**TÜRKÇE SAĞLIK DANIŞMANLIĞINDA BÜYÜK DİL MODELLERİNİN HASTA-DOKTOR İLETİŞİMİNDE KULLANIM POTANSİYELİ**

**THE POTENTIAL USE OF LARGE LANGUAGE MODELS IN PATIENT-DOCTOR COMMUNICATION IN TURKISH HEALTH CONSULTATION**

*Muhammed Kayra BULUT1* (ORCID: 0009-0000-3107-7121)

*Banu DİRİ,2\** (ORCID: 0000-\*\*\*\*-\*\*\*\*-\*\*\*\*)

1 Yıldız Teknik Üniversitesi, Bilgisayar Mühendisliği Bölümü, İstanbul, Türkiye

\*Sorumlu Yazar / Corresponding Author: Muhammed Kayra BULUT, kayrabulut39@gmail.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 Skor 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 BERTScore, 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 zamanlarda yapay zeka ve doğal dil işleme alanındaki ilerlemeler, Büyük Dil Modelleri (LLM) gibi teknolojilerin ortaya çıkmasını ve gelişimini hızlandırmıştır (Vaswani, 2017; Brown, 2020). Bu modeller, insan dilini anlama ve doğal dil oluşturma özellikleriyle öne çıkarken, bilhassa sağlık gibi ileri uzmanlığa ihtiyaç duyulan alanlarda kullanım potansiyelleri oldukça fazladır (Park ve diğ., 2020). Sağlık hizmetlerinde doğru ve hızlı bilgi paylaşımının hayati öneme hâiz olması, bu alanda LLM'lerin kullanımını daha da önemli hale getirmektedir (Jiang ve diğ., 2017).

Sağlık alanında LLM'lerin kullanımı, hasta-doktor iletişiminin kolaylaştırılması, tıbbi bilgilerin daha anlaşılır şekle getirilmesi ve sağlık uzmanlarına destek sağlanması gibi çeşitli avantajlar vadetmektedir (Chikhaoui ve diğ., 2022). Lâkin, bu teknolojinin sağlık gibi hayatî bir alanda kullanılabilmesi için, modellerin özel olarak eğitilmeleri ve performanslarının ayrıntılı bir biçimde değerlendirilmesi lâzımdır (Peng ve diğ., 2019).

Bu çalışmada, dört farklı LLM’in (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 performans miktarı, BLEU, BERT Skor, ANOVA ve CRONBACH gibi birçok ölçüt kullanılarak ölçülmüş (Zhang ve diğ., 2019), bunun yanında uzman değerlendirmeleri ve yapay zeka hakemliğinde kıyaslamalı analizler yapılmıştır (Chiang ve diğ., 2024).

*Araştırmanın Amacı ve Önemi*

Bu araştırmanın temel amacı, Türkçe sağlık verileri üzerinde Turkish-Llama-8b-v0.1 (ytu-ce-cosmos, 2024), Meta-Llama-3-8B (meta-llama, 2024), SambaLingo-Turkish-Chat (Sambanovasystems, 2024) ve Trendyol-LLM-7b-chat-v1.8 (Trendyol, 2024) modellerinin ince ayar yapılmasıyla oluşturulan doktor- doktor-llama-3-cosmos-8b , meta-llama-3-8b, doktor-LLama2-sambanovasystems-7b ve doktor-Mistral-trendyol-7b büyük dil modellerinin ince ayar performansını incelemek ve kıyaslamaktır.

LLM’lerin hasta-doktor ilişkisindeki potansiyelini ölçmek, tıbbi bilgileri hastaların daha kolay anlayacağı bir dille açıklama, hasta sorularını cevaplama ve sağlık çalışanlarına yardımcı olma hususlarındaki kullanım potansiyellerini ölçmek hedeflenmektedir. Genel amaçlı LLM’lerin kapladıkları devasa alanlar ve aynı zamanda devasa enerji ve işlem gücü tüketimi gibi dezavantajlarının en alt seviyeye indirildiği, belli bir işe özel LLM oluşturmak hedeflenmektedir (Chikhaoui ve diğ., 2022). Bununla beraber, Türkçeye özel ve hasta-doktor iletişimine odaklanmış daha ufak ve verimli LLM’lerin elzem olduğu ortaya konulmak istenmektedir (E. Cengiz, Altinuç & Temizel, 2022). Bu özelleştirilmiş modellerin hem doğruluk hem de boyut ve enerji tüketimi açısından daha avantajlı olduğu vurgulanacaktır.

Yapılan ince ayar eğitimi sonucunda, LLM’lerin Türkçe sağlık verileri üzerindeki potansiyelini göstermek amaçlanmaktadır. Bu değerlendirme süreci üç farklı ana yöntemle gerçekleştirilecektir: büyük yapay zeka modellerinin hakemliğinde değerlendirme (Claude, GPT-4, vb.), sentetik testlerle değerlendirme (örneğin BLEU, BERT Skor) ve gerçek uzman doktorların değerlendirmesi yoluyla kapsamlı bir analiz yapılması planlanmaktadır.

Sağlık hizmetlerinde doğru ve hızlı bilgi paylaşımının hayati önem taşıması, hasta-doktor iletişiminin geliştirilmesi ihtiyacı ve tıbbi bilgilerin daha anlaşılır hale getirilmesi gerekliliği bu araştırmanın temel odak noktalarını oluşturmaktadır. Bununla beraber, sağlık uzmanlarına destek vermesi ihtiyacı ve Türkçe sağlık alanında özelleşmiş yapay zeka çözümlerinin eksikliği de araştırmanın öneminini ortaya koyan diğer unsurlardır. Bu araştırma, sağlık alanında Türkçe dil modellerinin potansiyelini ortaya koyarak, daha verimli ve güvenilir sistemlerin geliştirilmesine önemli bir katkı vermeyi hedeflemektedir. Bilhassa Türkçe sağlık literatüründe yapay zeka destekli çözümlerin eksikliği göz önüne alındığında, bu çalışmanın alandaki önemli bir boşluğu dolduracağı öngörülmektedir.

Bu araştırmanın sonuçları, gelecekte Türkiye’deki sağlık hizmetlerinin dijital dönüşümüne katkı sağlayacak, hasta-doktor iletişimini güçlendirecek ve sağlık bilgilerinin daha etkili bir şekilde paylaşılmasını mümkün kılacak sistemlerin geliştirilmesine temel oluşturacaktır.

*Literatür Özeti*

LLM’ler ve sağlık verileriyle ilgili uygulamalar, yapay zekanın son yıllardaki popülerleşmesiyle birlikte araştırmacıların yoğun ilgisini çeken önemli bir alan haline gelmiştir. Bu alanda LLM’lerin sağlık verilerinde kullanımı ve performansı (Peng ve diğ., 2019), hasta-doktor iletişimindeki rolü (Park ve diğ., 2020), tıbbi bilgi çıkarımı ve özetleme, klinik karar destek sistemleri, etik ve yasal konular, Türkçe LLM’ler (E. Cengiz, Altinuç & Temizel, 2022) ve LLM’lerin başarım ölçütleri gibi önemli konular ön plana çıkmaktadır.

Yapılan önemli araştırmalar arasında, Peng ve arkadaşlarının 2019’da elektronik sağlık kayıtlarından klinik anlam çıkarma başarısını inceleyen çalışması (Peng ve diğ., 2019), Park ve arkadaşlarının 2020’de LLM’lerin hasta sorularını yanıtlama ve tıbbi bilgileri sarih şekilde açıklama konusundaki başarılarını ölçen araştırması (Park ve diğ., 2020) ve Chikhaoui ve arkadaşlarının 2022’de sağlık sektöründeki yapay zeka uygulamalarının etik ve yasal sıkıntılarını inceleyen çalışması (Chikhaoui ve diğ., 2022) bulunmaktadır. Türkçe dil modelleri alanında ise Kesgin ve arkadaşlarının 2024’te tamamen Türkçe verilerle eğittikleri cosmosGPT benzeri modeller (Kesgin ve diğ., 2024) dikkat çekmektedir.

Literatürden elde edilen bulgular, LLM’lerin sağlık alanında çok önemli bir potansiyele sahip olduğunu ortaya koymakla beraber, bu teknolojilerin güvenli ve etik kullanımı için daha fazla araştırma yapılmasının lazım olduğunu göstermiştir.

*Çalışmanın Kapsamı ve Katkıları*

Bu araştırma, Türkçe sağlık verileriyle dört farklı LLM’in performansını inceleyerek literatüre özgün katkılar sağlamayı amaçlamaktadır. Çalışmanın başlıca faydaları hasta-doktor iletişimi alanına özel LLM'ler oluşturmak, geliştirilen LLM'lerin kapsamlı kıyaslamasını yapmak, özgün hasta-doktor iletişimi veri kümesi oluşturmak, çok yönlü değerlendirme yaklaşımı geliştirmek ve enerji ve donanım verimliliğini artırmak olarak belirlenmiştir.

Çalışmanın kapsamında 321.179 soru-cevap çiftinden oluşan veri kümesi, dört farklı LLM’in ince ayarı ve değerlendirilmesi, sentetik testler, yapay zeka hakemliği ve uzman değerlendirmesi olmak üzere üç ayrı ölçüm yöntemi, Türkçe sağlık metinleri üzerinde performans analizi ve etik ve yasal çerçevenin değerlendirilmesi yer almaktadır. Bu kapsamlı çalışma, Türkçe sağlık alanında LLM'lerin güvenilir ve verimli kullanımı için önemli bir temel oluşturmakta ve sağlık iletişiminde yapay zeka teknolojilerinin gelişimine katkı sağlamaktadır.

# 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 (Bulut, 2024a). 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 setinin birleştirilip karıştırılmasıyla oluşturulmuştur (Bulut, 2024a). Bu veri kümelerinden ilki olan Patient Doctor Q&A TR 19583 (Bulut, 2024d), iCliniq platformundaki gerçek hasta soruları ve doktor yanıtlarının (Henry41, 2024) İngilizceden Türkçeye GPT-3.5-turbo ile çevrilmiş halidir. Bu çeviri yapılırken aynı zamanda noktalama işaretleri düzeltilmiş, büyük/küçük harf kullanımı standardize edilmiş ve medikal terimler korunmuştur.

İkinci veri kümesi olan Patient Doctor Q&A TR 95588 (Bulut, 2024c), chat\_doctor veri kümesinin (Avaliev, 2024) yine GPT-3.5-turbo ile İngilizceden Türkçeye çevrilmiş versiyonudur ve aynı düzenlemeler bu veri kümesi için de geçerlidir. Üçüncü veri kümesi Patient Doctor Q&A TR 5695 (Bulut, 2024b) ise doctor-id-qa veri kümesinin (Hermansyah, 2024) Endonezceden Türkçeye çevirisidir ve benzer şekilde düzenlenmiştir.

Son olarak Patient Doctor Q&A TR 167732 (Bulut, 2024e), doktorsitesi veri kümesinin (Bayram, 2024) temizlenmiş halidir ve bu Türkçe veri kümesinde önemli düzenlemeler yapılmıştır. Bu düzenlemeler vesilesiyle bağlam bağımlı cümleler, mahremiyete halel getirebilecek içerikler, doktorların reklamları ve çeşitli iletişim bilgileri kaldırılmıştır. Bilhassa önceki konuşmalara referans veren ifadeler, hasta ve doktorların kişisel bilgileri, özel hayata dair detaylar, muayenehane adresi, çalışma saatleri, özel tanıtımlar ve iletişim bilgileri gibi içerikler de temizlenmiştir.

*Veri Birleştirme*

Dört farklı veri kümesinin birleştirilmesi sürecinde, veri kümelerinin yalnızca soru ve cevap kolonları alınarak işlem yapılmıştır. Bu süreçte Patient Doctor Q&A TR 19583, Patient Doctor Q&A TR 167732, Patient Doctor Q&A TR 5695 ve Patient Doctor Q&A TR 95588 veri kümelerinin her birinden %90 eğitim ve %10 test verisi ayrılarak birleştirme işlemi gerçekleştirilmiştir.

Yapılan bu kapsamlı birleştirme işlemi sonucunda toplam 321.179 soru-cevap çiftinden oluşan gayet zengin bir veri kümesi ortaya çıkarılmıştır. Bu kapsamlı veri kümesinin %90'ı eğitim verisi olarak ayrılırken, kalan %10'luk kısım test verisi olarak kullanılmak üzere ayrılmıştır. Veri kümesinin güvenilirliğini ve kullanılabilirliğini artırmak maksadıyla veriler karıştırılmış ve son kontroller özenle yapılarak veri kümesinin bütünlüğü sağlanmıştır.

Bilhassa belirtmek gerekir ki, her bir veri kümesinden eşit oranda eğitim ve test verisi alınması stratejik bir karardır ve bu yaklaşım, modelin farklı kaynaklardan gelen verileri dengeli bir şekilde öğrenmesini ve yine modelin dengeli bir şekilde test edilmesini sağlayarak modelin performansını optimize etmeye yardımcı olmuştur. Bu metodoloji, veri kümesinin çeşitliliğini korurken aynı zamanda modelin genelleme yeteneğini de güçlendirmiştir.

*Kullanılan Modeller*

Çalışmada dört farklı büyük dil modeli kullanılmıştır:

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

Meta AI tarafından geliştirilen Llama ailesinin üçüncü nesli olan Meta-Llama-3-8B (Meta AI, 2024), 8 milyar parametre içermekte ve haleflerine göre daha gelişmiş doğal dili anlama ve metin oluşturma özelliklerine sahiptir (meta-llama, 2024). Her ne kadar 70 milyar parametreye sahip bir varyasyonu da mevcut olsa, bu varyasyonu çalıştıracak donanıma ulaşmanın zorluğu ve maliyetli olması nedeniyle 8 milyar parametreli versiyon kullanılmak zorunda kalınmıştır. Söz konusu modelin daha verimli hesaplama gereksinimleri ve çoklu dil desteği sunması, bilhassa düşük kaynaklı diller için önemli bir avantaj teşkil etmektedir.

Meta-Llama-3-8B, daha gelişmiş bağlam anlama yeteneği ve yüksek performanslı metin oluşturma kabiliyeti vesilesiyle son dönemde doğal dil işleme alanında önemli bir adım olarak değerlendirilmektedir (meta-llama, 2024). Bu model; soru cevaplama, metin anlama ve özetleme gibi birçok görevde kapsamlı bir yelpazede kullanılabilir. Bilhassa Türkçe gibi düşük kaynaklı dillerde de başarılı sonuçlar ortaya koyması, modelin yaygın bir kullanım potansiyeli olduğunu göstermektedir.

*doktor-LLama2-sambanovasystems-7b*

SambaNova Systems tarafından geliştirilen ve Türkçe dili için özel olarak ince ayar yapılarak eğitilmiş olan bu sohbet modeli, “SambaLingo-Turkish-Base” üzerine mesajlaşmaya uyumlu hâle getirilmiş ve DPO yöntemiyle eğitilmiştir (Sambanovasystems, 2024). Llama-2-7b modelini temel alan SambaLingo-Turkish-Base, 7 milyar parametre içermekte ve Türkçe diline özel olarak uyarlanmış durumdadır (Touvron ve diğ., 2024). Model, Cultura-X veri kümesinin Türkçe bölümünden alınan 42 milyar token ile eğitilmiş olup, SFT ve DPO olmak üzere iki aşamalı bir ince ayar sürecinden geçirilmiştir. Bu kapsamlı eğitim ve uyarlama süreci, modelin Türkçe dili üzerindeki performansını ve doğal dil işleme performansını belirgin derecede geliştirmektedir. Böylece model, Türkçe dili için geliştirilmiş gayet kapsamlı bir yapay zeka modeli olma niteliğini taşımaktadır.

*doktor-Mistral-trendyol-7b*

Trendyol tarafından geliştirilen ve bilhassa Türkçe dili üzerinde ince ayar yapılmış olan bu LLM, 7 milyar parametre içermekle birlikte Türkçe için özelleştirilmiştir (Trendyol, 2024). Bu model, Türkçe doğal dil işleme alanındaki önemli bir girişim sayılabilir; yerel dil anlama ve üretme konusundaki başarısıyla dikkatleri üzerine çekmektedir. Trendyol’un bu alandaki çalışmaları, Türkçe yapay zeka ekosisteminin güçlenmesine katkıda bulunmakta ve yerli teknolojinin gelişmesini desteklemektedir.

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

YTÜ Bilgisayar Mühendisliği COSMOS Araştırma Grubu tarafından geliştirilen ve Meta-Llama-3-8B modeli üzerinde ince ayar yapılan bu LLM, toplam 8 milyar parametreye sahip olup Türkçe verilerle eğitim sürecinden geçirilmiştir (ytu-ce-cosmos, 2024). Toplam 30GB boyutundaki Türkçe veri kullanılması, modelin Türkçe dil anlama ve üretme becerilerini farkedilir bir biçimde geliştirmektedir. Böylece YTÜ COSMOS Araştırma Grubu’nun bu çalışması, Türkçe doğal dil işleme ekosistemine önemli bir akademik katkı sağlamakta ve yerli yapay zeka teknolojilerinin ilerlemesine kapı aralamaktadır.

*İnce Ayar Süreci*

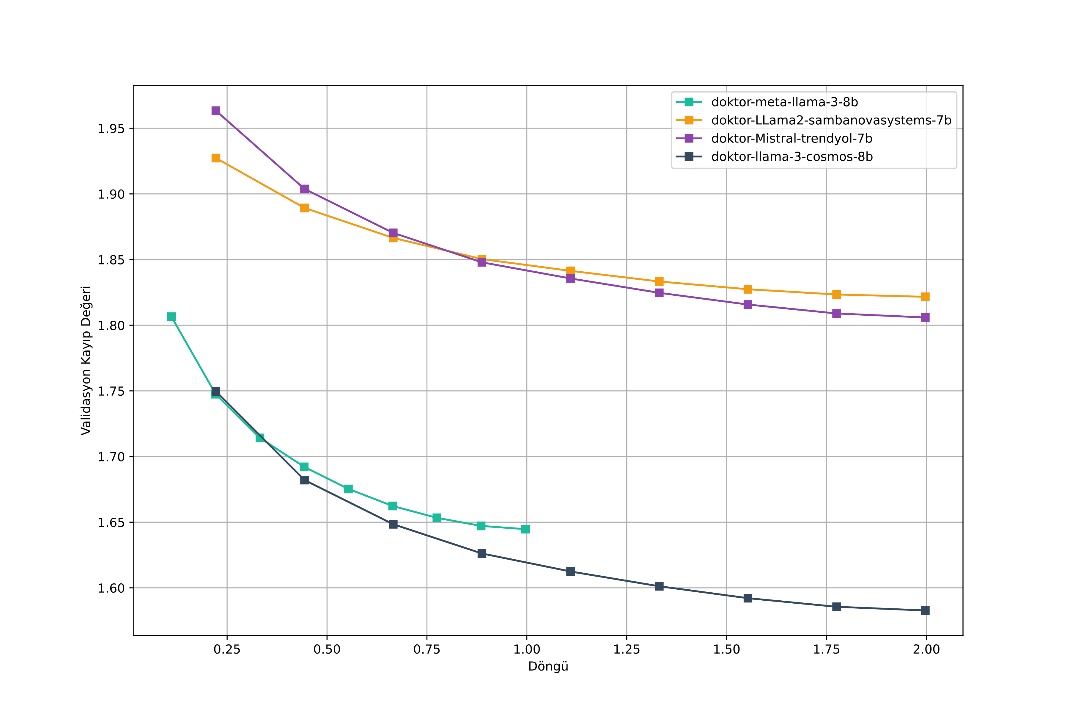
Bu bölümde, ince ayar sürecinde kullanılan hiperparametrelerden ve eğitim stratejisinden bahsedilecektir Özellikle derin öğrenme modellerinin performansını doğrudan etkileyen bu hiperparametrelerin doğru seçimi (Chen et al., 2022), modelin hem hızlı hem de kararlı bir şekilde öğrenmesini sağlamaktadır (Hoffmann et al., 2022).

*Hiperparametreler*

Öğrenme oranı (learning rate), modelin ağırlıklarının her adımda ne kadar güncelleneceğini belirleyen kritik bir hiperparametredir ve bu çalışmada 1×104 olarak belirlenmiştir. Bu değerin seçimi gayet önemlidir zira çok yüksek bir öğrenme oranı modelin aşırı büyük adımlarla öğrenerek en iyi noktayı kaçırmasına sebep olabilirken, çok düşük olması durumundaysa modelin çok yavaş öğrenerek yerel minimumlara takılabilme riski taşır (Hoffmann et al., 2022). Literatürdeki çalışmalar detaylı bir şekilde incelenerek, bu değerin 1×104 olarak belirlenmesinin en iyi sonuçları vereceği gözlemlenmiş ve bundan dolayı çalışmada bu değer tercih edilmiştir. Bu seçim, modelin hem yeterince hızlı öğrenmesine vesile olurken hem de aşırı büyük adımlardan kaçınarak daha stabil bir eğitim süreci geçirmesine olanak tanımaktadır (Devlin, 2018).

Isınma adımları (Warmup Steps), eğitim sürecinin başlangıç aşamasında öğrenme oranını tedrici olarak artıran önemli bir teknik olup, bu çalışmada 2000 adım olarak belirlenmiştir. Bu teknik, eğitimin başlangıcında düşük öğrenme oranıyla başlayarak modelin ani ağırlık güncellemelerinden korunmasını sağlarken, bununla birlikte gradyan patlaması problemini önlemektedir. Bu kademeli artış vesilesiyle, model veri kümesine ve öğrenme işine daha yumuşak bir şekilde uyum sağlama fırsatı bulur, bu da eğitim sürecinin başarısını ve modelin performansını önemli ölçüde artırır. Bilhassa karmaşık modellerde, bu ısınma periyodu modelin başlangıç aşamasında karşılaşabileceği potansiyel sorunları minimize ederek daha stabil bir eğitim süreci sağlar.

Meta-Llama-3-8B ve diğer modeller için eğitim döngüsü (epoch) sayıları farklı olarak belirlenmiş olup, Meta-Llama-3-8B için 1 döngü, diğer modeller için ise 2 döngü kullanılmıştır. Bu farklılığın temel saiki Şekil 1’de görüldüğü gibi, Meta-Llama modelinin doğrulama kaybı (validation loss) değerinin birinci döngünün sonlarına doğru düz bir çizgi çizmeye başlaması ve en iyi seviyeye ulaşmasıdır. Diğer modeller olan doktor-LLama2-sambanovasystems-7b, doktor-Mistral-trendyol-7b ve doktor-llama-3-cosmos-8b ise ikinci döngünün ortalarına doğru düz bir çizgi çizmeye başlamış, bu da eğitimin bu noktada optimum seviyeye ulaştığını göstermiştir. Her döngüde modeller tüm veri kümesini bir kez görürken, bu döngü sayıları aşırı öğrenmeyi önleyecek ve yeterli öğrenmeyi sağlayacak şekilde belirlenmiştir. Eğitim sürecinin başarısı, tüm modellerin başlangıçta yüksek olan validation loss değerlerinin eğitim ilerledikçe düşmesiyle kanıtlanmış, aynı zamanda bu değerler hesaplama kaynaklarının verimli kullanılmasını da göz önünde bulundurarak optimize edilmiştir.



Şekil 1. Doğrulama Kayıp Eğrileri

Maksimum sekans uzunluğu modeller arasında farklılık göstermekte olup, SambaLingo-Turkish-Chat modeli için 4096 token, diğer modeller için ise 8192 token olarak belirlenmiştir. Bu değer, modellerin tek seferde işleyebileceği maksimum token sayısını belirleyen önemli bir parametredir ve modelin uzun bağlamları anlayabilme yeteneğini doğrudan etkilemektedir. Aynı zamanda, bu parametre modelin bellek kullanımını da önemli ölçüde etkilediğinden, her modelin kendi mimarisi ve kapasitesi göz önünde bulundurularak optimize edilmiştir. Bu farklı token uzunlukları, modellerin performansını ve verimli çalışmasını sağlamak için özel olarak seçilmiş olup, her modelin kendi yapısal özelliklerine ve kullanım amaçlarına uygun şekilde ayarlanmıştır.

Öğrenme oranı planlayıcısı, eğitim sürecinde doğrusal bir yapıda tasarlanmış olup, modelin öğrenme sürecini optimize etmek için önemli bir rol oynamaktadır. Bu planlayıcı, eğitimin başlangıç aşamasında daha yüksek bir öğrenme oranıyla başlayarak modelin hızlı öğrenmesini sağlar ve zaman içerisinde bu oranı doğrusal bir şekilde azaltır. Bu doğrusal azalış sayesinde, eğitimin sonlarına doğru daha hassas bir öğrenme süreci gerçekleşir ve model daha ince ayarlamalar yapabilir. Bu yaklaşım, modelin başlangıçta hızlı öğrenmesini sağlarken, ilerleyen aşamalarda daha stabil ve hassas bir şekilde optimize olmasına olanak tanır, böylece eğitim sürecinin verimliliği ve modelin performansı maksimize edilmiş olur.

Ağırlık azaltma parametresi bu çalışmada 0.01 olarak belirlenmiş olup (Wu & Sun, 2022), bu değer modelin aşırı öğrenmesini engellemek için kullanılan önemli bir regularizasyon tekniğidir Bu teknik, model parametrelerinin büyüklüğünü kontrol ederek ağın genelleme yeteneğini artırır ve eğitim sürecinde gradyan patlamasını önlemeye yardımcı olur Ağırlık azaltma, eğitim sırasında ağırlıkları küçülterek zamanla daha küçük değerlere ulaşmalarını sağlar, bu da modelin aşırı özelleşmesini engelleyerek daha iyi bir genelleme performansı elde edilmesine katkıda bulunur. Bu değer, literatürdeki yaygın kullanım aralığı olan 0.1 ile 0.0001 arasındaki değerler göz önünde bulundurularak seçilmiş ve modelin optimal performans göstermesi için ayarlanmıştır.

Bu çalışmada optimizasyon algoritması olarak 8-bit AdamW kullanılmış olup, bu algoritma standart AdamW'nin 8-bitlik versiyonu olarak modelin ağırlıklarını güncellemek için optimize edilmiş bir yapıya sahiptir (Dettmers et al., 2021). Bu algoritma, bellek kullanımını önemli ölçüde azaltırken aynı zamanda eğitim hızını da artırmaktadır.

Modelin sayısal hesaplamalarında ise hassasiyet formatı olarak BF16 tercih edilmiştir. Bu format, FP32'ye göre daha az bellek kullanımı sağlarken, FP16'ya kıyasla daha geniş bir dinamik aralık sunmakta ve eğitim stabilitesini korumaktadır. Bu iki özelliğin bir arada kullanılması, modelin hem verimli bir şekilde eğitilmesini hem de yüksek performans göstermesini sağlamakta, aynı zamanda hesaplama kaynaklarının da optimal kullanımına olanak tanımaktadır.

*Eğitim Stratejisi*

Eğitim stratejisi, modellerin Türkçe tıbbi metinleri daha iyi anlayabilmesi ve performansını artırmak amacıyla özel olarak tasarlanmış olup, bu stratejinin merkezinde LoRA (Low-Rank Adaptation) tekniğinin kullanımı yer almaktadır. Bu teknik, modelin tüm parametrelerini güncellemek yerine düşük boyutlu matrisler ekleyerek eğitimi önemli ölçüde hızlandırmayı başarmıştır. LoRA'nın kullanımı sayesinde bellek kullanımı minimize edilmiş ve eğitim süresi önemli ölçüde kısaltılmıştır. Bu yaklaşım, özellikle büyük dil modellerinin eğitiminde karşılaşılan hesaplama maliyeti ve kaynak kullanımı gibi zorlukları aşmada etkili bir çözüm sunmuş, aynı zamanda modelin performansından ödün vermeden daha verimli bir eğitim süreci sağlamıştır.

Unsloth kütüphanesi, eğitim sürecini optimize etmek için özel olarak tercih edilmiş olup, özellikle ince ayar sürecindeki bellek kullanımını önemli ölçüde azaltmak için kullanılmıştır (Unsloth, 2024). Bu kütüphane, eğitim sürecini yaklaşık 2 kat hızlandırırken, bellek kullanımını %70 oranında azaltmayı başarmış ve modelin doğruluk oranından herhangi bir ödün vermeden kaynak verimliliği sağlamıştır. Unsloth'un kullanımı, özellikle Llama-3 ve Mistral gibi büyük dil modellerinin eğitiminde önemli avantajlar sağlamış, bu da eğitim sürecinin hem daha hızlı hem de daha verimli bir şekilde tamamlanmasına olanak tanımıştır.

Veri kümesi hazırlığı, modelin eğitim sürecinde önemli bir adım olarak gerçekleştirilmiş olup, bu süreçte öncelikle Endonezce ve İngilizce dillerindeki veriler Türkçe'ye çevrilmiş ve mevcut Türkçe veri kümesi ile birleştirilmiştir. Kapsamlı bir veri temizleme ve standardizasyon işlemi gerçekleştirilmiş, bu süreçte verilerin iletişim bilgileri, özel isimler ve bağlam bağımlı cümleleri titizlikle ayıklanmıştır. Bu hazırlık aşaması, yapay zeka modelinin başarısını doğrudan etkileyecek temel adımlardan biri olarak değerlendirilmiştir. Bu işlemler manzumesi icra edilirken daha büyük ve genel bir model olan gpt-3.5-turbo’dan yardım alınmıştır (OpenAI, 2024a).

Değerlendirme ve kaydetme stratejisi, modelin eğitim sürecini güvence altına almak ve performansını düzenli olarak izlemek için kapsamlı bir şekilde tasarlanmıştır. Model performansı, loss metriği kullanılarak her 57.800 adımda bir değerlendirilmiş, bu da modelin öğrenme sürecinin düzenli olarak takip edilmesini sağlamıştır. Olası teknik sorunlara karşı bir önlem olarak, model her 10.000 adımda bir eğitim yapılan cihazın beklenmedik kapanma durumlarına karşı otomatik olarak kaydedilmiştir. Bu düzenli kontrol noktaları sayesinde, herhangi bir kesinti durumunda eğitimin en son kaydedilen noktadan devam edebilmesi sağlanmış ve böylece eğitim sürecinin güvenliği ve sürekliliği garanti altına alınmıştır. Bu stratejik yaklaşım, hem modelin eğitim sürecinin güvenilirliğini artırmış hem de performans izleme ve değerlendirme süreçlerinin sistematik bir şekilde yürütülmesini sağlamıştır.

Google Colab platformunda Tesla A100 40 GB GPU kullanılarak gerçekleştirilen eğitim sürecinde, Cosmos LLama 19,4 saat, Meta LLama 11,38 saat, SambaLingo 17,46 saat ve Trendyol modeli 19,3 saat sürede eğitimlerini tamamlamış olup, bu stratejik yaklaşım ve donanım seçimi, modellerin verimli ve etkili bir şekilde eğitilmesini sağlamış, özellikle Tesla A100 GPU'nun sunduğu yüksek hesaplama gücü ve tensor core mimarisi sayesinde eğitim süreleri optimize edilmiş ve yapay zeka uygulamaları için benzersiz bir hızlanma elde edilmiştir (NVIDIA, 2024; Google, 2024b).

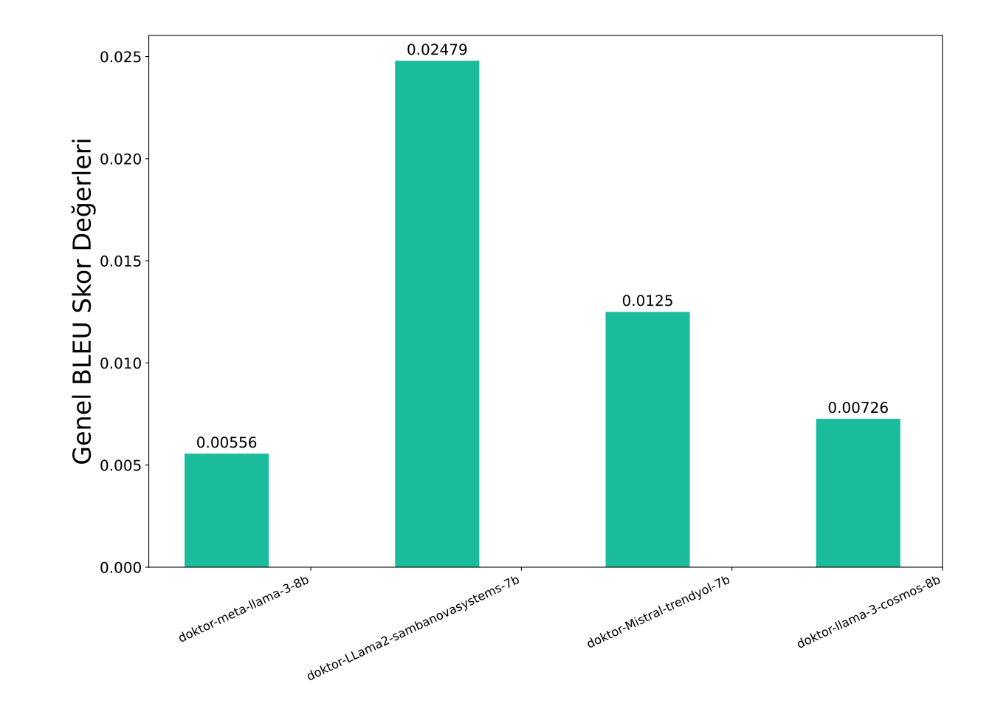
# BULGULAR

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.

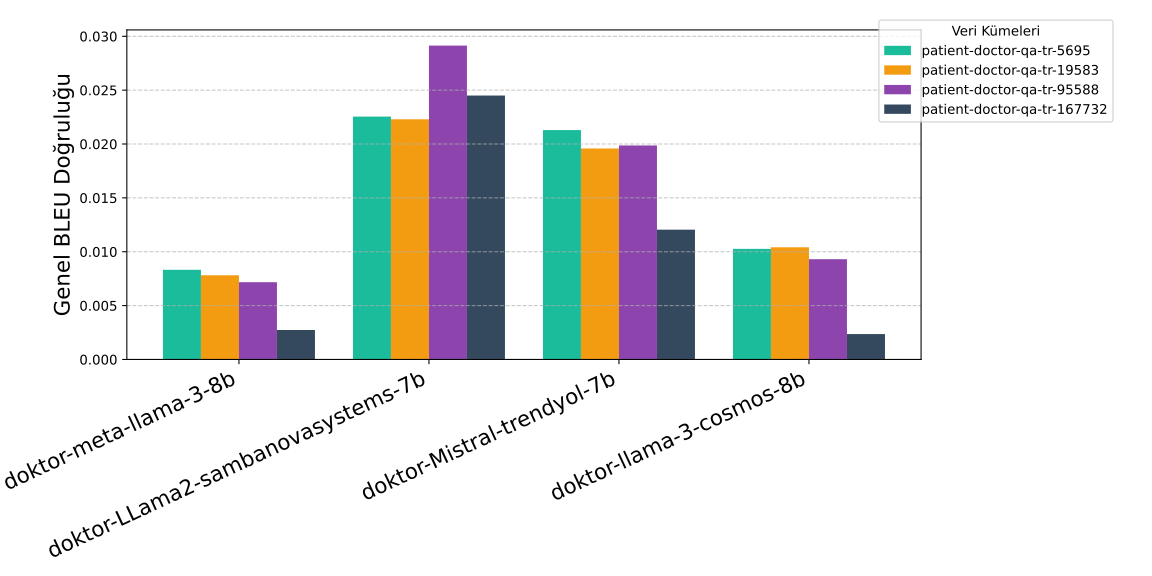
*BLEU Skor Sonuçları*



Şekil 2. Genel BLEU Skorları

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’deki sonuçların detaylı analizine göre, doktor-LLama2-sambanovasystems-7b modeli 0.02479 BLEU skoru ile en yüksek performansı gösterirken, doktor-Mistral-trendyol-7b modeli 0.0125 BLEU skoru ile ikinci sırada yer almıştır. Üçüncü sırada 0.00726 BLEU skoru ile doktor-llama-3-cosmos-8b modeli bulunurken, doktor-meta-llama-3-8b modeli 0.00556 BLEU skoru ile en düşük performansı sergilemiştir. Bu sonuçlar, modellerin ürettiği metinlerin referans metinlere olan yakınlığını göstermektedir.

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



Şekil 3. Veri Kümesi Özelinde BLEU Skorları

Şekil 3'te gösterildiği üzere, dört farklı modelin patient-doctor-qa-tr-5695, patient-doctor-qa-tr-19583, patient-doctor-qa-tr-95588 ve patient-doctor-qa-tr-167732 veri kümeleri üzerindeki BLEU doğruluk skorları karşılaştırılmıştır. Şekil 3'ün sonuçlarına göre, doktor-LLama2-sambanovasystems-7b modeli en yüksek performansı patient-doctor-qa-tr-95588 veri kümesinde yaklaşık 0.029 skorla göstermiştir. Model, diğer veri kümelerinde de tutarlı bir şekilde 0.020-0.025 aralığında başarı göstermiş ve tüm veri kümelerinde diğer modellere göre daha iyi performans sergilemiştir.

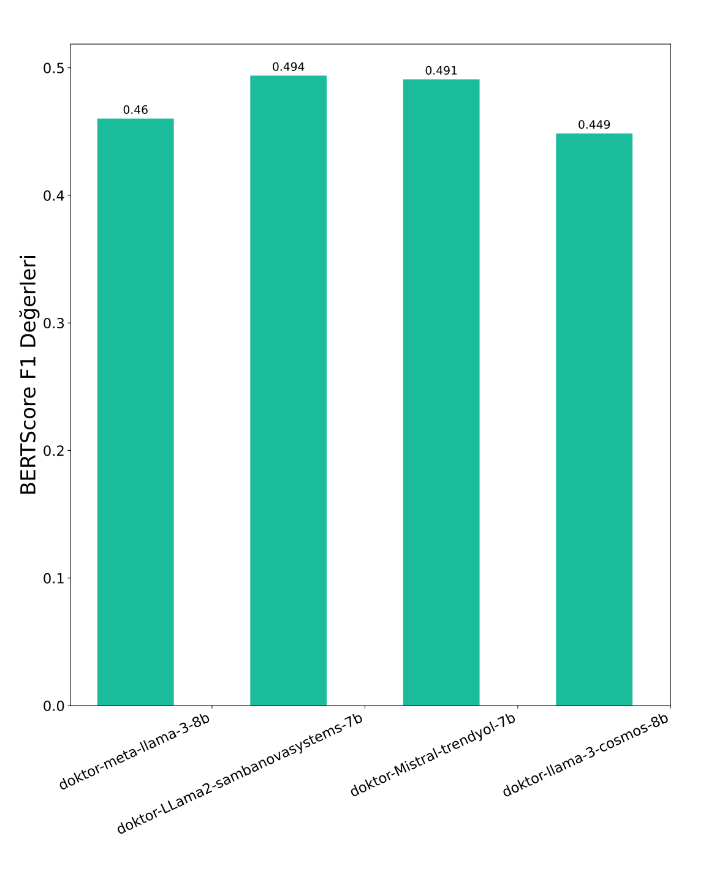
Şekil 3'teki veriler incelendiğinde, doktor-Mistral-trendyol-7b modelinin en iyi performansı patient-doctor-qa-tr-5695 veri kümesinde yaklaşık 0.021 skorla gösterdiği görülmektedir. Model, tüm veri kümelerinde 0.012-0.021 aralığında tutarlı bir başarı göstermiş ve genel olarak ikinci en iyi performansa sahip model olmuştur.

Şekil 3'te görüldüğü gibi, doktor-meta-llama-3-8b modeli veri kümeleri arasında 0.003-0.008 aralığında nispeten düşük ve tutarlı bir performans göstermiştir. Modelin en düşük performansı patient-doctor-qa-tr-167732 veri kümesinde gözlemlenmiştir.

Şekil 3'ün sonuçlarına göre, doktor-llama-3-cosmos-8b modeli tüm veri kümelerinde 0.002-0.010 aralığında değişen en düşük performansı göstermiştir. Model, en iyi sonucunu patient-doctor-qa-tr-5695 veri kümesinde elde etmiştir.

Genel olarak bakıldığında, 7b parametreli modeller olan doktor-LLama2-sambanovasystems-7b ve doktor-Mistral-trendyol-7b'nin, 8b parametreli modellere kıyasla daha üstün bir performans sergilediği görülmüştür. İlginç bir şekilde, veri kümesinin büyüklüğü ile modellerin performansı arasında doğrudan bir korelasyon gözlenmemiştir. Özellikle dikkat çeken bir nokta, Türkçe'ye özel eğitilmiş modeller olan sambanovasystems ve trendyol'un diğer modellere göre daha başarılı sonuçlar elde etmesidir.

*BERT Skor Sonuçları*

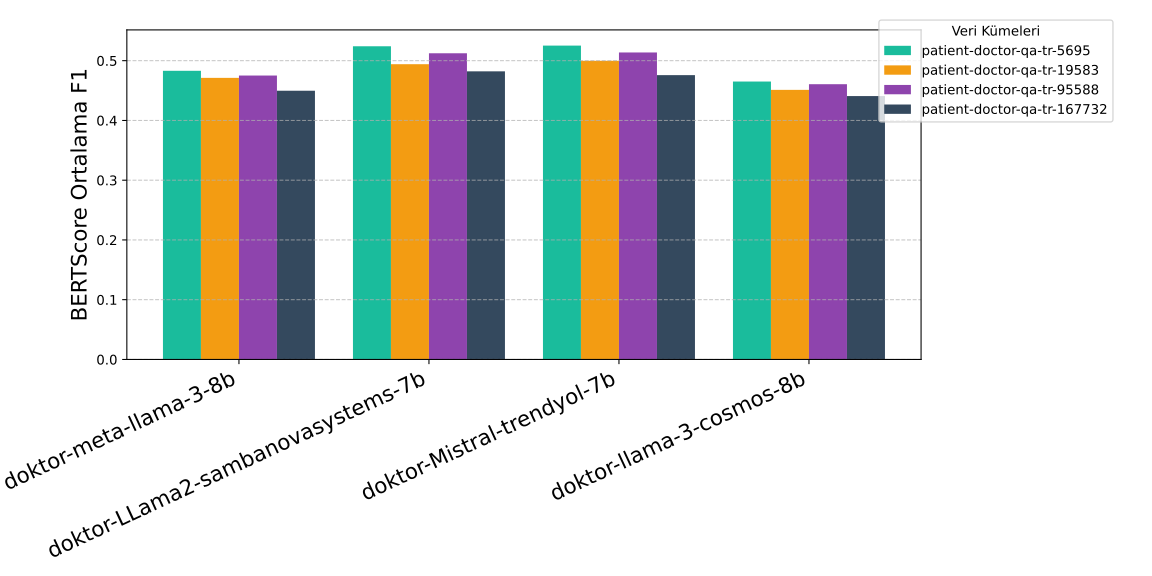


Şekil 4. BERT Skor F1 Skorları

Şekil 4’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:

Şekil 4'te gösterilen dört farklı dil modelinin BERT Skor F1 değerleri karşılaştırıldığında, doktor-LLama2-sambanovasystems-7b modelinin 0.494 F1 skoru ile en yüksek performansı gösterdiği görülmektedir. Bu sonuç, modelin ürettiği metinlerin anlamsal olarak referans metinlere en yakın olduğunu kanıtlamaktadır. Yine Şekil 4'teki verilere göre, doktor-Mistral-trendyol-7b modeli 0.491 F1 skoru ile ikinci sırada yer alırken, doktor-meta-llama-3-8b modeli 0.46 F1 skoru ile üçüncü sırada ve doktor-llama-3-cosmos-8b modeli 0.449 F1 skoru ile son sırada yer almaktadır.

Şekil 4'e göre, ilk iki model arasındaki minimal fark (0.003), her iki modelin de Türkçe sağlık metinlerini anlamsal olarak başarılı bir şekilde işleyebildiğini göstermektedir. 8B parametreli modellerin 7B parametreli modellere göre daha düşük performans göstermesi, parametre sayısının tek başına başarıyı belirlemediğini ortaya koymaktadır. Ayrıca, en yüksek ve en düşük skorlar arasındaki göreceli olarak küçük fark (0.045), tüm modellerin makul bir seviyede performans sergilediğine işaret etmektedir.



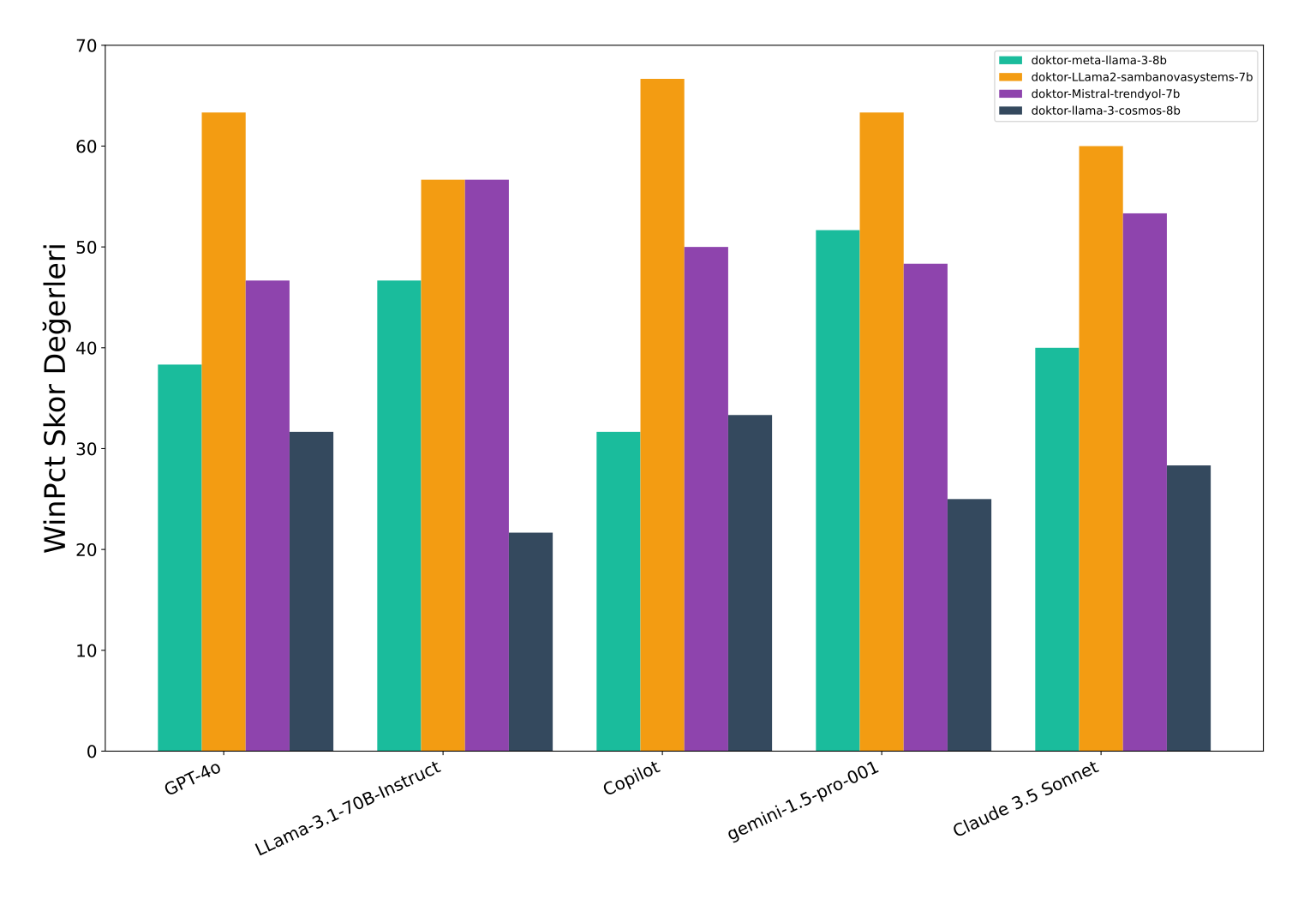
Şekil 5. Veri Kümesi Özelinde BERT Skor F1 Skorları

Şekil 5'te gösterilen dört farklı modelin farklı veri kümeleri üzerindeki BERT Skor F1 performansları karşılaştırıldığında, doktor-LLama2-sambanovasystems-7b modelinin en yüksek performansı patient-doctor-qa-tr-5695 veri kümesinde yaklaşık 0.52 skorla gösterdiği görülmektedir. Model, diğer veri kümelerinde de tutarlı bir şekilde 0.48-0.51 aralığında başarı göstermiş ve çoğu veri kümesinde diğer modellere göre daha iyi performans sergilemiştir. Şekil 5'teki verilere göre, doktor-Mistral-trendyol-7b modeli en iyi performansını patient-doctor-qa-tr-5695 veri kümesinde yaklaşık 0.53 skorla göstermiş ve tüm veri kümelerinde 0.475-0.525 aralığında tutarlı bir başarı sergileyerek genel olarak ikinci en iyi performansa sahip model olmuştur.

Şekil 5'in sonuçlarına göre, doktor-meta-llama-3-8b modeli veri kümeleri arasında 0.449-0.48 aralığında nispeten düşük ve tutarlı bir performans göstermiş, en düşük performansını ise patient-doctor-qa-tr-167732 veri kümesinde sergilemiştir. Diğer yandan, doktor-llama-3-cosmos-8b modeli tüm veri kümelerinde 0.44-0.465 aralığında değişen en düşük performansı göstermiş ve en iyi sonucunu patient-doctor-qa-tr-5695 veri kümesinde elde etmiştir.

Genel olarak baktığımızda, 7b parametreli modeller olan doktor-LLama2-sambanovasystems-7b ve doktor-Mistral-trendyol-7b'nin, 8b parametreli modellere kıyasla daha üstün bir performans sergilediği görülmüştür. Benzer bir şekilde, veri kümesinin büyüklüğü ile modellerin performansı arasında doğrudan bir korelasyon gözlenmemiştir. Özellikle dikkat çeken bir nokta, Türkçe'ye özel eğitilmiş modeller olan sambanovasystems ve trendyol'un diğer modellere göre daha başarılı sonuçlar elde etmesidir.

*Yapay Zeka Hakemliğinde Değerlendirme*



Şekil 6. 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 6'da gösterilmiştir:

Şekil 6'da gösterilen GPT-4 hakemliğinin sonuçlarına göre, doktor-LLama2-sambanovasystems-7b modeli %63.33 ile en yüksek başarıyı elde etmiştir. Bunu %46.67 ile doktor-Mistral-trendyol-7b takip ederken, doktor-meta-llama-3-8b %38.33 ve doktor-llama-3-cosmos-8b %31.67 ile daha düşük performans sergilemiştir. Bu sonuçlar, sambanovasystems modelinin GPT-4 değerlendirmesinde açık ara önde olduğunu göstermektedir.

LLaMA 3.1 70B hakemliğinde ilginç bir sonuç ortaya çıkmış, doktor-LLama2-sambanovasystems-7b ve doktor-Mistral-trendyol-7b modelleri %56.67 ile eşit ve en yüksek performansı göstermiştir. Doktor-meta-llama-3-8b %46.67 ile orta düzeyde kalırken, doktor-llama-3-cosmos-8b %21.67 ile belirgin şekilde düşük bir performans sergilemiştir.

Microsoft Copilot (Microsoft, 2024) hakemliğinde, doktor-LLama2-sambanovasystems-7b modeli %66.67 ile en yüksek skoru elde etmiştir. doktor-Mistral-trendyol-7b %50.00 ile ikinci sırada yer alırken, doktor-llama-3-cosmos-8b %33.33 ve doktor-meta-llama-3-8b %31.67 ile daha düşük performans göstermiştir. Bu sonuçlar, sambanovasystems modelinin tutarlı başarısını bir kez daha kanıtlamıştır.

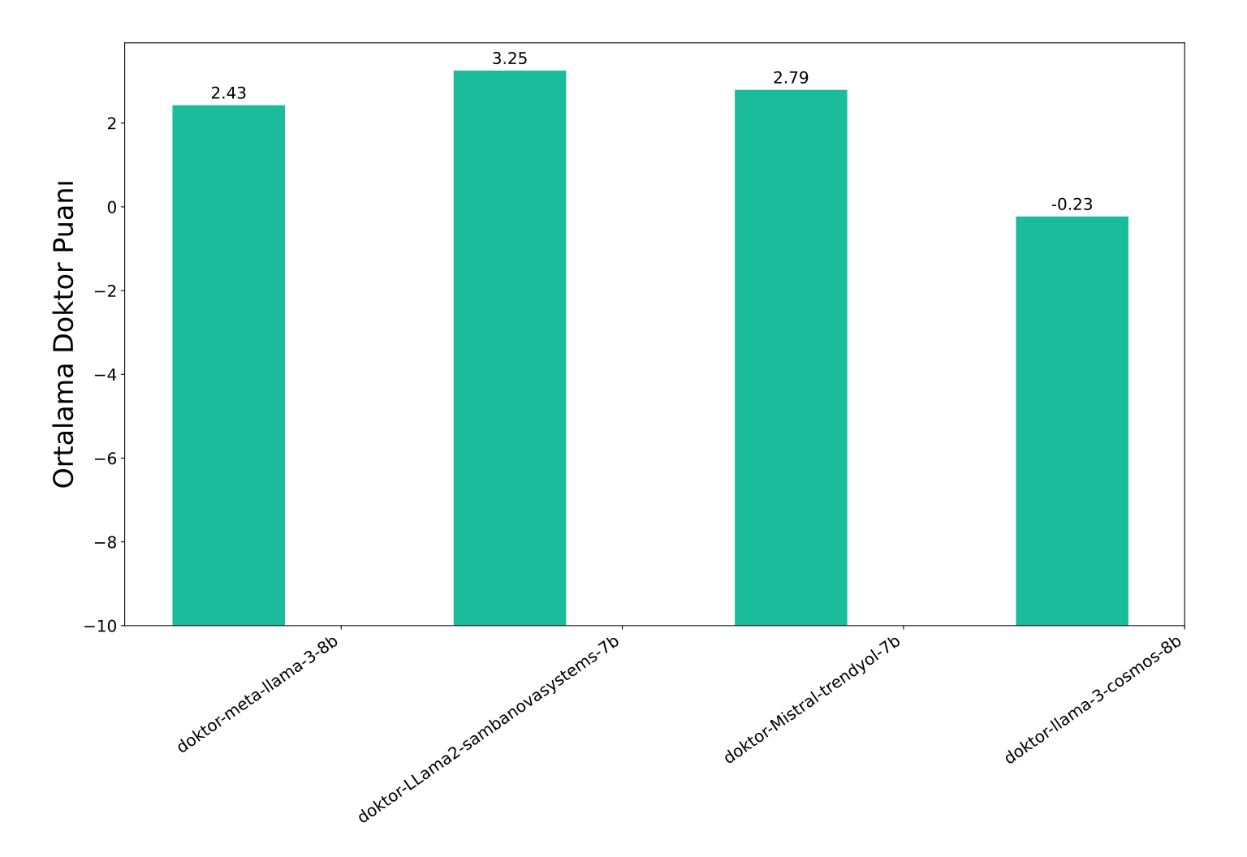
Gemini 1.5 Pro hakemliğinde de doktor-LLama2-sambanovasystems-7b %63.33 ile liderliğini sürdürmüştür. İlginç bir şekilde, bu değerlendirmede doktor-meta-llama-3-8b %51.67 ile ikinci sıraya yükselmiş, doktor-Mistral-trendyol-7b %48.33 ile üçüncü olmuş ve doktor-llama-3-cosmos-8b %25.00 ile son sırada yer almıştır.

Claude 3.5 Sonnet (Anthropic, 2024) hakemliğinde de benzer bir sıralama görülmüş, doktor-LLama2-sambanovasystems-7b %60.00 ile en yüksek performansı göstermiştir. Doktor-Mistral-trendyol-7b %53.33, doktor-meta-llama-3-8b %40.00 ve doktor-llama-3-cosmos-8b %28.33 ile sıralanmıştır. Bu sonuçlar, farklı hakem modeller arasında tutarlı bir değerlendirme olduğunu göstermektedir.

Tüm hakem değerlendirmelerinin genel bir analizi yapıldığında, doktor-LLama2-sambanovasystems-7b modelinin tüm hakem değerlendirmelerinde ya en yüksek puanı aldığı ya da en yüksek puanlardan birine sahip olduğu görülmektedir. Bu tutarlı başarı, modelin genel kalitesini ve güvenilirliğini kanıtlar niteliktedir. Diğer yandan, doktor-llama-3-cosmos-8b modeli tüm hakem değerlendirmelerinde genellikle en düşük performansı göstermiş, bu da modelin iyileştirmeye ihtiyaç duyduğuna işaret etmektedir. Doktor-Mistral-trendyol-7b ve doktor-meta-llama-3-8b modelleri ise değerlendirmelerde genellikle orta sıralarda yer almış, bu da bu modellerin kabul edilebilir ancak geliştirilebilir bir performans sergilediğini göstermektedir. Bilhassa dikkat çeken bir nokta, farklı hakem modeller arasında görülen tutarlı değerlendirme sonuçlarıdır, bu da değerlendirme sürecinin güvenilirliğini desteklemektedir.

*Uzman Değerlendirmesi Sonuçları*

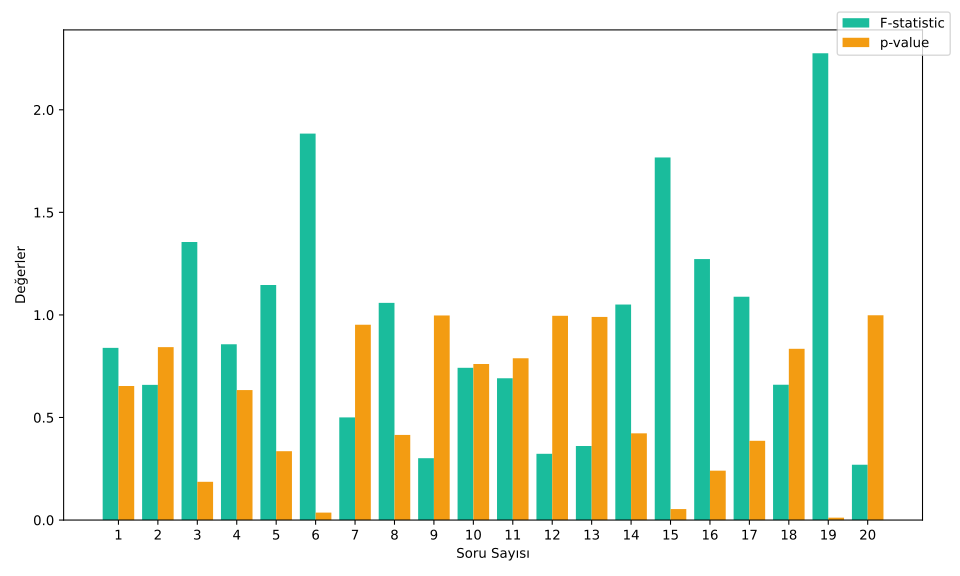
Uzman değerlendirmesi kapsamında, farklı uzmanlık alanlarından 20 doktor tarafından modellerin performansı değerlendirilmiştir. Değerlendirme sürecinde, rastgele seçilen 20 hasta sorusuna modellerin verdiği yanıtlar, -10 ile +10 arasında puanlanmıştır. Burada -10 yanıtın çok zararlı ve yanlış yönlendirici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



Şekil 7. Uzmanlar Tarafından Değerlendirme Sonuçları

Şekil 7'de sunulan değerlendirme sonuçları, özellikle sağlık alanında insan hayatını doğrudan etkileyen bir konuda yapay zeka modellerinin performansını göstermesi açısından büyük önem taşımaktadır. SambaLingo-Turkish-Chat modelinin 3.25 ortalama puanla en yüksek değerlendirmeyi alması, modelin Türkçe sağlık iletişiminde daha doğal ve anlaşılır yanıtlar üretebildiğini göstermektedir. Trendyol-LLM-7b-chat-v1.8'in 2.79 ve Meta-Llama-3-8B'nin 2.43 puanla orta düzeyde performans sergilemeleri, bu modellerin klinik ortamlarda kullanılabilir olmakla birlikte geliştirilmeye açık olduklarını işaret etmektedir. Turkish-Llama-8b-v0.1 modelinin -0.23 puanla negatif değerlendirme alması ise, sağlık gibi hassas bir alanda kullanılmadan önce ciddi iyileştirmelere ihtiyaç duyulduğunu göstermektedir. Bu sonuçlar, yapay zeka modellerinin doktor-hasta iletişiminde destekleyici bir araç olarak kullanılabileceğini, ancak insan doktorların yerini alamayacağını bir kez daha vurgulamaktadır.

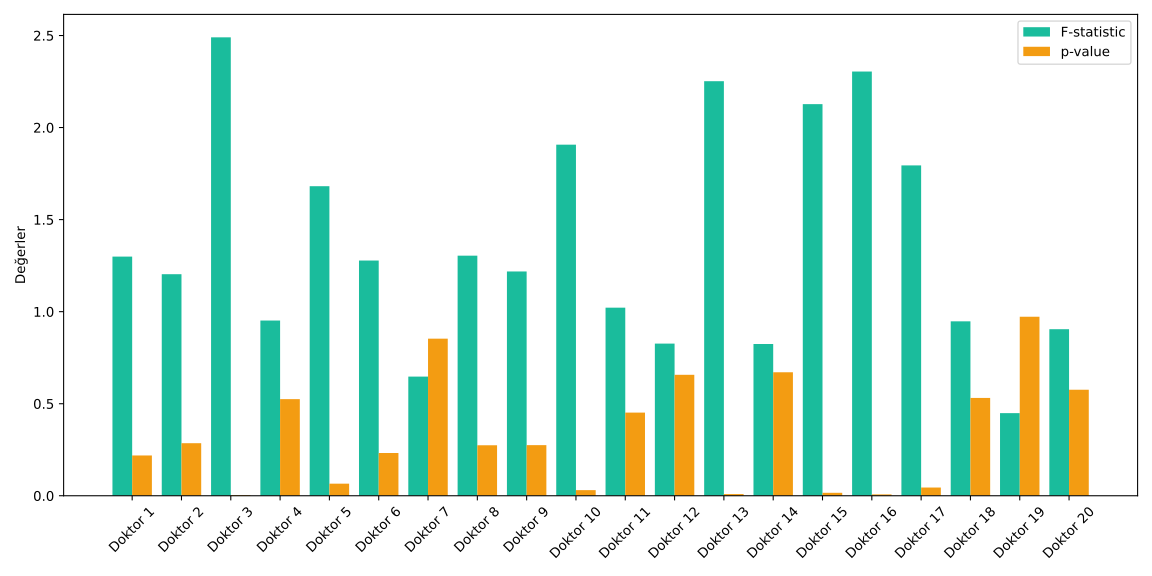
*ANOVA Sonuçları*



Şekil 8. Soruların ANOVA Skorları

Şekil 8’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.

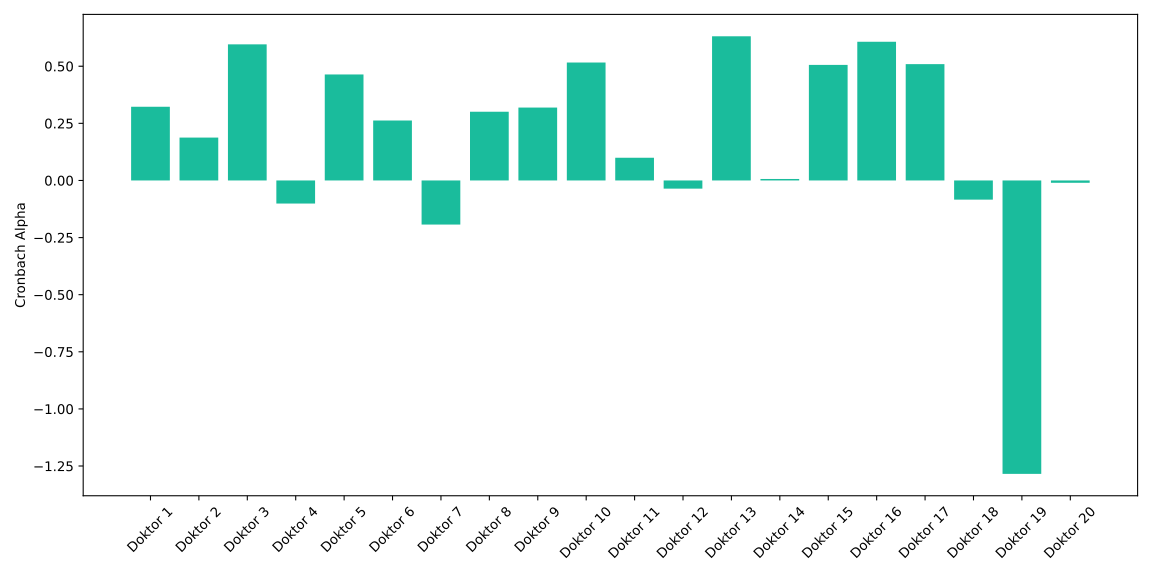


Şekil 9. Doktorların ANOVA Skorları

Şekil 9’da 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*

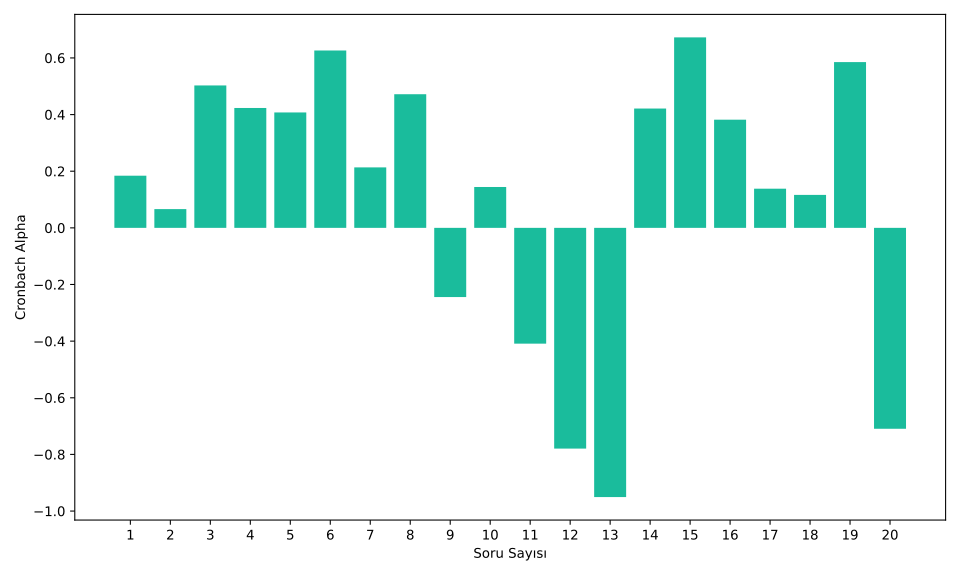


Şekil 10. Doktorların CRONBACH Skorları

Şekil 10'da sunulan 20 farklı doktorun değerlendirmelerine ait Cronbach Alfa değerleri incelendiğinde, -1.25 ile +0.75 arasında değişen bir dağılım gözlenmektedir. En yüksek tutarlılığı yaklaşık 0.6 değeriyle Doktor 3 gösterirken, en düşük tutarlılık yaklaşık -1.25 değeriyle Doktor 19'da görülmüştür. Değerlendirmeye katılan doktorların çoğunluğunun 0 ile 0.5 arasında değerler alması, genel olarak orta düzeyde bir tutarlılığın varlığına işaret etmektedir. Bu sonuçlar, doktorların yapay zeka modellerini değerlendirirken farklı bakış açılarına sahip olduklarını ve değerlendirme kriterlerini yorumlamada bireysel farklılıklar gösterdiklerini ortaya koymaktadır.

Şekil 11'de gösterilen soru bazındaki Cronbach Alfa değerleri analiz edildiğinde, en yüksek tutarlılığın yaklaşık 0.65 değeriyle 15. soruda, en düşük tutarlılığın ise yaklaşık -0.95 değeriyle 13. soruda görüldüğü tespit edilmiştir. Soruların yaklaşık yarısının pozitif, diğer yarısının negatif değerler alması, değerlendirme sürecinde önemli bir varyasyon olduğunu göstermektedir.

Bu sonuçlar, hem doktor hem de soru bazında önemli tutarlılık farklılıklarının varlığına işaret etmektedir. Bazı doktorlar ve sorular için tutarlılığın oldukça düşük olması ve genel olarak değerlendirmelerde orta düzeyde bir tutarlılık gözlenmesi, değerlendirme sürecinin standardizasyonunun artırılması ve değerlendiriciler arası tutarlılığın iyileştirilmesi gerektiğini açıkça ortaya koymaktadır.



Şekil 11. Soruların CRONBACH Skorları

# TARTIŞMA VE SONUÇ

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

*Bulguların Değerlendirilmesi*

Çalışmamızın sonuçları, dört farklı büyük dil modelinin Türkçe sağlık verileri üzerindeki performansını kapsamlı bir şekilde ortaya koymaktadır. doktor-LLama2-sambanovasystems-7b modeli, BLEU ve BERT Skor gibi tüm objektif ölçütlerde en yüksek başarımı göstermiş, uzman değerlendirmelerinde ve Elo (Elo, 1978) puanlaması ortalamasında da birinci sırada yer almıştır. doktor-Mistral-trendyol-7b modeli ikinci en iyi performansı sergilemiştir.

doktor-meta-llama-3-8b modeli, genel amaçlı bir model olmasına rağmen makul bir performans göstererek çoğu ölçütte üçüncü sırada yer alırken, doktor-llama-3-cosmos-8b modeli çoğu değerlendirme ölçütünde en düşük performansı sergilemiştir. ANOVA ve Cronbach alfa skorları, değerlendirmelerde doktor ve soru bazında belirgin varyasyonlar olduğunu göstermekle birlikte, genel olarak orta düzeyde bir tutarlılığın varlığına işaret etmektedir.

Özellikle dikkat çeken bir nokta, uzman değerlendirmeleri ile sentetik ölçütler (BLEU, ROUGE, vb.) ve yapay zeka hakemliklerinin sonuçları arasında güçlü bir korelasyon gözlenmesidir. Bu durum, farklı değerlendirme yöntemlerinin birbirini desteklediğini ve sonuçların güvenilirliğini artırdığını göstermektedir. Bu bulgular, Türkçe sağlık alanında kullanılacak dil modellerinin seçiminde ve geliştirilmesinde önemli içgörüler sunmaktadır.

*Kısıtlamalar ve Öneriler*

Çalışmamızda karşılaşılan temel kısıtlamalar arasında veri kümesi boyutu ve çeşitliliğinin sınırlı olması, hesaplama kaynaklarının kısıtlı olması, test edilen model sayısı ve çeşitliliğinin artırılabilir olması ve değerlendirme metriklerinin genişletilebilir olması yer almaktadır. Bu kısıtlamalar, özellikle sağlık alanı gibi hassas ve önemli bir konuda daha kapsamlı sonuçlar elde edilmesini sınırlandırmıştır. Ayrıca, mevcut hesaplama kaynaklarının sınırlı olması, daha büyük ve karmaşık modellerin test edilmesini zorlaştırmış ve potansiyel olarak daha iyi performans gösterebilecek birçok modelin değerlendirme dışı kalmasına neden olmuştur.

Bu kısıtlamaların aşılması için öncelikle daha geniş ve çeşitli veri kümelerinin oluşturulması, farklı model mimarilerinin test edilmesi ve daha kapsamlı değerlendirme ölçütlerinin geliştirilmesi önerilmektedir. Özellikle Türkçe sağlık alanında daha zengin ve çeşitli veri kümelerinin oluşturulması, modellerin performansını artırabilir ve daha güvenilir sonuçlar elde edilmesini sağlayabilir. Bunun yanı sıra, etik ve hukuki çerçevenin güçlendirilmesi, hasta mahremiyetinin korunması ve verilerin güvenli bir şekilde işlenmesi açısından büyük önem taşımaktadır. Gelecekteki çalışmalarda, bu önerilerin dikkate alınması ve uygulanması, Türkçe sağlık alanında daha etkili ve güvenilir yapay zeka modellerinin geliştirilmesine katkı sağlayacaktır.

*Gelecek Çalışmalar*

Gelecek çalışmalar için önerilerimiz, sağlık alanında yapay zeka uygulamalarının daha kapsamlı ve etkili hale getirilmesini hedeflemektedir. Özellikle görüntü, ses ve metin verilerini birlikte işleyebilen multimodal modellerin sağlık alanında kullanımının araştırılması, teşhis ve tedavi süreçlerinde daha bütüncül bir yaklaşım sağlayabilir. Aynı zamanda, daha büyük ve çeşitli Türkçe sağlık veri kümelerinin oluşturulması ve model mimarilerinin Türkçe dil yapısına göre optimize edilmesi, modellerin performansını önemli ölçüde artırabilir.

Veri güvenliği ve gizliliğine yönelik çözümlerin geliştirilmesi, hasta mahremiyetinin korunması açısından kritik öneme sahiptir. Bu bağlamda, Sağlık Bakanlığı gibi kurumlarla işbirliklerinin artırılması, hem veri kaynaklarının zenginleştirilmesi hem de yasal ve etik çerçevenin güçlendirilmesi açısından büyük önem taşımaktadır. Bu önerilerin hayata geçirilmesi, Türkiye'de sağlık alanında yapay zeka uygulamalarının daha güvenilir ve etkili bir şekilde kullanılmasına katkı sağlayacaktır.

*Sonuç*

Bu çalışma, LLM'lerin Türkçe sağlık alanında önemli bir potansiyele sahip olduğunu göstermektedir. Özellikle dile özgü eğitilmiş modellerin (doktor-LLama2-sambanovasystems-7b ve doktor-Mistral-trendyol-7b) daha başarılı olduğu görülmüştür. Ancak bu teknolojinin güvenli ve etik kullanımı için daha fazla araştırma gerekmektedir. Sağlık gibi önemli bir alanda yapay zeka kullanımının hem fırsatları hem de riskleri dikkatle değerlendirilmelidir.Çalışmamızda kullanılan farklı değerlendirme yöntemleri (sentetik metrikler, yapay zeka hakemlikleri ve uzman değerlendirmeleri), modellerin performansını çok yönlü olarak anlamamıza olanak sağlamıştır. Özellikle uzman değerlendirmeleri, modellerin pratik kullanımdaki potansiyelini ve sınırlarını ortaya koymuştur. doktor-LLama2-sambanovasystems-7b modelinin genel başarısı ve doktor-Mistral-trendyol-7b modelinin düşük zararlı içerik oranı, bu modellerin sağlık alanında kullanılabilirliğini göstermektedir.Bununla birlikte, çalışmamız LLM'lerin sağlık alanında kullanımına ilişkin bazı önemli zorlukları da ortaya çıkarmıştır. Hasta mahremiyeti, veri güvenliği, etik karar verme ve yasal sorumluluk gibi konular, bu teknolojinin yaygınlaşması önündeki en önemli engellerdir. Bu zorlukların aşılması için sağlık otoriteleri, teknoloji uzmanları ve hukukçuların işbirliği içinde çalışması gerekmektedir.

# KAYNAKLAR

E., Cengiz, C., Altinuç, S. O., & Temizel, A. (2022). Automated question generation and question answering from Turkish texts. Turkish Journal of Electrical Engineering and Computer Sciences, 30(5), 1931–1940.

Anthropic. (2024). Claude: A New AI Assistant by Anthropic. Erişim adresi: https://www.anthropic.com/news/claude-3-5-sonnet (Erişim tarihi 16/08/2024).

Avaliev, A. (2024). Chat Doctor Dataset. Erişim adresi: https://huggingface.co/datasets/avaliev/chat\_doctor (Erişim tarihi 18/08/2024).

Bayram, M. A. (2024). Türkçe Tıbbi Soru-Cevap Veri Seti [Veri seti]. https://doi.org/10.5281/zenodo.12770916 (Erişim adresi: https://zenodo.org/record/12770916).

Brown, T. B. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165.

Bulut, M. K. (2024a). Patient Doctor Q&A TR 321179. https://doi.org/10.5281/zenodo.12798934 (Erişim adresi: https://doi.org/10.5281/zenodo.12798934).

Bulut, M. K. (2024b). Patient Doctor Q&A TR 5695. Erişim adresi: https://huggingface.co/datasets/kayrab/patient-doctor-qa-tr-5695.

Bulut, M. K. (2024c). Patient Doctor Q&A TR 95588. Erişim adresi: https://huggingface.co/datasets/kayrab/patient-doctor-qa-tr-95588.

Bulut, M. K. (2024d). Patient Doctor Q&A TR 19583. Erişim adresi: https://huggingface.co/datasets/kayrab/patient-doctor-qa-tr-19583.

Bulut, M. K. (2024e). Patient Doctor Q&A TR 167732. Erişim adresi: https://huggingface.co/datasets/kayrab/patient-doctor-qa-tr-167732.

Chikhaoui, E., Alajmi, A., & Larabi-Marie-Sainte, S. (2022). Artificial intelligence applications in healthcare sector: ethical and legal challenges. Emerging Science Journal, 6(4), 717–738.

Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., & Zhou, D. (2024). Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference. arXiv preprint arXiv:2403.04132 [cs.AI].

Chen, Y., Nayman, N., Greenfeld, D., Gal, Y., & Berant, J. (2022). Towards learning universal hyperparameter optimizers with transformers. Advances in Neural Information Processing Systems, 35, 32053–32068.

Dettmers, T., Lewis, M., Shleifer, S., & Zettlemoyer, L. (2021). 8-bit optimizers via block-wise quantization. arXiv preprint arXiv:2110.02861.

Devlin, J. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Elo, A. E., & Sloan, S. (1978). The rating of chessplayers: Past and present. New York: Arco Pub.

Google. (2024a). Gemini: Google’s AI Model for Multimodal Understanding. Erişim adresi: https://deepmind.google/technologies/gemini/pro/ (Erişim tarihi 16/08/2024).

Google. (2024b). Google Colab. Erişim adresi: https://colab.google/ (Erişim tarihi 08/09/2024)..

Henry41. (2024). iCliniq Medical QA Dataset. Erişim adresi: https://www.kaggle.com/datasets/henry41148/icliniq-medical-qa.

Hermansyah, I. D. (2024). Doctor-ID-QA Dataset. Erişim adresi: https://huggingface.co/datasets/hermanshid/doctor-id-qa.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., de Las Casas, D., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., van den Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., Rae, J. W., Vinyals, O., & Sifre, L. (2022). Training compute-optimal large language models. arXiv preprint arXiv:2203.15556.

Kesgin, H. T., Yuce, M. K., Dogan, E., Uzun, M. E., Uz, A., Seyrek, H. E., Zeer, A., & Amasyali, M. F. (2024). Introducing cosmosGPT: Monolingual Training for Turkish Language Models.

Meta AI. (2024). LLaMA 3.1: Meta’s Next-Generation Large Language Model. Erişim adresi: https://huggingface.co/meta-llama/Meta-Llama-3.1-70B (Erişim tarihi 08/08/2024).

meta-llama. (2024). meta-llama/Meta-Llama-3-8B. Erişim adresi: https://huggingface.co/meta-llama/Meta-Llama-3-8B (Erişim tarihi 16/08/2024).

Microsoft. (2024). GitHub Copilot: AI-Powered Code Completion by Microsoft. Erişim adresi: https://copilot.microsoft.com/ (Erişim tarihi 16/08/2024).

NVIDIA. (2024). NVIDIA A100 Tensor Core GPU. Erişim adresi: https://www.nvidia.com/tr-tr/data-center/a100/ (Erişim tarihi 08/08/2024).

OpenAI. (2024a). GPT-3.5 Turbo. Erişim adresi: https://platform.openai.com/docs/models/gpt-3-5-turbo (Erişim tarihi 15/07/2024).

OpenAI. (2024b). GPT-4o: OpenAI’s Latest Language Model. Erişim adresi: https://openai.com/index/hello-gpt-4o/ (Erişim tarihi 16/08/2024).

Park, C.-W., Seo, S. W., Kang, N., Ko, B., Choi, B. W., Park, C. M., Chang, D. K., Kim, H., Kim, H., Lee, H., Jang, J., Ye, J. C., Jeon, J. H., Seo, J. B., Kim, K. J., Jung, K.-H., Kim, N., Paek, S., Shin, S.-Y., ... Yoon, H.-J. (2020). Artificial intelligence in health care: Current applications and issues. Journal of Korean Medical Science, 35(42), e379. https://doi.org/10.3346/jkms.2020.35.e379

Peng, Y., Yan, S., & Lu, Z. (2019). Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets. arXiv preprint arXiv:1906.05474.

Sambanovasystems. (2024). sambanovasystems/SambaLingo-Turkish-Chat. Erişim adresi: https://huggingface.co/sambanovasystems/SambaLingo-Turkish-Chat (Erişim tarihi 16/08/2024).

Trendyol. (2024). Trendyol/Trendyol-LLM-7b-chat-v1.8. Erişim adresi: https://huggingface.co/Trendyol/Trendyol-LLM-7b-chat-v1.8 (Erişim tarihi 16/08/2024).

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Canton Ferrer, C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., & Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.

Unsloth. (2024). Unsloth: Finetune Llama 3.1, Mistral, Phi & Gemma LLMs 2–5x faster with 80% less memory. Erişim adresi: https://github.com/unslothai/unsloth (Erişim tarihi 08/08/2024).

Vaswani, A. (2017). Attention is all you need. Advances in Neural Information Processing Systems.

Wu, S., & Sun, M. (2022). Exploring the efficacy of pre-trained checkpoints in text-to-music generation task. arXiv preprint arXiv:2211.11216.

ytu-ce-cosmos. (2024). ytu-ce-cosmos/Turkish-Llama-8b-v0.1. Erişim adresi: https://huggingface.co/ytu-ce-cosmos/Turkish-Llama-8b-v0.1 (Erişim tarihi 16/08/2024).

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). BERTScore: Evaluating text generation with BERT. arXiv preprint arXiv:1904.09675.