KSÜ Mühendislik Bilimleri Dergisi, 28(2), 2025
Araştırma Makalesi

KSU J Eng Sci, 28(2), 2025
Research Article

Kahramanmaras Sutcu Imam University
Journal of Engineering Sciences

Geliş Tarihi : 05.01.2025
Kabul Tarihi :23.03.2025

Received Date : 05.01.2025
Accepted Date : 23.03.2025

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 BULUT[1] (ORCID: 0009-0000-3107-7121)
Banu DİRİ[2*] (ORCID: 0000-0002-6652-4339)

[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 kullanılan 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 değerlendirmektedir. Modeller, 321.179 hasta-doktor soru-cevap çiftinden oluşan Patient Doctor Q&A TR 321179 veri kümesi üzerinde ince ayar yapılarak eğitilmiştir. Performans ölçümünde BLEU ve BERT skor gibi sentetik değerlendirmelerin yanı sıra, Elo puanlaması ile uzman doktorların yanıt kalitesi incelemeleri de kullanılmıştır. Sonuçlar, doktor-LLama2-sambanovasystems-7b modelinin genel başarı bakımından en iyi performansı sergilediğini göstermiş, bu model uzman doktor incelemelerinden de 3.25 puan almıştır. Öte yandan, doktor-Mistral-trendyol-7b modeli %18,4 ile en düşük zararlı yanıt oranına sahip model olarak öne çıkmıştır. Bu çalışma, Türkçe sağlık hizmetlerinde yapay zekâ 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 evaluates the performance of four different large language models used in Turkish healthcare consultancy: doctor-meta-llama-3-8b, doctor-LLama2-sambanovasystems-7b, doctor-Mistral-trendyol-7b, and doctor-llama-3-cosmos-8b. The models were fine-tuned using the Patient Doctor Q&A TR 321179 dataset, which consists of 321,179 patient-doctor question-answer pairs. Performance was measured using synthetic evaluations such as BLEU and BERT scores, as well as expert doctor reviews of response quality through Elo scoring. The results showed that the doctor-LLama2-sambanovasystems-7b model demonstrated the best overall performance, receiving a score of 3.25 from expert doctor evaluations. On the other hand, the doctor-Mistral-trendyol-7b model stood out with the lowest harmful response rate at 18.4%. This study highlights the potential of AI-powered 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

ToCite: BULUT, M. K., & DİRİ, B., (2025). TÜRKÇE SAĞLIK DANIŞMANLIĞINDA BÜYÜK DİL MODELLERİNİN HASTA-DOKTOR İLETİŞİMİNDE KULLANIM POTANSİYELİ. *Kahramanmaraş Sütçü İmam Üniversitesi Mühendislik Bilimleri Dergisi*, 28(2), 802-822.

GİRİŞ

Günümüzde sağlık hizmetlerinde hasta-doktor iletişiminin yetersizliği, tıbbi bilgilere erişimdeki zorluklar ve sağlık profesyonellerinin iş yükünün fazlalığı önemli sorunlar olarak karşımıza çıkmaktadır. Bu sorunlar, hastaların doğru ve zamanında sağlık hizmetine erişimini kısıtlamakta, tedavi süreçlerini olumsuz etkilemekte ve sağlık sisteminin verimliliğini düşürmektedir.

Son zamanlarda yapay zeka ve doğal dil işleme alanındaki ilerlemeler, Büyük Dil Modelleri (Large Language Model - LLM) gibi teknolojilerin ortaya çıkmasını ve gelişimini hızlandırmıştır (Brown, 2020; Vaswani, 2017). 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 vd., 2020). Sağlık hizmetlerinde doğru ve hızlı bilgi paylaşımının hayati öneme hâiz olması, bu alanda LLM'lerin kullanımını daha da önemli hale getirmektedir (Jiang vd., 2017).

Sağlık alanında LLM'lerin kullanımı, hasta-doktor iletişiminin kolaylaştırılması, tıbbi bilgilerin daha anlaşılır hale getirilmesi ve sağlık uzmanlarına destek sağlanması gibi çeşitli avantajlar vadetmektedir (Chikhaoui vd., 2022). Ancak, bu teknolojinin sağlık gibi hayati bir alanda kullanılabilmesi için, modellerin özel olarak eğitilmeleri ve performanslarının ayrıntılı bir biçimde değerlendirilmesi lazımdır (Peng vd., 2019). BERT bazlı değerlendirme yöntemleri, metin benzerliğini anlamlandırmada başarılı sonuçlar vermiştir (Zhang vd., 2019). Özellikle farklı modellerin performansını kıyaslamalı analiz yoluyla incelemek, model seçimi ve iyileştirmesi konusunda yol gösterici olmaktadır (Chiang vd., 2024).

LLM'lerin büyük boyutları ve yüksek enerji ile işlem gücü tüketimi, sağlık hizmetlerinde uygulanabilirliklerini sınırlayan faktörlerdendir. Bu nedenle, daha az parametreye sahip, enerji ve işlem gücü açısından daha verimli, işe özel LLM'lerin oluşturulması hedeflenmektedir. Parametre sayısının az olması, modellerin daha az işlem gücü gerektirmesi ve böylece enerji tüketiminin azaltılması anlamına gelmektedir. Bu durum, sağlık hizmetlerinde yapay zeka uygulamalarının daha sürdürülebilir ve erişilebilir olmasına katkı sağlayacaktır.

Bu araştırmanın önemi, hasta-doktor iletişiminin geliştirilmesi ihtiyacı, tıbbi bilgilerin daha anlaşılır hale getirilmesi gerekliliği ve sağlık uzmanlarına destek sağlanması gereğinden kaynaklanmaktadır. Ayrıca, Türkçe sağlık alanında yapay zeka destekli çözümlerin eksikliği göz önüne alındığında, bu çalışma alandaki önemli bir boşluğu doldurmayı hedeflemektedir. Özelleştirilmiş ve daha küçük parametreli modellerin geliştirilmesiyle, hem doğruluk hem de enerji verimliliği açısından avantajlı sistemlerin oluşturulması amaçlanmaktadır.

Bu çalışmanın temel amacı, Türkçe sağlık verileri üzerinde Büyük Dil Modellerinin (LLM) ince ayar performansını inceleyerek literatüre özgün katkılar sağlamayı amaçlamaktadır. Çalışmanın başlıca faydaları şunlardır:

- **Hasta-doktor iletişimi alanına özel LLM'ler oluşturmak:** Genel amaçlı LLM'lerin yerine, hasta-doktor iletişimine odaklanan özelleştirilmiş modeller geliştirilmiştir. Bu sayede, iletişimin hassasiyeti ve doğruluğu artırılmıştır.

- **Geliştirilen LLM'lerin kapsamlı kıyaslamasını yapmak:** Dört farklı modelin performansı, çeşitli ölçütler ve değerlendirme yöntemleri kullanılarak detaylı bir şekilde karşılaştırılmıştır. Bu kapsamlı analiz, modellerin güçlü ve zayıf yönlerinin belirlenmesine olanak tanımıştır.

- **Özgün hasta-doktor iletişimi veri kümesi oluşturmak:** Çalışma kapsamında, 321.179 soru-cevap çiftinden oluşan geniş ve özgün bir Türkçe hasta-doktor iletişimi veri kümesi oluşturulmuştur. Bu veri kümesi, gelecekteki araştırmalar için önemli bir kaynak niteliğindedir.

- **Çok yönlü değerlendirme yaklaşımı geliştirmek:** Modellerin performansı, sentetik testler, yapay zeka hakemliği ve uzman değerlendirmesi olmak üzere üç ayrı ölçüm yöntemiyle değerlendirilmiştir. Bu çok yönlü yaklaşım, modellerin gerçek dünya uygulamalarındaki başarısını daha güvenilir bir şekilde ölçmeye olanak tanımıştır.

- **Enerji ve donanım verimliliğini artırmak:** Trilyon seviyesindeki parametre sayısına sahip devasa modeller yerine, onların yaklaşık %0,4'ü kadar parametre içeren daha küçük ve verimli modeller kullanılmıştır. Bu sayede, enerji tüketimi ve donanım gereksinimleri önemli ölçüde azaltılmıştır.

Bu çalışma kapsamda, 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) modelleri üzerinde ince ayar yapılmış ve doktor-llama-3-cosmos-8b, meta-llama-3-8b, doktor-LLama2-sambanovasystems-7b ve doktor-Mistral-trendyol-7b büyük dil modelleri oluşturulmuştur. Modellerin performansları, BLEU, BERT skor ölçütleri kullanılarak ölçülmüş, bunun yanında uzman değerlendirmeleri ve yapay zeka hakemliğinde kıyaslamalı analizler yapılmıştır. Uzman değerlendirmelerinin istatistiksel anlamdaki karşılıkları da ANOVA ve CRONBACH ölçütleriyle analiz edilmiştir. Bu modeller seçilirken LLM Turkish Leaderboard'daki sıralamaları ve farklı mimarilere sahip olmaları göz önünde bulundurulmuştur (LMSYS Chatbot Arena Leaderboard, 2024). Makalenin ana hedefi, LLM'lerin hasta-doktor ilişkilerindeki potansiyelini ortaya koymaktır. Bu doğrultuda, modellerin tıbbi bilgileri hastaların daha kolay anlayabileceği bir dille açıklama, hasta sorularını yanıtlama ve sağlık çalışanlarına yardımcı olma gibi görevlerdeki performansları incelenmiştir. Sağlık hizmetlerinde doğru ve hızlı bilgi paylaşımının hayati önemi göz önünde bulundurularak, Türkçe sağlık alanında özelleşmiş yapay zeka çözümlerinin geliştirilmesi amaçlanmıştır.

Bununla birlikte, yapay zeka uygulamalarının sağlık sektöründe kullanımı, etik ve yasal zorlukları da beraberinde getirmektedir (Chikhaoui vd., 2022). Özellikle hasta mahremiyeti, veri güvenliği ve tıbbi bilgilerin doğruluğu gibi konular, bu teknolojilerin uygulanmasında önemli etik sorunlar oluşturabilir. Bu nedenle, araştırmada modellerin etik açıdan kullanılabilirliği de ortaya konulmuştur.

LİTERATÜR ÖZETİ

Büyük Dil Modellerinin (LLM) sağlık verileriyle ilgili uygulamaları, yapay zekanın son yıllardaki popülerleşmesiyle birlikte araştırmacılar tarafından yoğun ilgi gören önemli bir alan haline gelmiştir. Bu alanda LLM'lerin sağlık verilerinde kullanımı ve performansı (Peng vd., 2019), hasta-doktor iletişimindeki rolü (Park vd., 2020), tıbbi bilgi çıkarımı ve özetlemeleri (Akyon vd., 2021), klinik karar destek sistemleri, etik ve yasal konular (Chikhaoui vd., 2022), Türkçe LLM'ler (Bulut ve Diri, 2024; Kesgin vd., 2024) ve LLM'lerin performans ölçütleri gibi önemli konular ön plana çıkmaktadır.

Yapılan önemli araştırmalar arasında, Peng ve arkadaşlarının elektronik sağlık kayıtlarından klinik anlam çıkarma başarısını inceleyen çalışması bulunmaktadır (Peng vd., 2019). Bu çalışmada, önceden eğitilmiş dil modellerinin biyomedikal doğal dil işleme görevlerindeki performansları değerlendirilmiştir. Sonuçlar, BERT ve ELMo gibi modellerin, sağlık verilerinde başarılı bir şekilde kullanılabileceğini göstermiştir. Bu da LLM'lerin sağlık alanındaki potansiyelini ortaya koymaktadır.

Park ve arkadaşları, yapay zekanın sağlık hizmetlerindeki mevcut uygulamalarını ve karşılaşılan sorunları ele almıştır (Park vd., 2020). Özellikle yapay zeka teknolojilerinin tıbbi görüntü analizi, akıllı IoT cihazları, sinyal ve in-vitro tanı analizleri ve elektronik sağlık kayıtları gibi alanlardaki kullanımını incelemişlerdir. Çalışma, yapay zeka'nın sağlık sektöründe giderek daha fazla yer aldığını ve bu teknolojilerin etkili bir şekilde entegre edilmesi için klinik yaklaşımların ve yasal düzenlemelerin dikkate alınması gerektiğini vurgulamaktadır.

Yıldız ve Alper tarafından yapılan araştırma, ChatGPT'nin sağlık alanında Türkçe ve İngilizce yanıtlarını karşılaştırmaktadır (Yıldız ve Alper, 2023). Beş gastroenteroloji uzmanının değerlendirmesine göre, ChatGPT'nin Türkçe yanıtları İngilizce kadar doğru ancak daha az kapsamlı bulunmuştur. Buna rağmen yanıtlar "yetersiz" kategorisinde değerlendirilmemiştir. Çalışma, yapay zeka dil modellerinin Türkçe tıbbi amaçlar için kullanılabilirliğini göstermektedir.

Singhal ve arkadaşları tarafından geliştirilen Med-PaLM 2 çalışması, tıbbi LLM'lerin uzman düzeyinde performans potansiyelini göstermiştir (Singhal vd., 2025). PaLM 2 temel alınarak tıbbi alanda özel olarak ince ayarlanmış bu model, "ensemble refinement" ve "chain of retrieval" stratejileriyle geliştirilmiştir. MedQA testinde %86,5 başarı oranıyla önceki versiyonundan %19 daha iyi performans sergilemiştir. Klinik değerlendirmelerde, doktorlar sekiz klinik eksenden dokuzunda Med-PaLM 2'nin cevaplarını insan doktorlarınkine tercih etmişlerdir. Gerçek dünya tıbbi sorularına yönelik pilot çalışmada, uzmanlar Med-PaLM 2'yi %65 oranında genel pratisyenlere tercih etmiş ve cevapların güvenliği konusunda olumlu değerlendirmeler yapmışlardır. Bu çalışma, Türkçe sağlık danışmanlığı için geliştirilecek LLM'lere önemli bir referans sunmaktadır.

Labrak ve arkadaşlarının geliştirdiği "BioMistral" projesi, medikal alan için özelleştirilmiş açık kaynaklı dil modellerinden oluşan bir koleksiyondur (Labrak vd., 2024). Mistral tabanlı ve PubMed Central üzerinde ileri eğitimli bu model, 10 farklı tıbbi soru cevaplama görevinde mevcut açık kaynaklı modellerden daha iyi performans göstermiştir. Model 7 farklı dilde test edilmiş, doğruluk ve kalibrasyon analizleri yapılmıştır. BioMistral, Türkçe sağlık danışmanlığı modellerine içgörüler sunmakta ve çokdilli medikal doğal dil işleme uygulamaları için önemli bir referans oluşturmaktadır.

Li ve arkadaşlarının geliştirdiği "Agent Hospital" çalışması, hastaların, hemşirelerin ve doktorların tümünün büyük dil modelleri tarafından desteklenen otonom ajanlar olduğu sanal bir hastane simülasyonu sunmaktadır (Li vd., 2024). Bu simülasyon hastalık başlangıcından iyileşme sürecine kadar tüm tedavi aşamalarını kapsamaktadır. SEAL (Simulacrum-based Evolutionary Agent Learning) paradigması altında, doktor ajanlar binlerce hasta ajanını tedavi ederek ve tıbbi kaynakları inceleyerek zamanla gelişmektedir. Evrimleşen doktor ajanlar, on binlerce simüle hasta tedavisinden sonra MedQA kıyaslama testinde mevcut yöntemlerden daha iyi performans göstermiştir. Bu çalışma sağlık hizmetlerindeki yapay zeka uygulamalarına yenilikçi bir perspektif kazandırmaktadır.

Yılmaz'ın çalışması, Gemini ve ChatGPT 3.5 modellerinin blefarit (göz kapağı iltihabı) hakkında hasta eğitim materyali sunma yeteneğini değerlendirmiştir (Yılmaz, 2024). PEMAT aracı kullanılarak yapılan analizde, her iki model de hastalığın önemli yönlerini kapsamıştır. Ancak Flesch-Kincaid okunabilirlik skorları, hasta eğitim materyalleri için önerilen 60-70 aralığının oldukça altında kalmıştır (Gemini: 38,75; ChatGPT 3.5: 26,35). Bu sonuçlar, LLM'lerin sağlık bilgisi sunma potansiyeli taşıdığını, fakat mevcut içeriklerin hedef kitle için fazla karmaşık olabileceğini ve erişilebilirlik açısından geliştirilmesi gerektiğini göstermektedir.

Chikhaoui ve arkadaşları , sağlık sektöründe yapay zeka uygulamalarının etik ve yasal zorluklarını incelemişlerdir (Chikhaoui vd., 2022). Çalışmada, yapay zeka'nın tıp alanında kullanılmasının etik sorunları, özellikle veri gizliliği, hasta mahremiyeti ve sorumluluk konuları ele alınmıştır. Yazarlar, yapay zeka teknolojilerinin güvenli ve etik bir şekilde kullanılabilmesi için yasal çerçevelerin ve düzenlemelerin oluşturulmasının önemini vurgulamışlardır. Ayrıca, bu teknolojilerin sağlık hizmetlerinde etkin bir şekilde kullanılabilmesi için uluslararası işbirliğinin gerekliliğine dikkat çekmişlerdir.

Türkçe dil modelleri alanında, Bulut ve Diri, tamamen Türkçe verilerle eğitilmiş LLM tabanlı sanal doktor asistanlarının geliştirilmesini ele almışlardır (Bulut ve Diri, 2024). Bu çalışmada, LLama2, LLama3 ve Mistral tabanlı dört farklı model, Türkçe hasta-doktor yazılı iletişimindeki performansları açısından incelenmiştir. Sonuçlar, Türkçe sağlık hizmetlerinde yapay zeka destekli sanal doktor asistanlarının potansiyelini göstermekte ve Türkçeye özgü tıbbi sohbet robotlarının geliştirilmesine katkıda bulunmaktadır.

Benzer şekilde, Kesgin ve arkadaşları, Türkçe dil modellerinin geliştirilmesi ve değerlendirilmesi üzerine çalışmalar yapmışlardır (Kesgin vd., 2024). Bu araştırmada, Türkçe için özel olarak eğitilmiş cosmosGPT modelleri tanıtılmış ve farklı boyutlardaki Türkçe dil modellerinin performansları on farklı değerlendirme veri kümesinde incelenmiştir. Sonuçlar, Türkçe dil modellerinin performansını artırmak için özel olarak eğitilmiş modellerin, çok dilli modellere kıyasla daha iyi sonuçlar verdiğini göstermiştir.

LLM'lerin sağlık alanındaki uygulamalarının sabit soru-cevap sistemlerinin ötesine geçmesi gerektiğini gösteren önemli çalışmalar bulunmaktadır. Fan ve arkadaşları, 'AI Hospital' adlı çok ajanlı bir simülasyon çerçevesi geliştirerek, LLM'lerin gerçek klinik ortamlardaki performansını değerlendirmenin yeni bir yolunu sunmuştur. Bu sistem, doktorlar (başarısı ölçülmek istenen LLM'ler tarafından yönetilen) ve hastalar, muayene uzmanları ve baş hekimler (oyuncu olmayan karakterler – non-player characters - NPC) arasındaki dinamik etkileşimleri simüle etmektedir (Fan ve diğ, 2024).

Ayrıca, Akyon ve arkadaşları, Türkçe metinlerden otomatik soru üretimi ve soru cevaplama üzerine bir çalışma sunmuşlardır (Akyon vd., 2021). Çalışmada, mT5 modeli, Türkçe soru üretimi ve cevaplama için çoklu görevli bir ortamda eğitilmiştir. Bu model, TQuADv1, TQuADv2 ve XQuAD Türkçe bölümlerinde değerlendirilmiş ve çoklu görev yaklaşımının tek görevli ayara göre daha iyi performans gösterdiği belirlenmiştir. Bu çalışma, Türkçe doğal dil işleme alanında önemli bir katkı sağlamaktadır.

Bunun yanında Türkçe sağlık verileri oluşturma ve doğal dil işleme (NLP, Natural Language Processing) alanında önemli bir çalışma da, TurkMedNLI adlı Türkçe tıbbi doğal dil çıkarım veri kümesinin sunulmasıdır. Oğul ve

arkadaşları tarafından gerçekleştirilen bu çalışma, MedNLI veri kümesinin Türkçeye çevrilmesi ve Türkçe tıbbi doğal dil çıkarımı (Natural Language Inference - NLI) veri kümesinin oluşturulması üzerine odaklanmaktadır (Oğul vd., 2025). Büyük dil modellerini kullanarak tıbbi kısaltmaların doğru şekilde genişletilmesi, ölçü birimlerinin dönüştürülmesi ve klinik bağlamın korunması için özel bir çeviri ve işleme boru hattı geliştirilmiştir. Bu yaklaşım, Türkçe tıbbi Doğal Dil İşleme araştırmalarında veri eksikliği sorununu ele alarak, düşük kaynaklı diller için kaliteli tıbbi veri kümelerinin oluşturulmasına katkı sağlamaktadır.

Literatürde, Uçar ve arkadaşları tarafından gerçekleştirilen önemli bir çalışma bulunmaktadır (Ucar vd., 2025). Bu çalışmada, büyük dil modellerinin tıbbi soru cevaplama görevleri için ince ayar yapılarak performanslarının artırılması incelenmiştir. RoBERTa ve BERT gibi modeller, Healthline.com sitesinden elde edilen 6.800 tıbbi soru-cevap örneği kullanılarak eğitilmiş ve farklı modellerin performansları karşılaştırılmıştır. Sonuçlar, BERT Large Uncased modelinin en yüksek başarıyı elde ettiğini ve büyük dil modellerinin tıbbi uygulamalarda etkili bir şekilde kullanılabileceğini göstermiştir. Bu çalışma, büyük dil modellerinin tıbbi alanlarda uyarlanmasının önemini vurgulamakta ve Türkçe sağlık verileri üzerinde gerçekleştirdiğimiz çalışmayla benzerlik taşımaktadır.

Güneş ve Ülkir, çoklu kipli (multimodal) LLM`lerin görsel nöroanatomi sorularındaki performanslarını radyolog ve anatomistlerle karşılaştırmalı olarak değerlendirmiştir. GPT4-V, GPT-4o, LLaVA ve Gemini 1.5 Flash modellerinin 100 adet görsel nöroanatomi sorusundaki doğruluk oranları incelenmiştir. Sonuçlara göre, radyolog %90 doğruluk oranıyla en yüksek performansı sergilerken, anatomist %67 doğruluk oranına ulaşmıştır. çoklu kipli LLM'ler arasında en iyi performansı %45 doğruluk oranıyla GPT-4o göstermiş, onu %35 ile Gemini 1.5 Flash, %22 ile GPT4-V ve %15 ile LLaVA takip etmiştir. Bu çalışma, multimodal LLM'lerin tıp alanında önemli bir potansiyele sahip olduğunu ancak nöroanatomik bölgeleri doğru tanımlama konusunda henüz tıbbi uzmanların doğruluk seviyesine ulaşamadığını ortaya koymaktadır (Güneş ve Ülkir, 2024).

Literatürden elde edilen bulgular, LLM'lerin sağlık alanında çok önemli bir potansiyele sahip olduğunu ortaya koymaktadır. Bu teknolojiler, hasta-doktor iletişimini geliştirme, tıbbi veri analizi ve yorumlanması, otomatik soru cevaplama ve daha pek çok alanda önemli katkılar sağlayabilir. Ancak, bu teknolojilerin güvenli ve etik kullanımı için veri gizliliği, etik ilkeler ve yasal düzenlemeler gibi konularda daha fazla araştırma ve düzenlemeye ihtiyaç duyulmaktadır.

MATERYAL VE YÖNTEM

Bu bölümde araştırmanın temelini oluşturan veri kümesinin oluşturulma sürecinden, verinin temizlenmesi ve derleme aşamalarından bahsedilecektir.

Veri Kümesi ve Ön İşleme

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, 289.063 adet eğitim ve 32.116 adet test verisi olmak üzere toplam 321.179 soru-cevap çiftinden oluşmaktadır. Veri kümesinin içeriği, hastaların doktorlara sorduğu her türlü soruya karşılık ilgili doktorların verdiği cevaplardan oluşmaktadır. Bu veri kümesi dört farklı veri kümesinin birleştirilip karıştırılmasıyla oluşturulmuştur (Bulut, 2024a).

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. Toplam 19.583 veri içermektedir. 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. Bu veri kümesi 95.588 eğitim ve 11.949 test verisi olmak üzre toplam 107.537 veriden oluşmaktadır. 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. Bu veri kümesi 5.694 eğitim ve 633 test verisi olmak üzere toplam 6.327 veri içermektedir.

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 veri kümesiyse 167.732 eğitim ve 20.000 test verisi olmak üzre toplam 187.732 veriden oluşmaktadır. 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.

Dört farklı veri kümesinin birleştirilmesi sürecinde, veri kümelerinin yalnızca soru ve cevap kolonları alınarak işlem yapılmıştır. 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 olarak ayrılırken, kalan %10'luk kısım test verisi olarak kullanılmak üzere ayrılmıştır. Veri kümesinin güvenilirliğini ve kullanılabilirliğini artırmak maksadıyla hem eğitim hem de test verileri **Mersenne Twister** algoritması kullanılarak karıştırılmış ve son kontroller özenle yapılarak veri kümesinin bütünlüğü sağlanmıştır (Matsumoto ve Nishimura, 1998).

Tablo 1. Veri Kümesi Dağılımları

Veri Kümesi	İçerdiği Veri Miktarı	Patient Doctor Q&A TR 321179 Veri Kümesindeki Oranı (%)
Patient Doctor Q&A TR 19583	19.583	6,097
Patient Doctor Q&A TR 95588	107.537	33,482
Patient Doctor Q&A TR 5695	6.387	1,989
Patient Doctor Q&A TR 167732	187.732	58,45

Verilerin ana veri kümesi olan "Patient Doctor Q&A TR 321179" içindeki dağılımı Tablo 1'de verilmiştir. Veri kümesi eğitim ve test verisi olarak bölünürken de bu oranlar korunmuştur. Bununla beraber 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.

Veri kümesi yapısına örnek olması açısından şu soru cevap çiftleri verilebilir:

Soru 1: Merhaba, benim 21 aylık bir oğlum var. Doğduğundan beri karnının üzerinde ve bacak arasında lekeler var, sanki yanık lekesi gibi alalı. Normal doktor kontrolüne götürdük, doktor bir şey demedi ama ben çok rahatsız oluyorum bu durumdan. Bir hastalık belirtisi falan olabilir mi?

Cevap 1: Merhaba, doktorunuz eğer ciddi bir durum olsaydı bunun bilgisini size verirdi. Ancak tespit edemediği bir şey ise görmeden bir şey söylemek sağlıklı olmaz. Geçmiş olsun.

Soru 2: Merhaba hocam, babamda bel fıtığı var, fizik tedaviye gitti. Doktor romatizma da şüphelenip MR istedi, MR sonucunu size yazacağım, bilgilendirirseniz sevinirim. Solda sakroiliak eklem anterior superior kısmında sakral ve iliak yüzde minimal kontur düzensizliği ve subkondral minimal skleroz izlenmiş olup, dejeneratif değişiklikler lehine değerlendirilmiştir. Bunun dışında bilateral sakroiliak eklem aralık, ilişki ve konturları tabii izlenmiş olup, ekleme komşu kemiklerin morfoloji ve sinyal intensiteleri tabii izlenmiştir. İnceleme alanına giren sakral ve omber vertebralar ile iliak kemiklerin korteks devamlılıkları ile kortikal medüller sinyal intensite dağılımları doğaldır. İnceleme alanına giren kas ve yumuşak dokular, cilt-cilt altı yağlı dokular ve pelvik yapılar tabiidir.

Cevap 2: MR raporu önemli bir problem göstermiyor. Normal bir MR sayılabilir.

Soru 3: Merhaba, yaklaşık bir haftadır yaşadığınız çarpıntı şikayetiyle doktora gittim. EKG, kan tahlili vs temiz çıktı. Holter takıldı, 24 saat süreyle kalp atım hızı ort 105, max 140 şeklinde bir sonuç çıktı. HDL 32, LDL 135,

trigliserid olması gerekenden yüksek. Doktorum stresten olmuştur, önemli bir şey yok dedi. Ben tatmin olmadım, bu durum normal mi, aydınlatırsanız sevinirim. Teşekkür ederim. İyi çalışmalar.

Cevap 3: Geçmiş olsun. Holter takılıyken sizi rahatsız eden şikayetiniz yine oldu fakat holterde anormal bir şey çıkmadı ise endişe edecek bir şey yoktur. Fakat bazen çarpıntıya sebep olan kalp problemi tespit etmek zor olabilir. Eğer holter takılıyken şikayetiniz olmadı ise daha ileri araştırma veya takip gerekir. Çarpıntı şikayetiniz olduğunda EKG çekilebilirse kalp probleminden kaynaklanıp kaynaklanmadığı anlaşılır. Saygılarımla.

Veri kümesi bu verilerden çok daha uzun ya da çok daha kısa birçok soru-cevap çifti içermektedir. Yukarıdaki örneklem sırası, gerçek veri kümesindeki sıradan bağımsızdır.

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. 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 (Direct Preference Optimization) 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 vd., 2024). Model, Cultura-X veri kümesinin Türkçe bölümünden alınan 42 milyar token ile eğitilmiş olup, SFT (Supervised fine-tuning) 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.

- **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 fark edilir bir biçimde geliştirmektedir.

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 1×10^4 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 1×10^4 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).

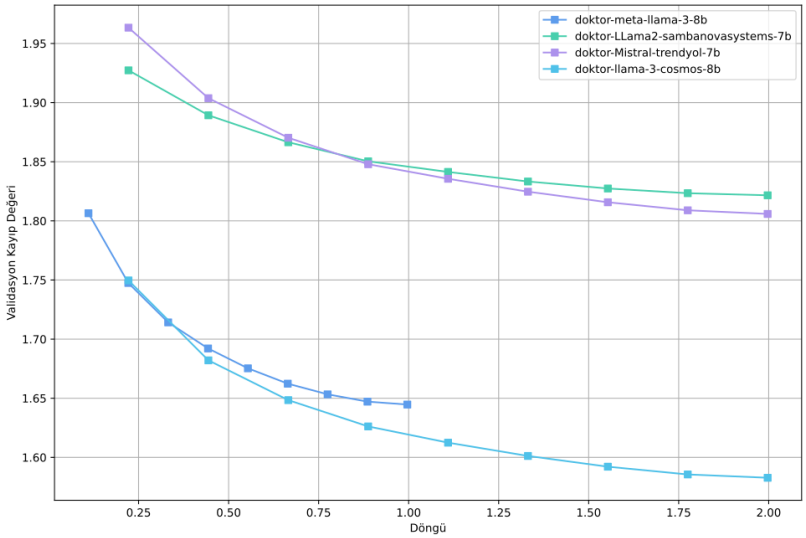
  Isınma adımları (Warmup Steps), eğitim sürecinin başlangıç aşamasında öğrenme oranını tedrici olarak artıran önemli bir teknik olup, bu çalışmada 2000 adım olarak belirlenmiştir. Bu teknik, eğitimin başlangıcında düşük öğrenme oranıyla başlayarak modelin ani ağırlık güncellemelerinden korunmasını sağlarken, 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.

  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 ve 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.



Şekil 1. Doğrulama Kayıp Eğrileri

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. Teknik kullanılırken rank değeri "r=8" olarak belirlenmiştir. 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 işlemlerin yapılmasında, 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).

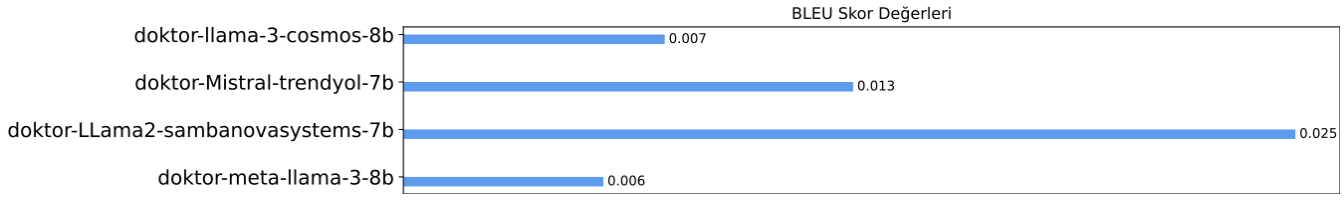
DENEYSEL SONUÇLAR

Bu bölümde, makale kapsamında gerçekleştirilen çalışmaların sonucunda elde edilen bulgular detaylı olarak sunulmaktadır. Araştırmanın temel amacı doğrultusunda, geliştirilen modellerin performansları farklı açılardan değerlendirilmiş ve karşılaştırmalı analizler yapılmıştır. Modellerin başarımları sentetik değerlendirme ölçütleri, yapay zeka hakemliğinde değerlendirme 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**

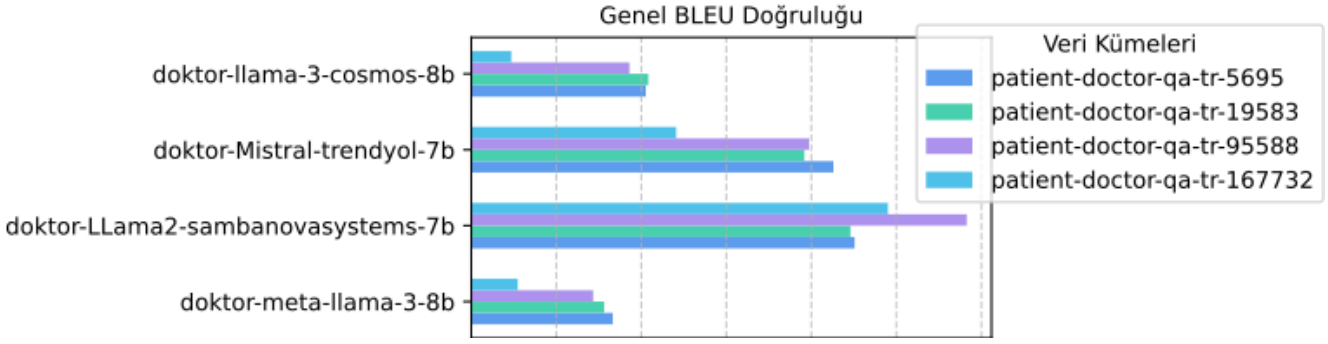


Şekil 2. Genel BLEU Skorları

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 0,025 BLEU skoru ile en yüksek performansı gösterirken, doktor-Mistral-trendyol-7b modeli 0,013 BLEU skoru ile ikinci sırada yer almıştır. Üçüncü sırada 0,007 BLEU skoru ile doktor-llama-3-cosmos-8b modeli bulunurken, doktor-meta-llama-3-8b modeli 0,006 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.

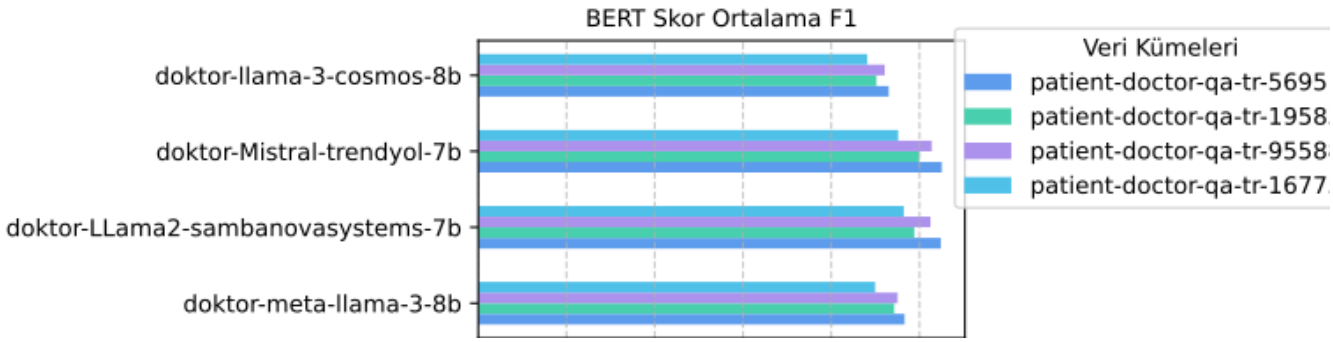
- **BERT Skor Sonuçları**

Şekil 4. BERT Skor F1 Değeri

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.

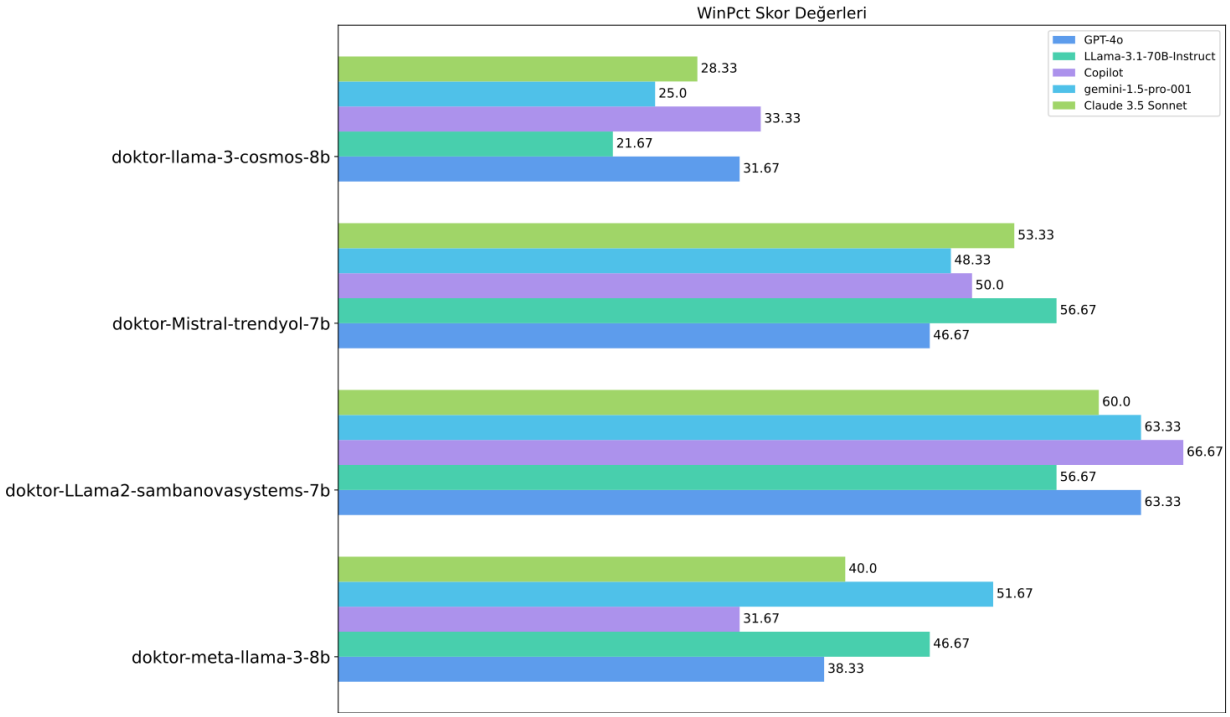
- **Yapay Zeka Hakemliğinde Değerlendirme**

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 (OpenAI, 2024c) 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 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.



Şekil 6. Yapay Zeka Hakemliğinde Kazanma Yüzdeleri

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 %25ile son sırada yer almıştır.

Claude 3.5 Sonnet (Anthropic, 2024) hakemliğinde de benzer bir sıralama görülmüş, doktor-LLama2-sambanovasystems-7b %60 ile en yüksek performansı göstermiştir. Doktor-Mistral-trendyol-7b %53,33, doktor-meta-llama-3-8b %40ve 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.

Bunun yanında, Tablo 2'deki veriler, farklı modellerin hastalara yönelik verdiği yanıtların "yararlı," "zararlı" ve "nötr" olmak üzere üç kategoride sınıflandırıldığını göstermektedir. Dikkate değer ilk bulgu, doktor-LLama2-sambanovasystems-7b modelinin en yüksek yararlı cevap oranına (%71,1) sahip olmasıdır. Bu, modelin hastaların sorularına çoğunlukla olumlu veya işe yarar nitelikte yanıt verebildiğini göstermesi bakımından önemlidir. Bununla birlikte aynı modelin %21,1 gibi hatırı sayılır bir zararlı cevap oranına sahip olması, özellikle sağlık gibi hassas bir alanda tek başına yeterince güvenli olmadığını da ortaya koymaktadır.

Tablo 2. Yarar-Zarar Oranı

Model	Yararlı Cevap (%)	Zararlı Cevap (%)	Nötr Cevap (%)
doktor-Mistral-trendyol-7b	69,4	18,4	12,2
doktor-meta-llama-3-8b	62,6	21,3	16,1
doktor-llama-3-cosmos-8b	42,1	33,8	24,1
doktor-LLama2-sambanovasystems-7b	71,1	21,1	7,8

Öte yandan doktor-Mistral-trendyol-7b modeli, yararlı cevap yüzdesi açısından (%69,4) ilk modele göre ufak bir farkla geride kalırken, zararlı cevap oranının %18,4 seviyesinde olması bakımından daha avantajlı bir konumda görünmektedir. Sağlık alanında "yanlış yönlendirici" veya "zararlı" addedilebilecek yanıtların sayısını olabildiğince azaltmak önemli olduğu için, bu model çok daha düşük bir risk profili sunmaktadır. Dolayısıyla hem toplamda epey yüksek bir yararlı cevap oranına sahip olması hem de nispeten az sayıda zararlı cevap üretmesi, onu uygulamada tercih edilebilir kılmaktadır.

doktor-meta-llama-3-8b modeline bakıldığında, yararlı cevap oranının %62,6 ile orta düzeyde kaldığı; zararlı cevap oranınınsa %21,3 düzeyine çıktığı görülür. Bu değerler, modeli önceki iki seçeneğe göre daha az güvenilir hâle getirmektedir. Son olarak doktor-llama-3-cosmos-8b modelinin %42,1 oranında yararlı cevaba kıyasla %33,8 zarar oranına sahip olması, bu modelin sağlık uygulamaları bağlamında iyileştirmeye en çok ihtiyaç duyan seçenek olduğu anlamına gelmektedir.

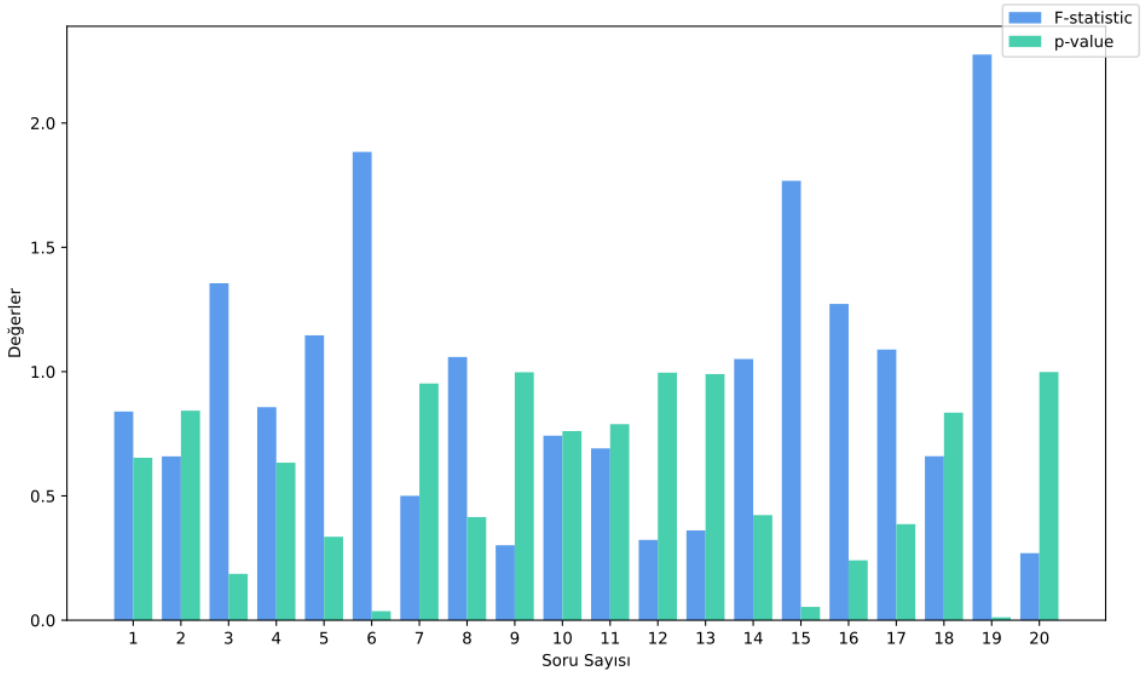
Genel olarak bakıldığında, sağlık sektöründe yapay zeka tabanlı danışmanlık veya destek sistemleri tasarlanırken yalnızca yüksek "yararlı cevap" oranına odaklanmak yeterli değildir; aynı zamanda "zararlı" veya yanıltıcı cevapların asgari düzeye indirilmesi şarttır. Nitekim tabloda da görüldüğü üzere, bazı modeller yüksek yararlılık sergilerken zararlı cevap oranlarını da yüksek tutabilmektedir. Bu nedenle karar vericiler, model seçiminde "en az zararlı" yanıtlara öncelik vermenin mi yoksa "en yüksek yararlılık" oranına ulaşmanın mı daha öncelikli olduğuna iyi karar vermelidir. Özellikle insan sağlığını ilgilendiren konularda, tek bir yanlış bilginin dahi ciddi sonuçlar doğurabileceği

göz önüne alındığında, risk yönetimini merkeze koyan bir yaklaşımın benimsenmesi öneme haizdir. sambanovasystems-7b modeli daha yüksek yararlı cevap oranına sahip olmasına rağmen sağlık gibi önemli bir alanda zaralı cevap verilmemesi daha elzemdir.

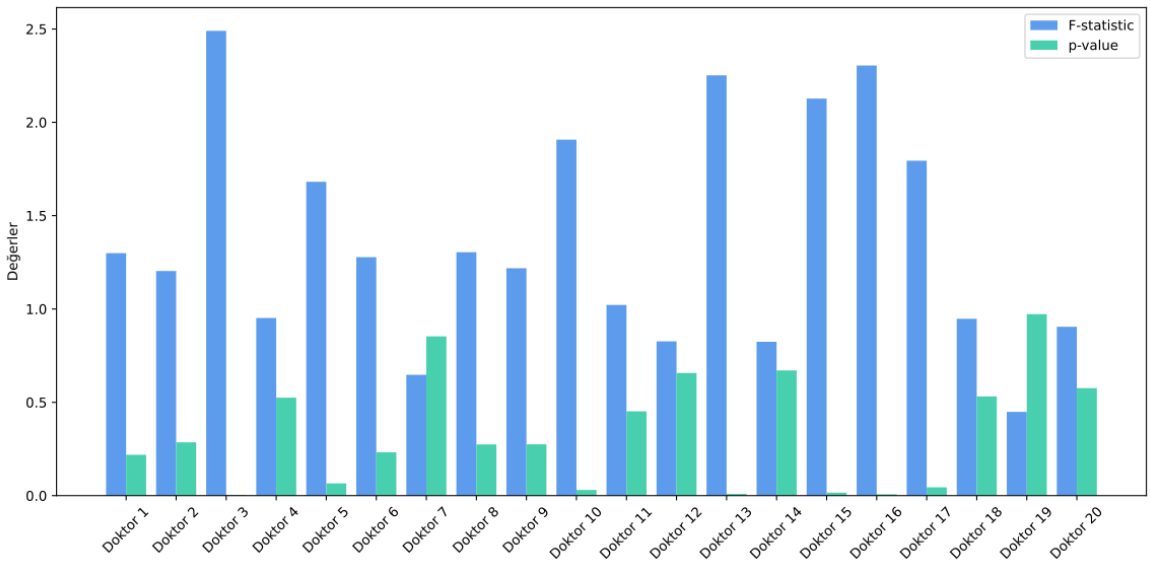
- **ANOVA Sonuçları**

Şekil 8'de 20 farklı doktorun değerlendirme sonuçları görülmektedir. F-istatistik bazı doktorlar için oldukça yüksektir. Özellikle Doktor 3, Doktor 13 ve Doktor 16'nın F-istatistik değerleri 2.0'ın üzerindedir. Bu durum doktorların değerlendirmeleri arasında istatistiksel olarak anlamlı farklılıklar olduğunu göstermektedir.

P-değerlerine baktığımızda, çoğu durumda 0,05'in altında olduğu görülmektedir. Bu da değerlendirmeler arasındaki farklılıkların istatistiksel olarak anlamlı olduğunu doğrulamaktadır. Özellikle Doktor 10, 15 ve 16'nın p-değerleri çok düşüktür.



Şekil 8. Soruların ANOVA Skorları

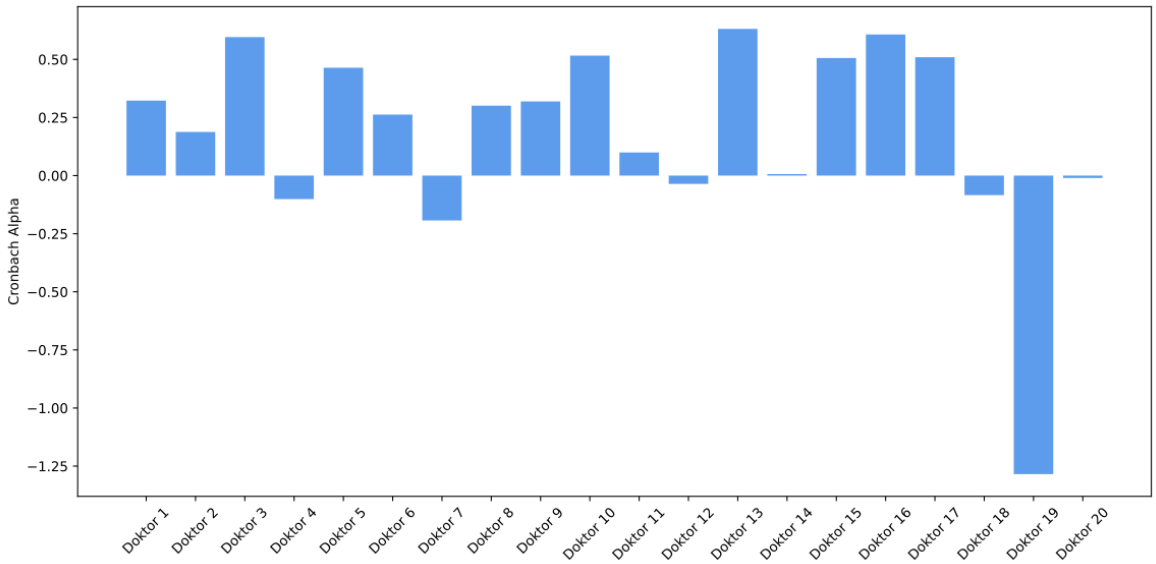


Şekil 9. Doktorların ANOVA Skorları

Şekil 9'da ise soru bazında ANOVA sonuçları gösterilmektedir. Soru 19'un yaklaşık 2,2 ile en yüksek F-istatistik değerine sahip olduğu görülmektedir. Bu durum, soruya verilen cevaplar arasında önemli farklılıklar olduğunu göstermektedir. Örneğin soru 12, 13 ve 20'nin 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.

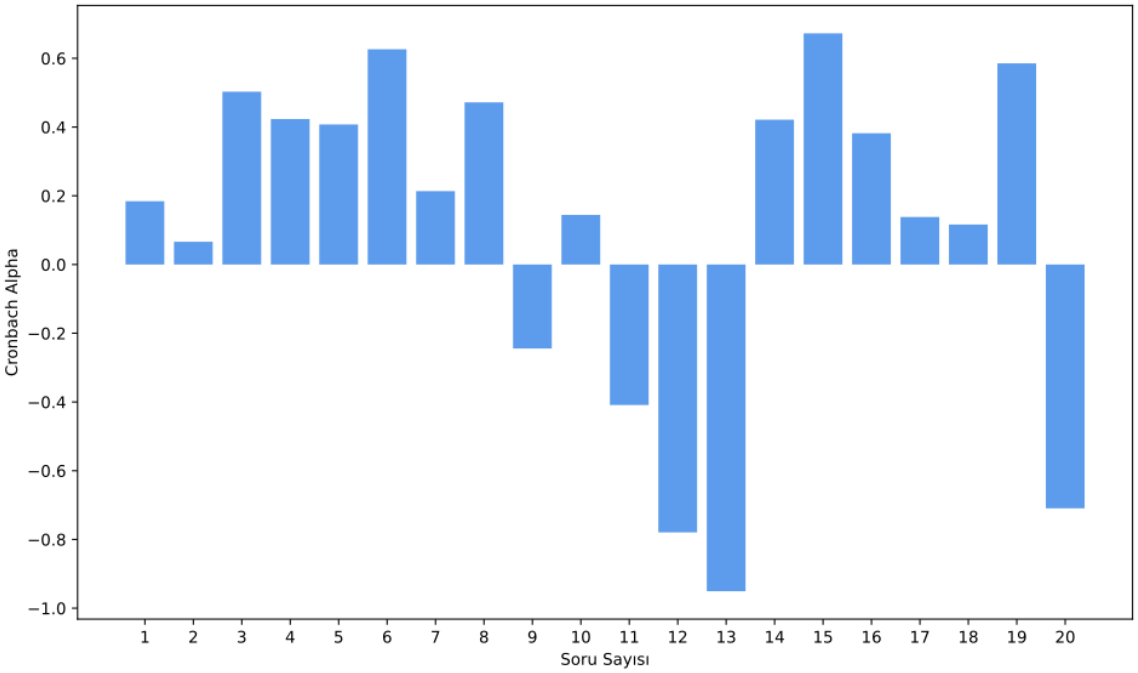
Şekil 11'de gösterilen soru bazındaki Cronbach Alfa değerleri analiz edildiğinde, en yüksek tutarlılığın yaklaşık 0,65 değeriyle 15. soruda, en düşük tutarlılığın ise yaklaşık -0,95 değeriyle 13. soruda görüldüğü tespit edilmiştir. Soruların yaklaşık yarısının pozitif, diğer yarısının negatif değerler 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.

- **Bulgular**

Yapılan denemeler sonucunda 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.



Şekil 11. Soruların CRONBACH Skorları

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.

Bu bulgular, özel alanlara özgü dil modellerinin genel amaçlı modellere göre daha iyi performans gösterdiğini desteklemektedir. Literatürde de benzer şekilde, alanına özel eğitilmiş modellerin performans avantajları vurgulanmaktadır. Örneğin, Peng ve arkadaşları, biyomedikal metinlerde önceden eğitilmiş dil modellerinin performansını incelemiş ve alanına özel eğitilen modellerin genel amaçlı modellere göre daha başarılı olduğunu göstermiştir (Peng vd., 2019). Benzer şekilde, Kesgin ve arkadaşları, Türkçe dil modellerinin geliştirilmesi ve değerlendirilmesinde alanına özel eğitilmiş modellerin daha iyi performans sergilediğini belirtmişlerdir (Kesgin vd., 2024).

doktor-meta-llama-3-8b modelinin genel amaçlı bir model olmasına rağmen makul bir performans göstermesi, genel amaçlı modellerin de belirli bir seviyede başarılı olabileceğini göstermektedir. Ancak, doktor-llama-3-cosmos-8b modelinin çoğu değerlendirme ölçütünde en düşük performansı sergilemesi, model ne kadar Türkçe için eğitilmiş olsa da, talimat odaklı eğitimin en az dil özelinde eğitim kadar önemli olduğunu göstermiştir. Bu durum, modellerin eğitiminde kullanılan veri kümesi büyüklüğü, kalitesi ve eğitim hedefinin görev özelindeki performans üzerinde önemli bir etkiye sahip olduğunu göstermektedir.

Bunlarla beraber, GPT-4o (OpenAI, 2024b) gibi yüksek parametre sayısına sahip genel amaçlı büyük dil modelleri, çok fazla işlem gücü ve VRAM gerektirmektedir. Bu durum, bu modellerin pratik uygulamalarda kullanımını sınırlayabilmektedir. Buna karşılık, bizim kullandığımız alanına özel ve daha az parametreye sahip modeller, daha az donanım gereksinimiyle çalışabilmekte ve enerji tüketimini azaltmaktadır. Böylece, daha verimli ve erişilebilir bir çözüm sunmaktadır.

SONUÇ, TARTIŞMA VE KISITLAMALAR

Sonuç

Bu çalışma, Türkçe sağlık verileri üzerinde ince ayar yapılan büyük dil modellerinin, hasta-doktor iletişiminin kalitesini artırmada ve tıbbi bilgiye daha güvenilir erişim sağlamada önemli bir potansiyel taşıdığını ortaya koymuştur. Elde edilen sonuçlar, dile ve alana özgü ince ayarın, genel amaçlı modellerin ötesinde somut avantajlar sunduğunu göstermektedir. İnsan uzman değerlendirmeleri, otomatik ölçütler ve yapay zeka hakemli karşılaştırmaların entegre bir şekilde kullanılması sayesinde, bu modellerin performansındaki belirgin artış net bir şekilde gözlemlenmiştir. Özellikle, doktor-LLama2-sambanovasystems-7b modeli genel başarı açısından öne çıkarken, doktor-Mistral-trendyol-7b modeli ise ürettiği içeriklerin zararlı olma riskinin düşük tutulması bakımından dikkat çekmiştir. Bu bulgular, yalnızca model çıktılarının nicel performansını değil, aynı zamanda sağlık alanında uygulanabilirliğin önemli boyutları olan hasta mahremiyeti, veri güvenliği, etik karar verme ve yasal sorumluluk gibi unsurları da kapsamlı bir şekilde değerlendirmeye olanak tanımıştır.

Çalışmamızın asıl katkısı, mevcut literatürde çoğunlukla benzer alanlarda gerçekleştirilen çalışmaların ötesine geçerek, Türkçe sağlık verilerinin özgün dil yapısı ve terminolojisinin dikkate alındığı, çok boyutlu bir değerlendirme yaklaşımının uygulanmasıdır. Bu yöntem, yalnızca model çıktılarının nicel ölçülerini sunmakla kalmayıp, sağlık alanında uygulanabilirliğin önemli boyutları olan hasta mahremiyeti, veri güvenliği, etik karar verme ve yasal sorumluluk gibi unsurları da tartışmaya açmıştır.

Bununla birlikte, elde edilen bulgular, LLM'lerin klinik bağlamda kullanılması için gerekli olan ince ayarın sağladığı avantajların yanı sıra, mevcut teknolojinin uygulanması sırasında karşılaşılan temel engelleri de gözler önüne sermektedir. Bu açıdan çalışma, yalnızca mevcut performansın tekrarı değil; aynı zamanda Türkçe sağlık alanında model adaptasyonunun, güvenlik ve etik gerekliliklerle birlikte nasıl iyileştirilebileceği konusunda yeni bakış açıları sunmaktadır.

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.

Tartışma

Çalışmamızın literatürdeki benzer araştırmalardan farkı ve üstünlüğü, kullanılan veri kümesinin kapsamı, model kıyaslama stratejilerinin derinliği, istatistiksel analiz yöntemlerinin kullanımı ve etik boyutun değerlendirilmesinde yatmaktadır.

Öncelikle, Uçar ve arkadaşlarının çalışmasında yaklaşıma dayalı olarak 6.800 örnek kullanılmasına karşın, çalışmamızda çeşitli kaynaklardan birleştirilen ve temizleme işlemleri sonrasında 321.179 soru-cevap çiftinden oluşan çok daha geniş ve zengin bir veri kümesi kullanılmıştır (Ucar vd., 2025). Bu durum, modellerin daha geniş bir örneklem üzerinde eğitilmesini ve dolayısıyla sonuçların genellenebilirliğini artırmaktadır. Bu çalışmada aynı mimariye sahip BERT tabanlı modeller kullanılıp karşılaştırılmıştır. Buna karşın, çalışmamızda sağlık verilerine özgü ince ayar gerçekleştirmek amacıyla, Meta-Llama-3-8B, SambaLingo-Turkish-Chat, Trendyol-LLM-7b-chat-v1.8 ve Turkish-Llama-8b-v0.1 gibi farklı mimarileri de içeren modeller, aynı veri kümesi üzerinde görev özelinde kıyaslandı. Böylece, hem aynı temel mimari üzerinde modellerin dil özelinde eğitilme başarımı hem de farklı mimarilerin başarımları detaylı bir şekilde ortaya konulabildi.

Ayrıca, pek çok çalışmada model performansları yalnızca temel ölçütler (doğruluk, BLEU, ROUGE vb.) üzerinden ölçülürken, biz çalışmamızda Elo puanlaması, kazanma yüzdesi gibi daha karmaşık işler için daha ölçücü olan

karşılaştırma yöntemlerinin yanı sıra, ANOVA, Cronbach alfa gibi istatistiksel yöntemleri de kullanarak, elde edilen sonuçların tutarlılığını ve güvenilirliğini ölçmeye yönelik bir değerlendirme çerçevesi oluşturduk. Bu vesileyle, modellerin performans farkları anlamlı istatistiksel delillerle desteklenmiş ve karşılaştırmalar daha sağlam temellere oturtulmuştur.

Diğer yandan, Chikhaoui ve arkadaşları, yapay zekanın sağlık alanındaki etik ve hukuki zorluklarını detaylı olarak ele almışlardır (Chikhaoui vd., 2022). Çalışmamız esas olarak modellerin teknik performansına odaklanmış olsa da, hasta-doktor iletişimi gibi hassas bir uygulama alanında, uzman değerlendirmeleri aracılığıyla model cevaplarının etik uygunluğu, yanlılık ve zarar üretme riskleri gibi konular da incelenmiştir. Böylece, hem sayısal hem de niteliksel değerlendirme yöntemlerini entegre ederek sağlık hizmetlerinde kullanılabilirliğe dair kapsamlı bir bakış açısı sunulmuştur.

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.

KAYNAKLAR

Akyon, F. C., Cavusoglu, D., Cengiz, C., Altinuc, S. O., & Temizel, A. (2021). Automated question generation and question answering from Turkish texts. arXiv preprint arXiv:2111.06476.

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).

LMSYS Chatbot Arena Leaderboard. (2024). LMSYS Chatbot Arena Leaderboard. Retrieved August 16, 2024, from https://chat.lmsys.org/?leaderboard

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.

Bulut, M. K., & Diri, B. (2024f). Artificial Intelligence Revolution in Turkish Health Consultancy: Development of LLM-Based Virtual Doctor Assistants. In 2024 8th International Artificial Intelligence and Data Processing Symposium (IDAP) (pp. 1–6). IEEE.

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.

Fan, Z., Tang, J., Chen, W., Wang, S., Wei, Z., Xi, J., ... & Zhou, J. (2024). AI Hospital: Benchmarking large language models in a multi-agent medical interaction simulator. arXiv preprint arXiv:2402.09742.

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).

Güneş, Y. C., & Ülkir, M. (2024). Comparative Performance Evaluation of Multimodal Large Language Models, Radiologist, and Anatomist in Visual Neuroanatomy Questions. Uludağ Üniversitesi Tıp Fakültesi Dergisi, 50(3), 551-556.

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.

Labrak, Y., Bazoge, A., Morin, E., Gourraud, P. A., Rouvier, M., & Dufour, R. (2024). Biomistral: A collection of open-source pretrained large language models for medical domains. arXiv preprint arXiv:2402.10373.

Li, J., Lai, Y., Li, W., Ren, J., Zhang, M., Kang, X., ... & Liu, Y. (2024). Agent hospital: A simulacrum of hospital with evolvable medical agents. arXiv preprint arXiv:2405.02957.

Matsumoto, M., & Nishimura, T. (1998). Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. ACM Transactions on Modeling and Computer Simulation (TOMACS), 8(1), 3-30.

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).

Oğul, İ. Ü., Soygazi, F., & Bostanoğlu, B. E. (2025). TurkMedNLI: a Turkish medical natural language inference dataset through large language model based translation. PeerJ Computer Science, 11, e2662.

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 Language Model. Erişim adresi: https://openai.com/index/hello-gpt-4o/ (Erişim tarihi 16/08/2024).

OpenAI. (2024c). GPT-4: OpenAI's Language Model. Erişim adresi: https://openai.com/index/gpt-4/ (Erişim tarihi 21/08/2024).

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).

Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Amin, M., ... & Natarajan, V. (2025). Toward expert-level medical question answering with large language models. Nature Medicine, 1-8.

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.

Ucar, A., Nayak, S., Roy, A., Taşcı, B., & Taşcı, G. (2025). A Comprehensive Study on Fine-Tuning Large Language Models for Medical Question Answering Using Classification Models and Comparative Analysis. arXiv preprint arXiv:2501.17190.

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.

Yıldız, M. S., & Alper, A. (2023). Can ChatGPT-4 diagnose in Turkish: a comparison of ChatGPT responses to health-related questions in English and Turkish. Journal of Consumer Health on the Internet, 27(3), 294-307.

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.