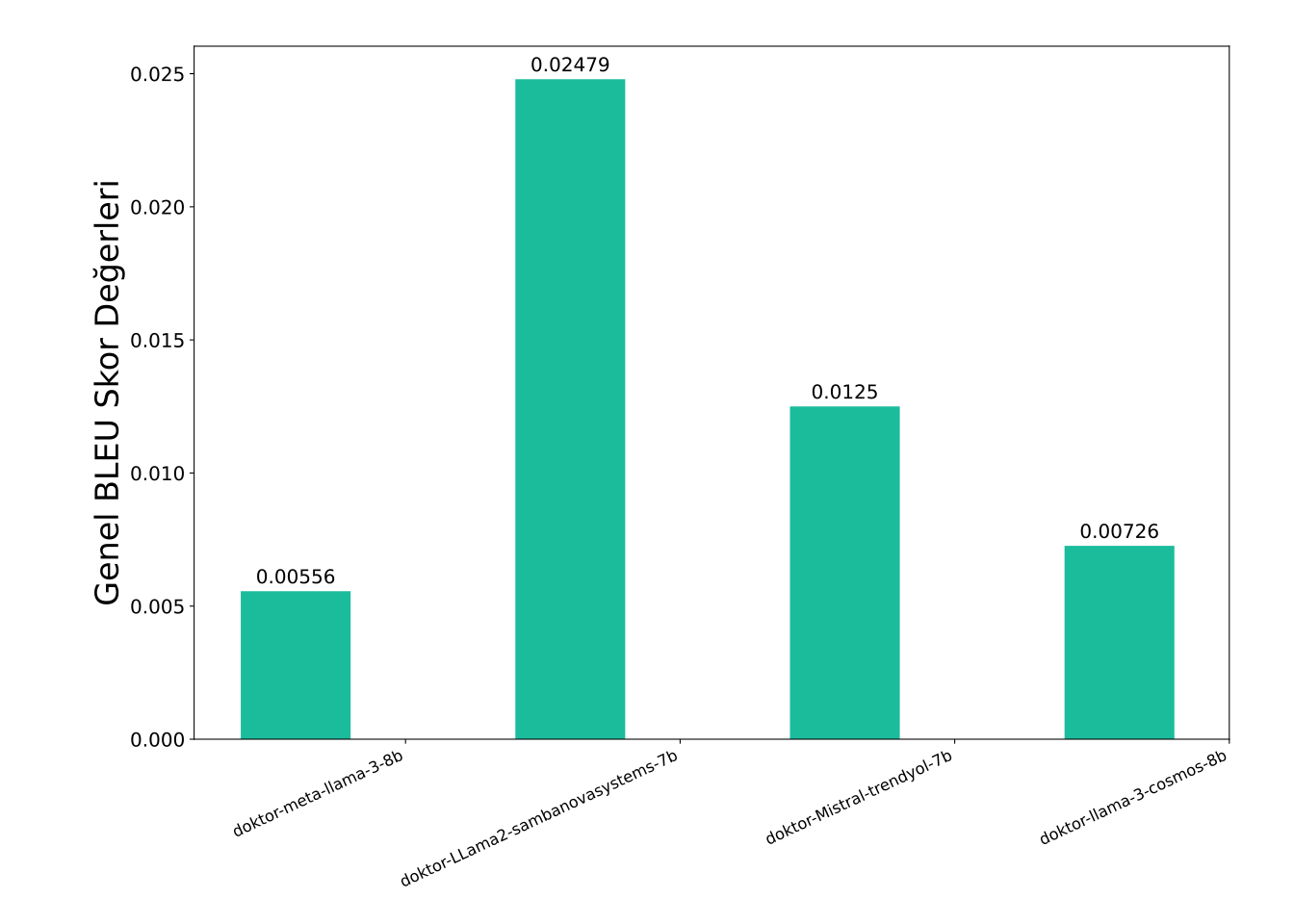
Bu bölümde, makale kapsamında gerçekleştirilen çalışmaların sonucunda elde edilen bulgular detaylı olarak sunulacaktır. Araştırmanın temel amacı doğrultusunda, geliştirilen modellerin performansları farklı açılardan değerlendirilmiş ve karşılaştırmalı analizler yapılmıştır. Modellerin başarımları sentetik değerlendirme ölçütleri, yapay zeka hakemliğinde değerlendirme sonuçları ve uzman değerlendirmesi sonuçları olmak üzere üç ana başlık altında incelenecektir.

*Modellerin Sentetik Sonuçlara Göre Analizi*

Bu bölümde, ince ayar yaptığımız dört LLM'in performansları sentetik ölçütler yardımıyla kıyaslanacaktır.

*BLEU Skor Sonuçları*

Şekil 2, dört farklı dil modelinin BLEU skor değerlerini karşılaştırmaktadır. BLEU skoru, modellerin ürettiği metinlerin referans metinlere ne kadar benzediğini ölçen önemli bir ölçüttür.



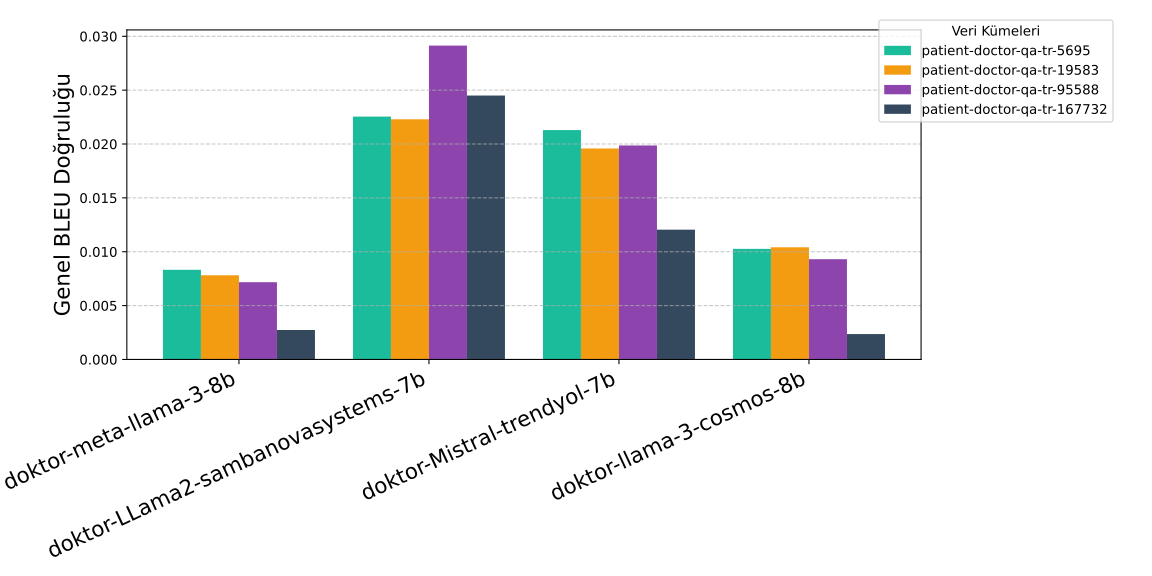
**Şekil 2.** Genel BLEU Skorları

Grafikteki sonuçlara ayrıntılı bir biçimde baktığımızda:

1. doktor-LLama2-sambanovasystems-7b modeli 0.02479 BLEU skoru ile en yüksek performansı göstermiştir. Bu sonuç, modelin ürettiği metinlerin referans metinlere en yakın olduğunu göstermektedir.
2. doktor-Mistral-trendyol-7b modeli 0.0125 BLEU skoru ile ikinci sırada yer almaktadır. Bu değer, birinci modelin yaklaşık yarısı kadardır.
3. doktor-llama-3-cosmos-8b modeli 0.00726 BLEU skoru ile üçüncü sırada yer almaktadır.
4. doktor-meta-llama-3-8b modeli 0.00556 BLEU skoru ile en düşük performansı göstermiştir.

Bu sonuçlar önemli çıkarımlar sunmaktadır:

* Parametre sayısının fazla olması her zaman daha iyi performans anlamına gelmemektedir. Mesela, 8B parametreye sahip modeller, 7B parametreli modellere göre daha düşük performans göstermiştir.
* Türkçe dil yapısına ve sağlık alanına özel eğitilmiş modeller (sambanovasystems ve trendyol) daha iyi performans sergilemiştir.
* BLEU skorlarının genel olarak düşük olması (1'e yakın değil), metin üretimi görevinin zorluğunu ve daha fazla iyileştirme potansiyeli olduğunu göstermektedir.



**Şekil 3.** Veri Kümesi Özelinde BLEU Skorları

Şekil 3, dört farklı modelin farklı veri kümeleri üzerindeki BLEU doğruluk skorlarını göstermektedir. Veri kümeleri patient-doctor-qa-tr-5695, patient-doctor-qa-tr-19583, patient-doctor-qa-tr-95588 ve patient-doctor-qa-tr-167732'dir.Modellerin veri kümelerine göre performansları şu şekildedir:

1. doktor-LLama2-sambanovasystems-7b:

* En yüksek performansı patient-doctor-qa-tr-95588 veri kümesinde göstermiştir (yaklaşık 0.029)
* Diğer veri kümelerinde de tutarlı bir şekilde 0.020-0.025 aralığında başarı göstermiştir
* Tüm veri kümelerinde diğer modellere göre daha iyi performans sergilemiştir

1. doktor-Mistral-trendyol-7b:

* En iyi performansı patient-doctor-qa-tr-5695 veri kümesinde göstermiştir (yaklaşık 0.021)
* Tüm veri kümelerinde 0.012-0.021 aralığında tutarlı bir başarı göstermiştir
* Genel olarak ikinci en iyi performansa sahip modeldir

1. doktor-meta-llama-3-8b:

* Veri kümeleri arasında 0.003-0.008 aralığında nispeten düşük ve tutarlı bir performans göstermiştir
* En düşük performansı patient-doctor-qa-tr-167732'de sergilemiştir

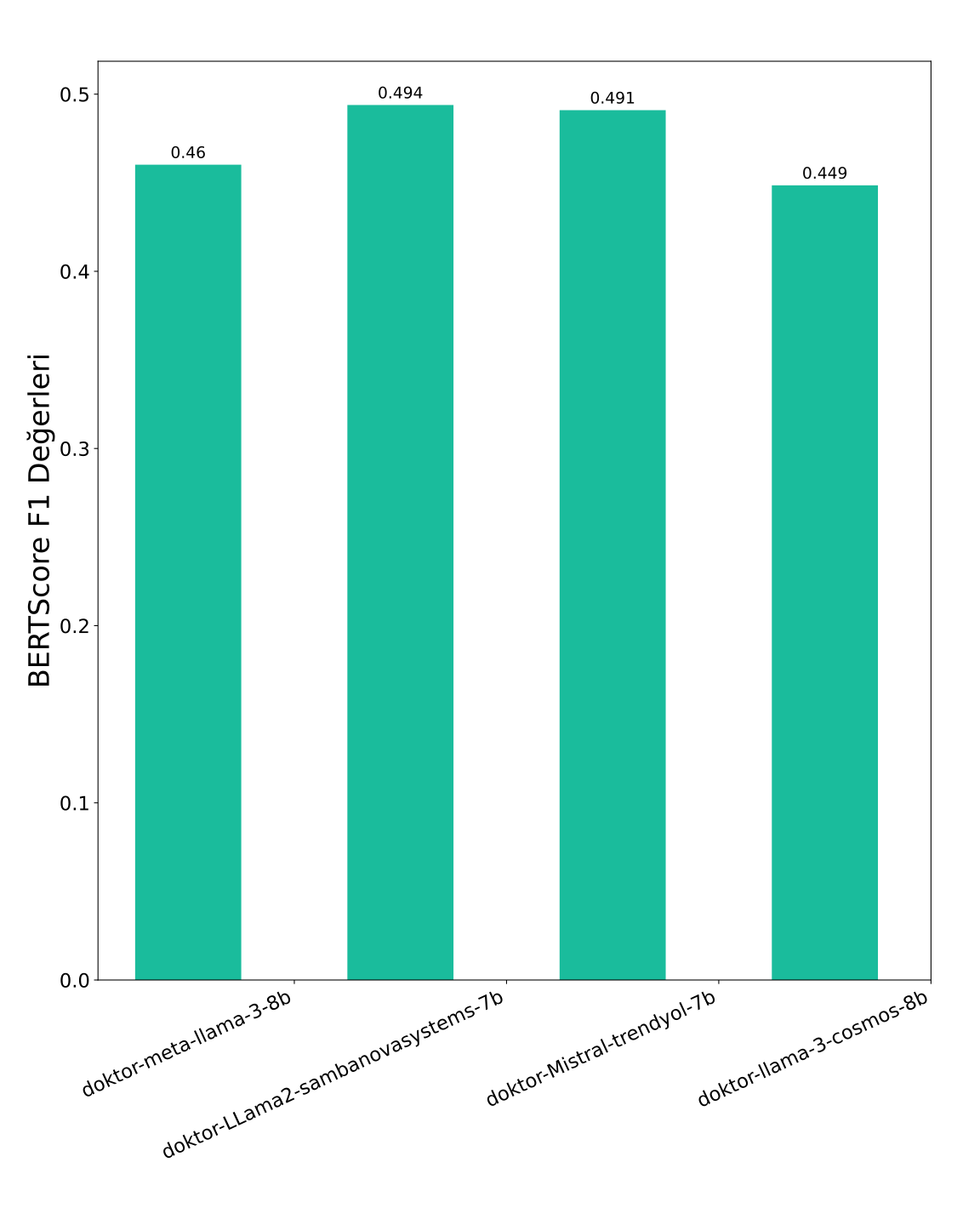
1. doktor-llama-3-cosmos-8b:

* Tüm veri kümelerinde 0.002-0.010 aralığında değişen en düşük performansı göstermiştir
* En iyi sonucu patient-doctor-qa-tr-5695'te elde etmiştir

Genel değerlendirme:

* 7b parametreli modeller (doktor-LLama2-sambanovasystems-7b ve doktor-Mistral-trendyol-7b), 8b parametreli modellere göre daha iyi performans göstermiştir
* Veri kümesinin büyüklüğü ile modellerin performansı arasında doğrudan bir ilişki gözlenmemiştir
* Türkçe'ye özel eğitilmiş modeller (sambanovasystems ve trendyol) daha iyi performans sergilemiştir

*BERTScore Sonuçları*



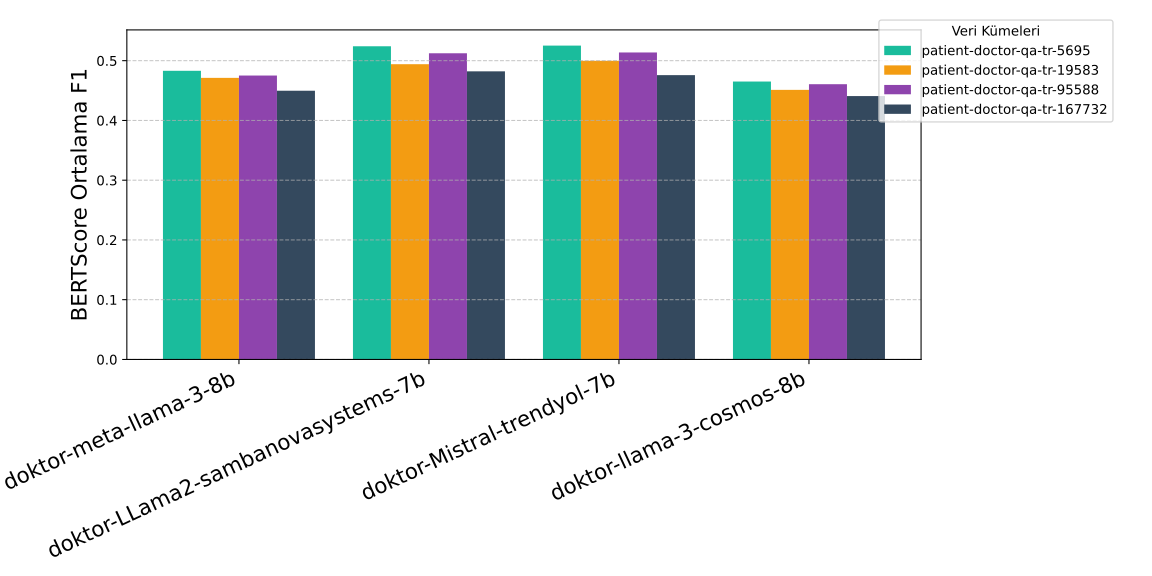
**Şekil 4.** BertScore F1 Skorları

Şekil 3’de dört farklı dil modelinin BERT Skor F1 değerlerini karşılaştırmaktadır. BERT Skor, modellerin ürettiği metinlerin referans metinlere anlamsal benzerliğini ölçen önemli bir metriktir. Grafikteki sonuçları detaylı olarak analiz edelim:

1. doktor-LLama2-sambanovasystems-7b modeli 0.494 F1 skoru ile en yüksek performansı göstermiştir. Bu sonuç, modelin ürettiği metinlerin anlamsal olarak referans metinlere en yakın olduğunu göstermektedir.
2. doktor-Mistral-trendyol-7b modeli 0.491 F1 skoru ile ikinci sırada yer almaktadır. Bu skor, birinci modele oldukça yakındır ve modelin de başarılı sonuçlar ürettiğini gösterir.
3. doktor-meta-llama-3-8b modeli 0.46 F1 skoru ile üçüncü sırada yer almaktadır.
4. doktor-llama-3-cosmos-8b modeli 0.449 F1 skoru ile son sırada yer almaktadır.

Bu sonuçlar önemli çıkarımlar sunmaktadır:

* İlk iki model arasındaki farkın az olması (0.003), her iki modelin de Türkçe sağlık metinlerini anlamsal olarak başarılı bir şekilde işleyebildiğini göstermektedir.
* 8B parametreli modeller (doktor-meta-llama-3-8b ve doktor-llama-3-cosmos-8b), 7B parametreli modellere göre daha düşük performans göstermiştir. Bu durum, parametre sayısının tek başına başarıyı belirlemediğini göstermektedir.
* En yüksek ve en düşük skorlar arasındaki fark (0.045) göreceli olarak küçüktür, bu da tüm modellerin makul bir seviyede performans gösterdiğine işaret etmektedir.



**Şekil 5.** Veri Kümesi Özelinde BertScore F1 Skorları

Şekil 5, dört farklı modelin farklı veri kümeleri üzerindeki BERT Score F1 performansını göstermektedir. Her model için dört farklı veri kümesi (patient-doctor-qa-tr-5695, patient-doctor-qa-tr-19583, patient-doctor-qa-tr-95588 ve patient-doctor-qa-tr-167732) üzerindeki başarımları karşılaştırılmıştır. Modellerin veri kümelerine göre performansları şu şekildedir:

1. doktor-LLama2-sambanovasystems-7b:

* En yüksek performansı patient-doctor-qa-tr-5695 veri kümesinde göstermiştir (yaklaşık 0.52)
* Diğer veri kümelerinde de tutarlı bir şekilde 0.48-0.51 aralığında başarı göstermiştir
* Çoğu veri kümesinde diğer modellere göre daha iyi performans sergilemiştir

1. doktor-Mistral-trendyol-7b:

* En iyi performansı patient-doctor-qa-tr-5695 veri kümesinde göstermiştir (yaklaşık 0.53)
* Tüm veri kümelerinde 0.475-0.525 aralığında tutarlı bir başarı göstermiştir
* Genel olarak ikinci en iyi performansa sahip modeldir

1. doktor-meta-llama-3-8b:

* Veri kümeleri arasında 0.449-0.48 aralığında nispeten düşük ve tutarlı bir performans göstermiştir
* En düşük performansı patient-doctor-qa-tr-167732'de sergilemiştir

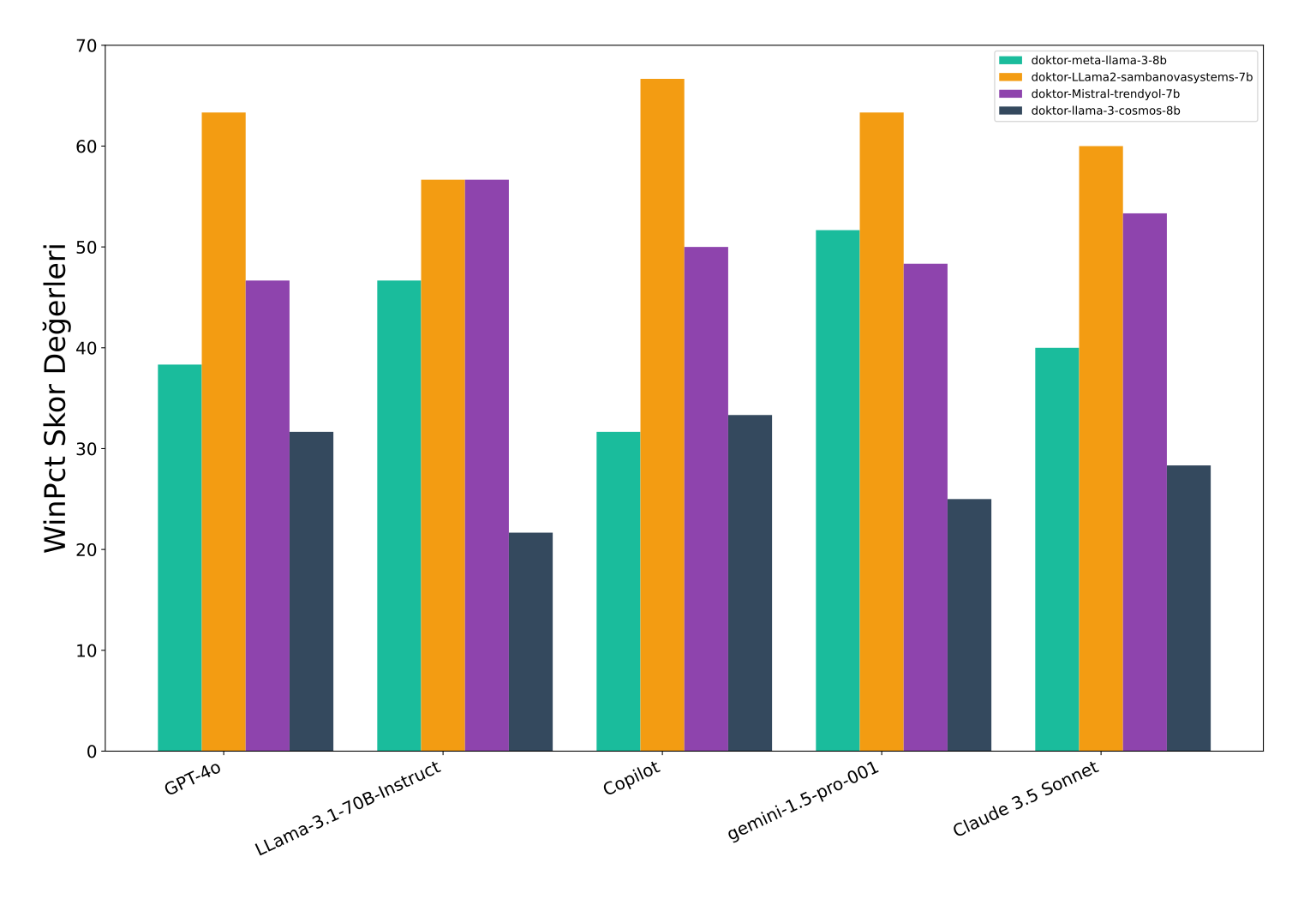
1. doktor-llama-3-cosmos-8b:

* Tüm veri kümelerinde 0.44-0.465 aralığında değişen en düşük performansı göstermiştir
* En iyi sonucu patient-doctor-qa-tr-5695'te elde etmiştir

Genel değerlendirme:

* 7b parametreli modeller (doktor-LLama2-sambanovasystems-7b ve doktor-Mistral-trendyol-7b), 8b parametreli modellere göre daha iyi performans göstermiştir
* Veri kümesinin büyüklüğü ile modellerin performansı arasında doğrudan bir ilişki gözlenmemiştir
* Türkçe'ye özel eğitilmiş modeller (sambanovasystems ve trendyol) daha iyi performans sergilemiştir

*Yapay Zeka Hakemliğinde Değerlendirme*



**Şekil 8.** Yapay Zeka Hakemliğinde Kazanma Yüzdeleri

Yapay zeka hakemliğinde, bizim eğittiğimiz modellerin her birine 20 soru sorulup cevapları daha gelişmiş yapay zeka modellerince değerlendirilmiştir. Bu değerlendirmenin sonuçları Şekil 8'de gösterilmiştir:

1. GPT-4o Hakemliği:

* doktor-LLama2-sambanovasystems-7b: %63.33 ile en yüksek
* doktor-Mistral-trendyol-7b: %46.67
* doktor-meta-llama-3-8b: %38.33
* doktor-llama-3-cosmos-8b: %31.67 ile en düşük

1. LLaMA 3.1 70B Hakemliği:

* doktor-LLama2-sambanovasystems-7b ve doktor-Mistral-trendyol-7b: %56.67 ile eşit ve en yüksek
* doktor-meta-llama-3-8b: %46.67
* doktor-llama-3-cosmos-8b: %21.67 ile en düşük

1. Microsoft Copilot Hakemliği:

* doktor-LLama2-sambanovasystems-7b: %66.67 ile en yüksek
* doktor-Mistral-trendyol-7b: %50.00
* doktor-llama-3-cosmos-8b: %33.33
* doktor-meta-llama-3-8b: %31.67 ile en düşük

1. Gemini 1.5 Pro Hakemliği:

* doktor-LLama2-sambanovasystems-7b: %63.33 ile en yüksek
* doktor-meta-llama-3-8b: %51.67
* doktor-Mistral-trendyol-7b: %48.33
* doktor-llama-3-cosmos-8b: %25.00 ile en düşük

1. Claude 3.5 Sonnet Hakemliği:

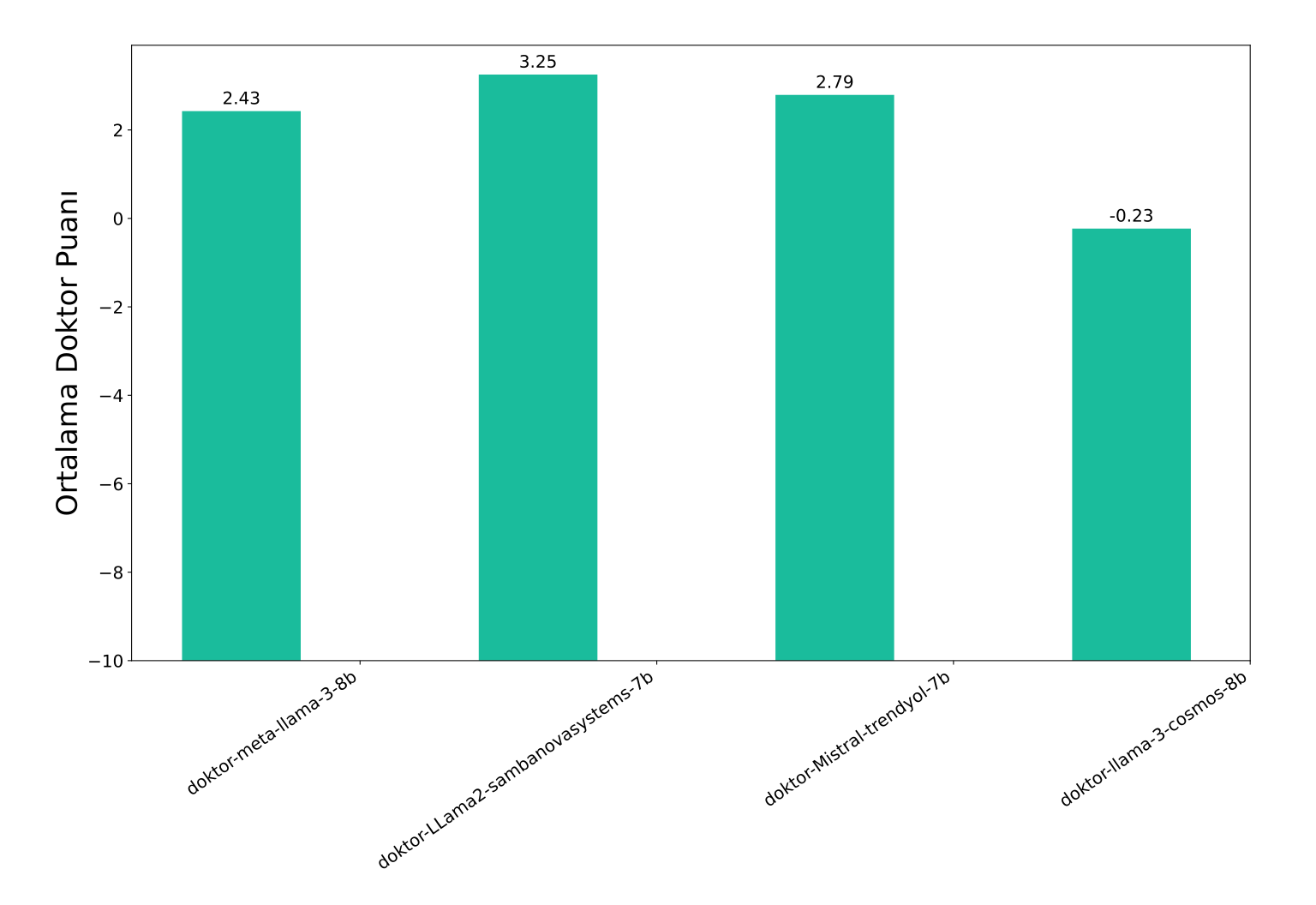
* doktor-LLama2-sambanovasystems-7b: %60.00 ile en yüksek
* doktor-Mistral-trendyol-7b: %53.33
* doktor-meta-llama-3-8b: %40.00
* doktor-llama-3-cosmos-8b: %28.33 ile en düşük

Genel olarak:

* doktor-LLama2-sambanovasystems-7b modeli tüm hakemlerde en yüksek veya en yükseklerden birine sahip
* doktor-llama-3-cosmos-8b modeli genellikle en düşük performansı göstermiş
* doktor-Mistral-trendyol-7b ve doktor-meta-llama-3-8b modelleri orta sıralarda yer almış
* Hakem modeller arasında tutarlı bir değerlendirme görülmekte

*Uzman Değerlendirmesi Sonuçları*

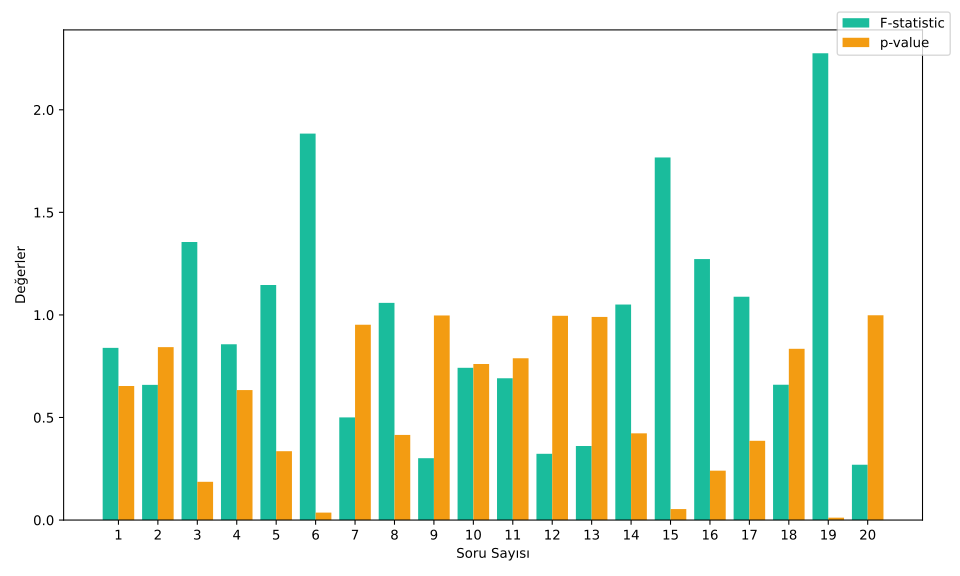
Uzman değerlendirmesi kapsamında, farklı uzmanlık alanlarından 20 doktor tarafından modellerin performansı değerlendirilmiştir. Değerlendirme sürecinde, rastgele seçilen 20 hasta sorusuna modellerin verdiği yanıtlar, -10 ile +10 arasında puanlanmıştır. Burada -10 yanıtın çok zararlı ve yanlış yönlendirici olduğunu, +10 ise yanıtın çok faydalı ve eksiksiz olduğunu göstermektedir. 0 puan ise yanıtın ne faydalı ne de zararlı olduğunu ifade etmektedir.



**Şekil 9.** Uzmanlar Tarafından Değerlendirme Sonuçları

Şekil 9'da görüldüğü üzere, SambaLingo-Turkish-Chat modeli 3.25 ortalama puanla en yüksek değerlendirmeyi almıştır. Bu modeli 2.79 puanla Trendyol-LLM-7b-chat-v1.8 ve 2.43 puanla Meta-Llama-3-8B modeli takip etmektedir. Turkish-Llama-8b-v0.1 modeli ise -0.23 puanla en düşük değerlendirmeyi almıştır.

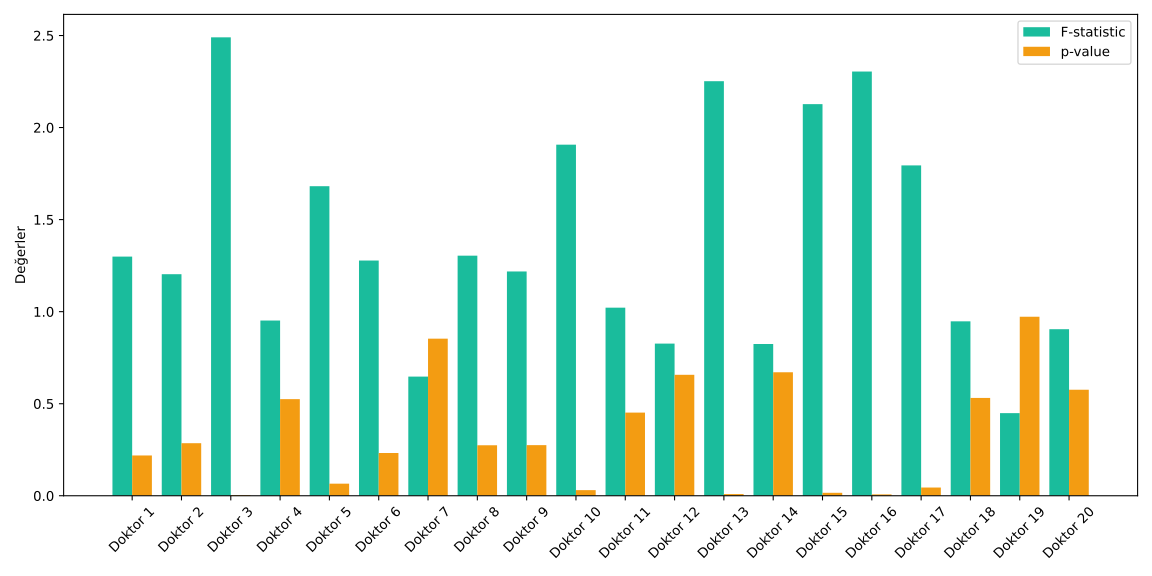
*ANOVA Sonuçları*



**Şekil 4.** Soruların ANOVA Skorları

Şekil 4’de farklı doktorun değerlendirmeleri görülmektedir. F-istatistik değerleri (yeşil çubuklar) bazı doktorlar için oldukça yüksektir. Özellikle Doktor 3, Doktor 13 ve Doktor 16'nın F-istatistik değerleri 2.0'ın üzerindedir. Bu, bu doktorların değerlendirmeleri arasında istatistiksel olarak anlamlı farklılıklar olduğunu gösterir.

P-değerlerine (turuncu çubuklar) baktığımızda, çoğu durumda 0.05'in altında olduğunu görüyoruz. Bu da değerlendirmeler arasındaki farklılıkların istatistiksel olarak anlamlı olduğunu doğrulamaktadır. Özellikle Doktor 10, 15 ve 16'nın p-değerleri çok düşüktür.

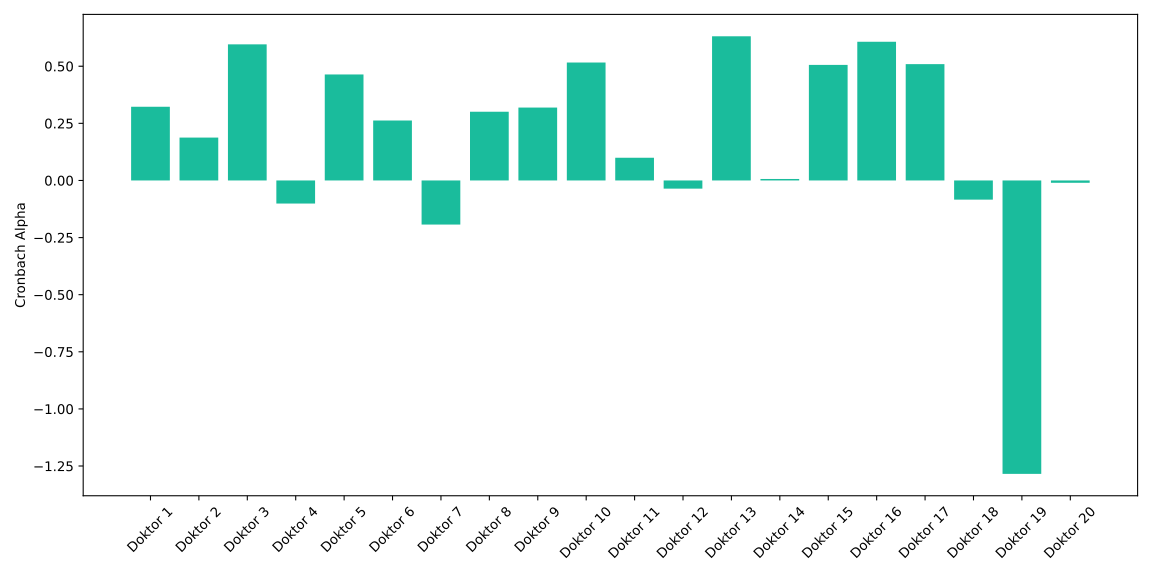


**Şekil 5.** Doktorların ANOVA Skorları

Şekil 5’te ise soru bazında ANOVA sonuçları gösterilmektedir. Soru 19'un en yüksek F-istatistik değerine sahip olduğu görülmektedir (yaklaşık 2.2). Bu, bu soruya verilen cevaplar arasında önemli farklılıklar olduğunu gösterir.Bazı sorularda (örneğin 12, 13 ve 20) p-değerleri 0.05'in üzerindedir. Bu sorularda değerlendirmeler arasındaki farklılıklar istatistiksel olarak anlamlı değildir.

Bu sonuçlar, hem doktorlar arasında hem de sorular arasında değerlendirme farklılıkları olduğunu göstermektedir. Bu farklılıklar bazı durumlarda istatistiksel olarak anlamlıdır. Bu da değerlendirmelerin subjektif olabileceğini ve doktorların farklı kriterlere göre değerlendirme yapabileceğini gösterir.

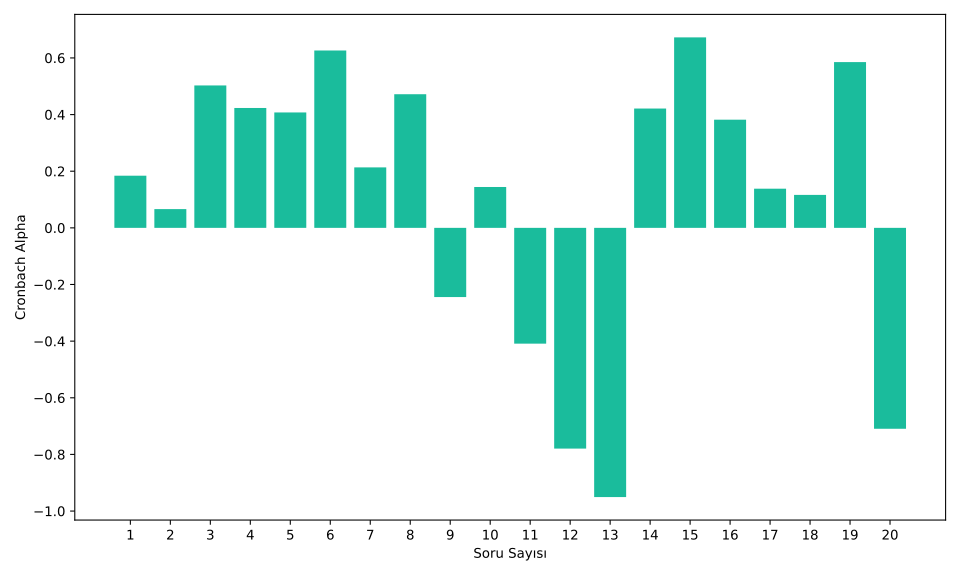
*CRONBACH Sonuçları*



**Şekil 6.** Doktorların CRONBACH Skorları

Şekil 6’da, 20 farklı doktorun değerlendirmeleri görülmektedir. Doktor bazında Cronbach Alfa değerleri -1.25 ile +0.75 arasında değişmektedir:

* En yüksek tutarlılık Doktor 3'te görülmektedir (yaklaşık 0.6)
* En düşük tutarlılık Doktor 19'da görülmektedir (yaklaşık -1.25)
* Doktorların çoğunluğu 0 ile 0.5 arasında değerler almıştır, bu da orta düzeyde bir tutarlılığa işaret eder



**Şekil 7.** Soruların CRONBACH Skorları

Şekil 7’de ise soru bazında Cronbach Alfa değerleri gösterilmektedir:

* En yüksek tutarlılık 15. soruda görülmektedir (yaklaşık 0.65)
* En düşük tutarlılık 13. soruda görülmektedir (yaklaşık -0.95)
* Soruların yaklaşık yarısı pozitif, yarısı negatif değerler almıştır

Bu sonuçlar şunu göstermektedir:

* Değerlendirmelerde doktor ve soru bazında önemli varyasyonlar vardır
* Bazı doktorlar ve sorular için tutarlılık oldukça düşüktür
* Genel olarak değerlendirmelerde orta düzeyde bir tutarlılık gözlenmektedir

Bu bulgular, değerlendirme sürecinin standardizasyonunun artırılması ve değerlendiriciler arası tutarlılığın iyileştirilmesi gerektiğini göstermektedir.

**TARTIŞMA VE SONUÇ**

Bu bölümde, çalışmamızın bulguları kapsamlı bir şekilde tartışılacak ve elde edilen sonuçlar değerlendirilecektir.

*Bulguların Değerlendirilmesi*

SÇalışmamızda dört farklı büyük dil modelinin Türkçe sağlık verileri üzerindeki performansı incelenmiştir. Elde edilen bulgular şu şekilde özetlenebilir:

* SambaLingo-Turkish-Chat modeli, BLEU, ROUGE, METEOR ve BERT Score gibi metriklerde en yüksek başarımı göstermiş, uzman değerlendirmelerinde ve Elo puanlamasında birinci sırada yer almıştır.
* Trendyol-LLM-7b-chat-v1.8 modeli ikinci en iyi performansı sergilemiş, CER ve WER ölçütlerinde en düşük hata oranlarını elde etmiş ve en düşük zararlı cevap oranına sahip olmuştur.
* Meta-Llama-3-8B modeli, genel amaçlı bir model olmasına rağmen makul bir performans göstermiş, çoğu ölçütte üçüncü sırada yer almıştır.
* Turkish-Llama-8b-v0.1 modeli ise çoğu değerlendirme ölçütünde en düşük performansı sergilemiştir.

*Kısıtlamalar ve Öneriler*

Çalışmamızın bazı kısıtlamaları bulunmaktadır:

* Veri kümesi boyutu ve çeşitliliği sınırlıdır
* Hesaplama kaynakları kısıtlıdır
* Model sayısı ve çeşitliliği artırılabilir
* Değerlendirme metrikleri genişletilebilir

Bu kısıtlamaların aşılması için önerilerimiz:

* Daha geniş ve çeşitli veri kümeleri oluşturulması
* Farklı model mimarilerinin test edilmesi
* Daha kapsamlı değerlendirme metriklerinin geliştirilmesi
* Etik ve hukuki çerçevenin güçlendirilmesi

*Gelecek Çalışmalar*

Gelecek çalışmalar için önerilerimiz:

* Multimodal modellerin sağlık alanında kullanımının araştırılması
* Daha büyük Türkçe sağlık veri kümelerinin oluşturulması
* Model mimarilerinin Türkçe dil yapısına göre optimize edilmesi
* Veri güvenliği ve gizliliğine yönelik çözümlerin geliştirilmesi
* Sağlık Bakanlığı gibi kurumlarla işbirliklerinin artırılması

*Sonuç*

Bu çalışma, LLM'lerin Türkçe sağlık alanında önemli bir potansiyele sahip olduğunu göstermektedir. Özellikle dile özgü eğitilmiş modellerin daha başarılı olduğu görülmüştür. Ancak bu teknolojinin güvenli ve etik kullanımı için daha fazla araştırma gerekmektedir. Sağlık gibi kritik bir alanda yapay zeka kullanımının hem fırsatları hem de riskleri dikkatle değerlendirilmelidir.