T.C.
YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

SAĞLIK VERİLERİ ÜZERİNDE LLM'LERİN İNCE AYAR PERFORMANSI

Muhammed Kayra BULUT

YÜKSEK LISANS TEZİ
Bilgisayar Bilimleri Anabilim Dalı
Bilgisayar Mühendisliği Programı

Danışman
Prof. Dr. Banu DİRİ

Şubat, 2025

T.C.
YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

SAĞLIK VERİLERİ ÜZERİNDE LLM'LERİN İNCE AYAR PERFORMANSI

Muhammed Kayra BULUT tarafından hazırlanan tez çalışması 01.02.2025 tarihinde aşağıdaki jüri tarafından Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Bilimleri Anabilim Dalı Bilgisayar Mühendisliği Programı **YÜKSEK LISANS TEZİ** olarak kabul edilmiştir.

Prof. Dr. Banu DİRİ
Yildiz Technical University
Danışman

Jüri Üyeleri

Prof. Dr. Banu DİRİ, Danışman
Yildiz Technical University _____

Prof. Dr. Name SURNAME, Üye
Yildiz Technical University _____

Doç. Dr. Name SURNAME, Üye
Yildiz Technical University _____

Danışmanım Prof. Dr. Banu DİRİ sorumluluğunda tarafımca hazırlanan SAĞLIK VERİLERİ ÜZERİNDE LLM'LERİN İNCE AYAR PERFORMANSI başlıklı çalışmada veri toplama ve veri kullanımında gerekli yasal izinleri aldığımı, diğer kaynaklardan aldığım bilgileri ana metin ve referanslarda eksiksiz gösterdiğimi, araştırma verilerine ve sonuçlarına ilişkin çarpıtma ve/veya sahtecilik yapmadığımı, çalışmam süresince bilimsel araştırma ve etik ilkelerine uygun davrandığımı beyan ederim. Beyanımın aksinin ispatı halinde her türlü yasal sonucu kabul ederim.

Muhammed Kayra BULUT

İmza

Gazze'de soykırım gören
mazlumlara...

TEŞEKKÜR

Tez süresince gece, gündüz, tatil demeden her zaman bana yardımcı olan tez hocam Banu DİRİ'ye, takıldığım yerlerde sorularıma cevap bulabildiğim Prof. Dr. Fatih Amasyalı, Arş. Gör. Himmet Toprak Kesgin'e ve Mehmet Ali Bayram'a teşekkürü borç bilirim.

Muhammed Kayra BULUT

İÇİNDEKİLER

KISALTMA LİSTESİ

BDM	Büyük Dil Modeli (Large Language Model)
BLEU	İki Dilli Değerlendirme Vekili (Bilingual Evaluation Understudy)
BP	Kısalık Düzeltme Faktörü (Brevity Penalty)
CER	Karakter Hata Oranı (Character Error Rate)
CNN	Evrimsel Sinir Ağı
DDİ	Doğal Dil İşleme
DPO	Doğrudan Tercih Optimizasyonu
GPT	Üretici Önceden Eğitilmiş Dönüştürücü (Generative Pre-trained Transformer)
GPU	Grafik İşleme Ünitesi
LLM	Large Language Model (Büyük Dil Modeli)
LoRA	Düşük Dereceli Uyarlama (Low-Rank Adaptation)
METEOR	Açık Sıralamaya Sahip Çeviri Değerlendirme Ölçütü (Metric for Evaluation of Translation with Explicit ORdering)
RNN	Tekrarlayan Sinir Ağı
ROUGE	Özetleme Değerlendirmesi için Anımsama Odaklı Vekil (Recall-Oriented Understudy for Gisting Evaluation)
SFT	Denetimli İnce Ayar
TPU	Tensor İşleme Ünitesi
UL2	Birleşik Dil Öğrenimi
VRAM	Video Rastgele Erişimli Bellek (Video Random Access Memory)
WER	Kelime Hata Oranı (Word Error Rate)
YZ	Yapay Zeka

ŞEKİL LİSTESİ

TABLO LİSTESİ

ÖZET

SAĞLIK VERİLERİ ÜZERİNDE LLM'LERİN İNCE AYAR PERFORMANSI

Muhammed Kayra BULUT

Bilgisayar Bilimleri Anabilim Dalı
Yüksek Lisans Tezi

Danışman: Prof. Dr. Banu DİRİ

Bu çalışma, Türkçe sağlık danışmanlığında dört farklı büyük dil modelinin (LLama2, LLama3 ve Mistral temelli) doktor-hasta yazılı iletişimindeki performanslarını incelemektedir [1–4]. Bu iletişim için hasta-doktor soru-cevap veri kümesi oluşturulmuştur ve modeller, bu veri kümesi üzerinde eğitilmiş ve ince ayar yapılmıştır [5]. Performans değerlendirmesi için kullanılan metrikler, ROUGE, Elo puanlaması [6], Kazanma yüzdesi ve Uzman değerlendirmesidir. Karşılaştırmalı analiz sonucunda, SambaLingo-Turkish-Chat modeli yanıt doğruluğu ve bağlama uygunluk açısından başarılıyken, Trendyol-LLM-7b-chat-v1.8 modeli işin etik kısmı da dikkate alınınca daha başarılı olmuştur. Bu çalışma, Türkçe sağlık hizmetlerinde yapay zeka destekli sanal doktor asistanlarının potansiyelini göstermekte ve Türkçe'ye özgü tıbbi sohbet robotlarının geliştirilmesine katkıda bulunmaktadır.

Anahtar Kelimeler: Doğal dil işleme, Tıbbi yapay zeka, Türkçe sağlık hizmetleri, BDM ince ayarı, Doktor-hasta iletişimi, Karşılaştırmalı model değerlendirmesi

YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

ABSTRACT

FINE TUNING PERFORMANCE OF LLMS ON HEALTH DATA

Muhammed Kayra BULUT

Department of Computer Science
Master of Science Thesis

Supervisor: Prof. Dr. Banu DIRI

This study examines the performance of four different large language models (based on LLama2, LLama3, and Mistral) in Turkish health counseling, specifically in written doctor-patient communication [1–4]. A patient-doctor question-answer dataset was created for this communication, and the models were trained and fine-tuned on this dataset [5]. The metrics used for performance evaluation include ROUGE, Elo rating [6], Win percentage, and Expert evaluation. As a result of the comparative analysis, while the SambaLingo-Turkish-Chat model was successful in terms of response accuracy and contextual relevance, the Trendyol-LLM-7b-chat-v1.8 model proved more successful when ethical considerations were taken into account. This study demonstrates the potential of AI-assisted virtual doctor assistants in Turkish healthcare services and contributes to the development of Turkish-specific medical chatbots.

Keywords: Natural language processing, Medical AI, Turkish healthcare, LLM fine-tuning, Doctor-patient communication, Comparative model evaluation

YILDIZ TECHNICAL UNIVERSITY
GRADUATE SCHOOL OF SCIENCE AND ENGINEERING

1
GİRİŞ

LLM'ler, doğal dil işleme alanında son yıllarda önemli gelişmeler kaydetmiş ve çeşitli uygulamalarda kullanılmaya başlanmıştır. Özellikle sağlık sektöründe, bu modellerin potansiyel kullanımı büyük ilgi görmektedir. Bu tez çalışması, sağlık verileri üzerinde LLM'lerin ince ayar (fine-tuning) performansını incelemeyi amaçlamaktadır. Sağlık hizmetlerinde doğru ve hızlı bilgi paylaşımı hayati önem taşımaktadır. Hasta-doktor iletişiminin geliştirilmesi, tıbbi bilgilerin daha anlaşılır hale getirilmesi ve sağlık profesyonellerine destek sağlanması gibi alanlarda LLM'lerin kullanımı, sağlık hizmetlerinin kalitesini artırma potansiyeli taşımaktadır. Lâkin, bu modellerin sağlık gibi hassas bir alanda kullanılabilmesi için, özel olarak eğitilmeleri ve performanslarının detaylı bir şekilde değerlendirilmesi gerekmektedir. Bu noktada, LLM'lerin sağlık alanında kullanımıyla ilgili etik ve hukuki konuların da dikkatle ele alınması gerekmektedir. Hasta mahremiyeti, veri güvenliği ve tıbbi bilgilerin doğruluğu gibi konular, bu teknolojilerin uygulanmasında önemli etik sorunlar oluşturabilir. Ayrıca, LLM'lerin sağlık alanında kullanımı, hekimlerin mesleki sorumluluklarını ve tıbbi uygulamaların yasal çerçevesini etkileyebilecek yeni hukuki sorunları da beraberinde getirebilir. Etik açıdan bakıldığında, LLM'lerin sağlık verilerini işlerken ve hasta-doktor iletişiminde kullanılırken, erdem etiği, ödev etiği ve sonuçsalcı etik gibi farklı etik yaklaşımların göz önünde bulundurulması gerekmektedir. Örneğin, bu modellerin kullanımında dürüstlük, adalet, doğruluk ve hasta haklarına saygı gibi etik ilkelerin korunması büyük önem taşımaktadır. Hukuki açıdan ise, LLM'lerin sağlık alanında kullanımı, mevcut yasal düzenlemelerin yeniden değerlendirilmesini gerektirebilir. Özellikle kişisel sağlık verilerinin korunması, hasta hakları ve tıbbi uygulamaların yasal sorumluluğu gibi konularda yeni düzenlemelere ihtiyaç duyulabilir. Bu bağlamda, LLM'lerin kullanımıyla ilgili yasal çerçevenin net bir şekilde belirlenmesi, hem hastaların hem de sağlık profesyonellerinin haklarının korunması açısından kritik öneme sahiptir. Ayrıca, şu an için LLM'lerin sağlık alanında kullanımının, yalnızca bu teknolojileri ve sağlık sektörünü iyi bilen uzmanlar tarafından gerçekleştirilmesi

gerektiğini vurgulamak gerekir. Bu modellerin potansiyel riskleri ve sınırlamaları hakkında yeterli bilgiye sahip olmayan kişiler tarafından kullanılması, yanlış teşhis veya tedavi önerilerine yol açabilir ve hasta sağlığını tehlikeye atabilir. Bu çalışmada, Türkçe sağlık verileri üzerinde dört farklı LLM'in ince ayar performansı incelenmektedir. Bu modeller şunlardır:

- Meta-Llama-3-8B [1]
- SambaLingo-Turkish-Chat [3]
- Trendyol-LLM-7b-chat-v1.8 [4]
- Turkish-Llama-8b-v0.1 [2]

Bu modeller, LLama2, LLama3 ve Mistral tabanlı büyük dil modelleridir. Çalışmada, bu modellerin Türkçe sağlık verileri üzerindeki performansları değerlendirilmiştir. Modellerin eğitimi ve değerlendirilmesi için çalışmada kullanılan "patient-doctor-qa-tr-321179" veri kümesi, doktorsitesi veri kümesinin [7] temizlenmesiyle elde edilen patient-doctor-qa-tr-321179 veri kümesi [8] ve patient-doctor-qa-tr-5695 [9], patient-doctor-qa-tr-95588 [10], patient-doctor-qa-tr-19583 [11] veri kümelerinin birleştirilmesiyle 321.179 soru-cevap çiftinden oluşan nihai veri seti elde edilmiştir. Elde edilen veri kümesi Hugging Face platformuna yüklenmiştir. Çalışmamızda, modellerin performansını ölçmek için ROUGE, BLEU, BERT SCORE, CER, WER ve METEOR gibi çeşitli metrikler kullanılmıştır. Ayrıca, modellerin birbirleriyle karşılaştırılması için Elo puanlaması ve kazanma yüzdesi gibi yöntemler de uygulanmıştır. Son olarak, doktorlar tarafından yapılan uzman değerlendirmeleri ile model cevaplarının gerçeğe ve etiğe uygunluğu incelenmiştir.

1.1 Literatür Özeti

LLM'ler ve sağlık verileriyle uygulamaları, son yıllarda YZ'nin popülerleşmesiyle beraber araştırmacıların ilgisini çeken bir alan olmuştur. Bu bölümde, konuyla alakalı mevcut literatürün kapsamlı bir özeti verilecektir. Literatür taraması, araştırmamızın temelini oluşturan ve bilim dünyasındaki mevcut bilgi dağarcığını anlamamıza vesile olan gayet önemli bir adımdır. Bu süreçte, konuyla ilgili akademik makaleler, konferans bildirileri, kitaplar ve birçok güvenilir kaynak incelenmiştir. Bilhassa hakemli dergiler öncelikli olarak ele alınmış, sonrasında diğer kaynaklardan da istifade edilmiştir. LLM'lerin sağlık alanındaki kullanımıyla alakalı ve Türkçe LLM'lerle ilgili literatür, genel olarak şu ana başlıklar altında incelenebilir:

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

LLM'lerin sağlık verilerinde kullanımı başlığı altında yapılan çalışmalar, bu modellerin tıbbi belgeleri anlamlandırma ve yorumlama hususundaki potansiyelini göstermektedir. Mesela, Peng ve arkadaşları tarafından yapılan çalışmada, LLM'lerin elektronik sağlık kayıtlarından klinik anlam çıkarma başarısı incelenmiştir [12]. Peng ve arkadaşları, BERT ve ELMo gibi önceden eğitilmiş dil modelleriyle, biyomedikal ve klinik veriler üzerinde 5 farklı iş ve 10 farklı veri kümesi içeren gayet kapsamlı bir kıyaslama çalışması yapmışlardır. Sonuçlar, bilhassa PubMed makaleleri ve MIMIC-III klinik notlarıyla önceden eğitilmiş BERT modellerinin, genel olarak en iyi performansı verdiğini gözler önüne sermiştir. Hasta-doktor ilişkisi alanındaysa, Park ve arkadaşları, LLM'lerin hasta sorularına cevap verme ve tıbbi bilgileri daha sarih bir dille açıklama konusundaki başarılarını ölçmüştür [13]. Çalışmada, LLM'lerin anlaması zor tıbbi terimleri basitleştirme ve hastaya özel açıklamalar yapma konusunda potansiyel vadeden sonuçlar alınmıştır. Bu bulgular, LLM'lerin tıbbi metin anlama, tıbbi metinden bilgi çıkarımı ve iletişim görevlerinde kullanılabileceğini gözler önüne sermektedir. Klinik karar destek sistemlerinde, YZ teknolojilerinin potansiyeli yadsınamaz. Jiang ve arkadaşları tarafından yapılan gayet kapsamlı bir araştırmada, YZ ve makine öğrenmesi tekniklerinin sağlık hizmetlerindeki mevcut uygulamaları ve geleceği tartışılmıştır [14]. Çalışma, YZ'nin tanı koyma, tedavi önerme ve prognoz değerlendirme konularındaki potansiyelini gözler önüne sermektedir. Bilhassa görüntüleme, genetik ve elektronik sağlık kayıtları gibi farklı veri türlerinin analiz edilmesinde YZ tekniklerinin gayet başarılı sonuçlar ortaya koyduğu belirtilmiştir. Bununla birlikte, araştırmacılar YZ sistemlerinin potansiyel sınırlamalarına ve insan faktörünün henüz tamamen yok sayılamayacağına da dikkat çekmişlerdir. Sağlık sektöründe YZ uygulamaları, titizlikle ele alınması gereken gayet önemli etik ve yasal sıkıntıları beraberinde getirmektedir. Chikhaoui ve arkadaşlarının araştırmasındaki gibi, temel etik sıkıntılar arasında hasta mahremiyeti ve veri korumasına yönelik potansiyel korkular, algoritmik önyargı ve adalet sıkıntıları,

YZ sistemlerinin şeffaflığı ve açıklanabilirliğiyle ilgili sorular ve insan gözetiminin sürdürülmesi ihtiyacı ve asıl sorumlunun kim olduğu sorunu yer almaktadır [15]. Yasal taraftansa, YZ sistemleri tıbbi karar verme gibi süreçleri yönettiğinde ve bu süreçlerde sıkıntılar meydana geldiğinde sorumluluk ve yükümlülük konularındaki soru işaretleri hala giderilebilmiş değildir. Ayrıca, mevcut sağlık yasaları ve politikalarının YZ teknolojilerini yeterince kapsamadığı yasal boşluklar bulunmaktadır. Bu nedenle, LLM'lerin sağlık hizmetlerinde güvenli ve sürekli bir biçimde kullanılabilmesi için çok disiplinli bir yaklaşım ve gayet kapsamlı düzenlemeler lazımdır. Yazarlar, özellikle sağlık hizmetlerinde YZ için etik kılavuzlar ve uygun yasalar geliştirmenin öneminden bahsetmektedir. Avrupa Komisyonu'nun Güvenilir YZ için Etik Kılavuzları gibi uğraşlarının doğru yönde atılmış adımlar olduğunu söylemekle beraber, sağlık hizmetlerinin özgün bağlamını ele almak amacıyla çok fazla çalışma yapılmasının lazım olduğunu söylemektedirler. Türkçe dil modellerinin geliştirilmesi ve değerlendirilmesi konusunda da önemli çalışmalar yapılmıştır. Uludoğan ve arkadaşları tarafından geliştirilen TURNA modeli [16], Türkçe için hem metin anlama hem de metin oluşturma işlerinde kullanılabilen ilk birleşik dil modelidir. *1.1* milyar parametreye sahip olan bu model, *UL2* çerçevesine dayalı bir kodlayıcı-kod çözücü mimarisi kullanmaktadır ve 43 milyar tokenlık gayet kapsamlı bir Türkçe veri kümesi üzerinde eğitilmiştir. TURNA'nın performansı, üç adet metin oluşturma görevi ve beş adet metin anlama görevi içeren kapsamlı bir değerlendirme süreci ile test edilmiş ve hem metin anlama hem de metin oluşturma görevlerinde mT5 ve mBART gibi çok dilli modelleri geride bıraktığı gösterilmiştir [17, 18]. Doğan ve arkadaşları tarafından yapılan bir çalışmada [19], yedi farklı Türkçe dil modeli seçilerek bu modellerin, bağlamda öğrenme ve soru cevaplama performansları değerlendirilmiştir. Çalışmada, standart veri kümeleri Türkçeye çevrilerek kullanılmış ve modeller hem otomatik metrikler hem de insan değerlendirmesi ile karşılaştırılmıştır. Sonuçlar, Trendyol-chat modelinin genel olarak en iyi performansı gösterdiğini ortaya koymuştur. Kesgin ve arkadaşları tarafından geliştirilen cosmosGPT modelleri [20] ise tamamen Türkçe veriler kullanılarak eğitilen iki farklı boyutta model sunmaktadır. Çalışmanın sonuçları, cosmosGPT modellerinin, kendilerinden çok daha büyük çok dilli modellerle aşık atabilecek performans gösterdiğini ortaya koymuştur. Bu bulgular, dil modellerinde parametre sayısındaki büyüklüğün tek başına belirleyici olmadığını ve dile özel eğitimin önemli olduğunu göstermektedir. Türkçe sağlık verileri üzerinde LLM'lerin performansını inceleyen çalışmalar ise nispeten sınırlıdır. Bu alandaki boşluk, bizim araştırmamızın önemini ve katkı potansiyelini göstermektedir. Sonuç olarak, mevcut literatür LLM'lerin Türkçe sağlık alanında önemli bir başarı potansiyeline haiz olduğunu ortaya koymaktadır. Lakin, bu modellerin güvenilir ve etik bir

şekilde kullanılabilmesi için daha fazla araştırma ve geliştirme faaliyetine ihtiyaç duyulmaktadır. Bizim çalışmamız, özellikle Türkçe sağlık verileri üzerinde LLM'lerin performansını ve potansiyelini inceleyerek bu alana katkı sağlamayı amaçlamaktadır.

1.2 Tezin Amacı

Bu tezin temel amacı, Türkçe sağlık verileri üzerinde LLM'lerin ince ayar performansını incelemek ve değerlendirmektir. Çalışmamızın spesifik hedefleri şunlardır:

1. LLM'lerin hasta-doktor ilişkisindeki potansiyelini ölçmek. Bu modeller, tıbbi bilgileri hastaların anlayacağı bir dille açıklama, hasta sorularını cevaplama ve sağlık çalışanlarına yardımcı olma konularında kullanılabilir.
2. Genel amaçlı LLM'lerin kapladıkları devasa alanlar ve devasa enerji ve işlem gücü tüketimi gibi dezavantajlarının en alt seviyeye indirildiği duruma özel LLM üretmek.
3. Türkçeye özel ve hasta-doktor ilişkisine odaklanmış daha küçük ve verimli LLM'lerin elzem olduğunu göstermek. Bu tür özelleştirilmiş modeller, hem doğruluk, hem de boyut ve enerji tüketimi açısından daha avantajlıdır, böylece sağlık hizmetlerinde daha yaygın ve etkin kullanım sağlanabilir.
4. Yapılan ince ayar eğitimi sonucunda, LLM'lerin Türkçe sağlık verileri üzerindeki potansiyelini göstermek. Bu çalışma, özelleştirilmiş modellerin performansını ve pratik uygulanabilirliğini değerlendirmeyi hedeflemektedir.
5. LLM'lerin performansını üç farklı yöntemle değerlendirmek:
    - Büyük yapay zeka modellerinin hakemliğinde değerlendirme
    - Sentetik testlerle değerlendirme (BLEU, ROUGE, BERT SCORE, CER, WER, METEOR gibi metrikler kullanarak)
    - Gerçek doktor uzmanların değerlendirmesi

Bu amaçlar doğrultusunda, tezimiz Türkçe sağlık verileriyle eğitilen LLM'lerin performansını kapsamlı bir şekilde değerlendirerek, bu modellerin sağlık hizmetlerinde güvenilir, etik ve verimli bir şekilde kullanılabilmesi için gerekli temel bilgiyi sağlamayı amaçlamaktadır. Ayrıca, özelleştirilmiş ve daha küçük modellerin yapabileceklerini göstererek, ilerideki araştırma ve uygulamalar için ufuk açmayı hedeflemektedir.

1.3 Orijinal Katkı

Bu tez çalışması, Türkçe sağlık verileri üzerinde LLM'lerin performansını inceleyerek literatüre özgün katkılar sağlamayı hedeflemektedir. Çalışmamızın başlıca orijinal katkıları şunlardır:

1. Hasta-doktor iletişimi alanına özel LLM'ler oluşturmak: Bu çalışma, Türkçe sağlık verileri kullanılarak hasta-doktor iletişimine özel olarak eğitilmiş LLM'ler oluşturmayı hedeflemektedir. Bu özelleştirilmiş modeller, genel amaçlı LLM'lere kıyasla daha doğru ve bağlama uygun yanıtlar üretmektedir.

2. Geliştirilen LLM'lerin kapsamlı karşılaştırması: Çalışmamız, dört farklı LLM'in performansını çeşitli metrikler ve değerlendirme yöntemleri kullanarak karşılaştırmaktadır. Bu karşılaştırma, Türkçe sağlık verilerinde hangi modelin daha kullanıma uygun olduğunu belirlemeye yardımcı olacaktır.

3. Özgün hasta-doktor iletişimi veri kümesi oluşturmak: Tez kapsamında, Türkçe hasta-doktor iletişimine özel bir veri kümesi oluşturulmuştur. Bu veri kümesi, gerçek hasta-doktor verilerinden oluşan 321.179 soru-cevap çiftini içermektedir ve LLM'lerin eğitimi ve değerlendirilmesi için kullanılmıştır.

4. Çok yönlü değerlendirme yaklaşımı: Geliştirilen LLM'lerin performansı, üç farklı yöntemle değerlendirilmiştir:

   - Büyük yapay zeka modellerinin karşılaştırmalı değerlendirmesi
   - Sentetik testler (BLEU, ROUGE, BERT SCORE gibi metrikler kullanarak)
   - Gerçek doktor uzmanların değerlendirmesi

   Bu çok yönlü yaklaşım, modellerin performansının daha ayrıntılı ve gerçeğe yakın bir şekilde değerlendirilmesini sağlamaktadır. Ayrıca, çalışmamızda sentetik testlerin gerçek doktor değerlendirmelerinin yerine kullanılıp kullanılamayacağını da inceleyeceğiz.

5. Enerji ve donanım verimliliği: Çalışmamız, özelleştirilmiş ve daha küçük LLM'ler geliştirerek, sağlık alanında kullanılacak modellerin enerji ve donanım gereksinimlerini azaltmayı ve hastalar için modelleri daha ulaşılabilir hale getirmeyi hedeflemektedir.

2
LLM'ler ve Transformer Mimarisi

Son zamanlarda yapay zeka ve DDİ alanında yaşanan gelişmeler, LLM gibi teknolojilerin ortaya çıkışını hızlandırmıştır. Bu modeller, insan dilini anlama ve üretme yetenekleriyle dikkat çekerken, GPT-4o, Claude 3.5 Sonnet, Gemini Copilot ve gibi modellerle bu alanda yeni bir çığır açılmıştır [21–24]. Bu bölümde, LLM'lerin ne olduğu, gelişimi ve sağlık alanında fine-tune ile kullanılma potansiyelini ele alacağız.

2.1 LLM Nedir?

LLM, doğal dil işleme alanında kullanılan gelişmiş yapay zeka modelleridir. Bu modeller, özellikle yüksek düzey DDİ görevleri için tasarlanmış ve devasa donanım kaynağı kullanan yapay zeka sistemleridir.

LLM'lerin temel yapısını, transformer modeli mimarisi oluşturur. Bu mimari, paralel işlemedeki hesaplama kazancı vesilesiyle en büyük ve en güçlü LLM'lerin oluşturulmasına olanak sağlar. Transformer mimarisine sahip modeller, kendi kendine dikkat mekanizması (self attention) kullanarak, giriş kelimelerini birbiriyle karşılaştırır ve her kelime için bir önem puanı hesaplar [25].

LLM'ler, aşağıdaki özelliklere ve yeteneklere sahiptir:

- **Kapsamlı Veri Seti Üzerinde Eğitim**: LLM'ler, devasa miktarda veri kullanılarak eğitilir ve bu vesileyle kapsamlı bir bilgi dağarcığına sahip olurlar.
- **Esnek Kullanım Yelpazesi**: Bu modeller, metin özetleme, metin artırma, metin çevirme, metinleri farklı üsluplarla dönüştürme ve hatta görüntü, pdf gibi farklı veri tiplerini işleyebilme gibi çeşitli görevleri gerçekleştirebilir.
- **Bağlam Anlama**: Transformerler vesilesiyle, LLM'ler kelimelerin ve cümlelerin bağlamını, nüanslarını ve aralarındaki ilişkileri anlayabilir.

- **Sürekli Gelişim**: LLM'ler, eğitildikçe ve LLM'lerin kullanım sıklığı arttıkça kullanıcıların geri bildirimleriyle daha doğru ve güvenilir hale gelir.

LLM'lerin, çok çeşitli sektör ve alanlarda kullanım potansiyeli vardır. Bunlardan bazıları:

- Arama motorları
- Doğal dil işleme
- Sağlık hizmetleri
- Robotik
- Kodlama
- Reklam ve pazarlama
- E-ticaret
- Eğitim
- Finans
- İnsan kaynakları
- Hukuk

LLM'lerin kullanımı bazı etik ve güvenlik kaygılarını ve sıkıntılarını da beraberinde getirmektedir:

- **Veri Gizliliği**: Kurumlar, ücretsiz LLM'lere özel veya gizli verilerini ve bilgilerini vermemelidir.
- **Çevresel Etki**: LLM'leri eğitmek ve LLM'lerin bakımını yapmak için çok yüksek oranda enerji tüketilir, bu da uzun vadede sürdürülebilirlik açısından engel oluşturma potansiyeline sahiptir.
- **Taraflılık ve Etik Sorunlar**: LLM'lerin oluşturduğu verilerde taraflılık ve etik sorunlar oluşabilir. Bundan dolayı LLM teknolojisini kullanırken, bu teknolojinin nasıl sorunlara yol açabileceğini tam olarak anlamak gerekir.

2.2 Transformers Mimarisi Nedir?

Transformer mimarisi, DDİ ve yapay zeka alanında çığır açan bir yapay sinir ağı mimarisidir. İlk olarak *Vaswani*'nin 2017 yılındaki çalışmasında tanıtılmıştır [25]. GPT, LLama serisi gibi birçok başarılı dil modeli bu mimariyi kullanır [26–29]. Transformer mimarisi, dil verilerini işlemek için geleneksel yinelemeli veya katmanlı sinir ağları yerine, yoğun bir şekilde **dikkat** mekanizmalarını kullanarak metinleri analiz eder. Bu mimari, birbirlerine uzak tokenların bağımlılıkları daha iyi yakalar ve paralel hesaplamalarla daha hızlı çalışır. Transformer mimarisi, bilhassa büyük veri kümeleri ve GPU/TPU tabanlı paralel hesaplama gücü ile daha hızlı ve başarılı sonuçlar verir. Doğal dil çevirisi, metin/belge özetleme, metin/belge sınıflandırma ve konuşma tanıma gibi çeşitli DDİ işlerinde kullanılabilir. Ayrıca, esnek yapısı vesilesiyle DDİ dışında görüntü işleme ve ses işleme gibi diğer alanlarda da hatrı sayılır derecede başarılı çalışabilmektedir.

2.2.1 Transformer Mimarisi ve Bileşenleri

Transformerlar, girdi olarak verilen diziyi alıp bir çıktı dizisi oluşturan veya girdi dizisini değiştiren bir tür sinir ağı mimarisidir. Bu modeller, bağlamı iyi derecede bilerek ve dizi bileşenleri arasındaki alakayı anlayarak çalışır. Transformerlar, konuşma tanımadan tutun makine çevirisine ve protein dizisi analizine kadar her tür dizi dönüşümü için kullanılmaya oldukça elverişlidir. Transformer modelleri, DDİ işlerini iyi derecede yapabilen derin öğrenme modelleridir ve yüzlerce kelimeden oluşan metinlerde bile, kelimeler arasındaki bağımlılıkları yakalayarak, işleyip analiz edebilirler. Transformer modelleri, sinir ağlarının öz dikkat mekanizmasını kullanarak çok uzun dizileri, RNN'lerden ve CNN'lerden farklı olarak, bir bütün halinde işleyebilir.

Transformerların bazı bileşenleri şunlardır:

- **Konumsal Kodlama (Positional Encoding)**: Transformer modelleri, bir sonraki verinin pozisyonunu belirtmek için konumsal kodlamayı kullanır. Bu, modelin giriş verisindeki kelimelerin veya token'ların, giriş verisindeki sırasını anlamasını sağlar.
- **Dönüştürücü Bloku (Transformer Block)**: Transformer modelinin temel yapı taşıdır. Bu blok, aşağıdaki parçaları içerir:
  - Çoklu başlıklı dikkat mekanizması (multi-head attention)
  - Doğrusal katmanlar (feed-forward neural networks)
  - Katman normalizasyonu (layer normalization)

  - Artık bağlantılar (residual connections)

- **Çıktı Katmanı**: Model çıktılarını hesaplanmada kullanılır ve aşağıdaki parçaları içerir:
  - **Doğrusal Katman**: Girdileri ağırlıklarla çarpar.
  - **Softmax Fonksiyonu**: Modelin çıktısını olasılık dağılımına çevirir.

2.2.2 Transformer Mimarisi Uygulamaları

GPT, önceden eğitilmiş Transformer modeli olarak 2018 yılında Dünya'ya tanıtılmıştır ve çeşitli DDİ işlerinde kullanılmıştır [27]. BERT, 2018 yılında tanıtılan diğer büyük önceden eğitilmiş modeldir ve doğal dili daha iyi anlaması için tasarlanmıştır [30]. GPT-2, GPT'nin geliştirilmiş ve daha büyük bir versiyonudur aynı zamanda büyük versiyonu *1024* bağlam uzunluğuna ve *1.5 Milyar* parametreye sahiptir [28]. DistilBERT, BERT'in %40 daha küçük ve hızlı bir versiyonudur ve buna rağmen BERT'in %97'si kadar başarılıdır [31]. BART ve T5 ise orijinal Transformer mimarisi kullanan büyük ve önceden eğitilmiş modellerdir ve yine DDİ alanında kullanılırlar [32, 33]. GPT-3, fine-tuning gerektirmeyen çeşitli görevlerde iyi performans gösterebilen *2048* bağlam uzunluğuna ve *175 Milyar* parametreye sahip olan daha büyük bir GPT-2 versiyonudur [26]. Transformer modellerini GPT benzeri, BERT benzeri ve BART/T5 benzeri modeller olmak üzere üç kategoriye ayırabiliriz. Tüm Transformer modelleri (GPT, BERT, BART, T5 vb.) devasa boyuttaki metin verisiyle öz denetimli bir şekilde eğitilmiştir. Bu tür modeller, genellikle transfer öğrenme adındaki bir yöntemle özel bir görev için ince ayar (fine-tuning) yapılır [34]. İnce ayar işlemi, en baştan eğitme işine göre daha az veri, zaman ve kaynak gerektirir. Aynı zamanda önceden eğitilmiş olan modelin eski anlamlı kelime ağırlıklarından da faydalanılmış olur.

3
YÖNTEM

Bu bölümde, çalışmamızda kullanılan yöntemler detaylı bir şekilde açıklanacaktır. Araştırmamızın temelini oluşturan veri kümesi oluşturma süreci, araştırmada kullanılan LLM'ler ve ince ayar işlemleri hakkında bilgi verilecektir.

3.1 Veri Kümesi Oluşturma

Çalışmamızda kullanılan veri kümesi, Hugging Face platformunda bulunan **Patient Doctor Q&A TR 321179** adlı özel bir Türkçe hasta-doktor soru-cevap veri kümesidir [5]. 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.

3.1.1 Veri Kümesinin Oluşturulması ve Birleştirilmesi

Patient Doctor Q&A TR 321179 veri kümesi, dört farklı veri kümesinin birleştirilip sonrasında karıştırılması sonucunda oluşturulmuştur [5]. Bu veri kümeleri şunlardır:

- Patient Doctor Q&A TR 19583 [11]
- Patient Doctor Q&A TR 167732 [35]
- Patient Doctor Q&A TR 5695 [9]
- Patient Doctor Q&A TR 95588 [10]

3.1.1.1 Patient Doctor Q&A TR 167732

Patient Doctor Q&A TR 167732 veri seti, **doktorsitesi** veri setinin [7] temizlenmiş halidir. Temizleme işlemi kapsamında telefon numaraları, mail adresleri, açık adresler, bağlam bağımlı cümleler kaldırılmış, noktalama işaretleri düzeltilmiş

ve büyük/küçük harf kullanımı standardize edilmiştir. Bu işlem OpenAI'nın geliştirdiği *gpt-3.5-turbo-0125* ile yapılmıştır [36]. Bu veri seti 4 sütun içermektedir: Ünvan, Alan, Soru ve Cevap. İçerik Türkçe dilindedir ve çeşitli tıbbi konuları kapsamaktadır [35].

3.1.1.2 Patient Doctor Q&A TR 5695

Patient Doctor Q&A TR 5695 veri seti, **doctor-id-qa** veri setinin Türkçeye çevrilmiş halidir [37]. Çeviri işlemi yine *gpt-3.5-turbo-0125* ile yapılmıştır [36]. Bu veri seti 2 sütundan oluşmaktadır: Soru ve Cevap. Tüm içerik Türkçe dilindedir ve çeşitli tıbbi konuları kapsamaktadır [9].

3.1.1.3 Patient Doctor Q&A TR 19583

Patient Doctor Q&A TR 19583 veri seti, iCliniq platformundaki gerçek hasta soruları ve doktor yanıtlarının Türkçeye çevrilmiş halini içeren bir koleksiyondur. Bu veri seti, aslında **iCliniq Medical QA** veri setinin [38] Türkçeye çevrilmiş halidir. Çeviri işlemi yine *gpt-3.5-turbo-0125* ile yapılmıştır [36]. 3 sütun içermektedir: Başlık, Soru ve Cevap [11].

3.1.1.4 Patient Doctor Q&A TR 95588

Patient Doctor Q&A TR 95588 veri seti, **chat_doctor** veri setinin [**avaliev2024chat_doctor**] Türkçeye çevrilmiş halidir. Çeviri işlemi yine *gpt-3.5-turbo-0125* ile yapılmıştır [36]. Bu veri seti 3 sütun içermektedir: Talimat, Soru ve Cevap. Tüm içerik Türkçe dilindedir ve çeşitli tıbbi konuları kapsamaktadır [10].

3.1.1.5 Birleştirilmiş Veri Setinin Özellikleri

Birleştirilen veri seti, tıbbi araştırmalar, DDİ ve tıbbi eğitim gibi alanlarda kullanılabilecek zengin bir kaynak niteliğindedir. Ama, soru ve yanıt kalitesindeki farklar ve potansiyel önyargılar gibi sıkıntılar göz ardı edilmemelidir. Bu birleştirilmiş veri seti, araştırmacılara ve eğitimcilere, Türkçe tıbbi iletişim verilerini kullanarak daha farklı alan ve modellerde test etme imkanı sunmaktadır.

3.1.2 Veri Ön İşleme

Veri kümesinin kalitesini artırmak ve veriyi modellerin eğitimine hazır hale getirmek için çeşitli ön işleme adımları uygulanmıştır. Bu adımlar, veri madenciliği

ve makine öğrenmesinin temel ve önemli bir parçasıdır. Uygulanan ön işleme adımları şunlardan oluşmaktadır:

- **Metin temizleme**: Gereksiz boşlukların kaldırılması.
- **Tekrar etme kontrolü**: Tekrar eden soru-cevapların tespit edilmesi ve kaldırılması. Bu, veri kümesindeki faydasız tekrarları n önüne geçer ve modellerin eğitim performansını iyileştirir.
- **Eksik veri kontrolü**: Boş veya anlamsız içeriklerin filtrelenmesi. Bu ön işleme adımı, veri kümesinin bütünlüğünü korur ve modelin eğitim performansını yükseltir.
- **Veriyi tek bir metin haline getirme**: Soru-cevap çiftleri modeli eğitmek üzere aşağıdaki formatlarda yapılandırılmıştır: SambaLingo-Turkish-Chat ve Trendyol-LLM-7b-chat-v1.8 modellerinde özel tokenlar bulunmadığı için

  ```
  Sen bir doktorsun. Soruları buna göre cevapla.
  ### Soru:
  {}

  ### Cevap:
  {}
  ```

  Meta-Llama-3-8B ve Turkish-Llama-8b-v0.1 modellerinde özel tokenlar bulunduğu için

  ```
  Sen bir doktorsun. Soruları buna göre cevapla.
  ### <|reserved_special_token_0|>:
  {}

  ### <|reserved_special_token_1|>:
  {}
  ```

Bu format, modelin doktor rolünü benimsemesini ve soruları bu bağlamda cevaplamasını sağlamak için özel olarak tasarlanmıştır. Boş küme parantezleri (), veri setindeki gerçek soru ve cevapların yerleştirileceği alanları temsil eder. Veri ön işleme aşaması, modelin eğitim sürecinde karşılaşabileceği potansiyel sorunları minimize etmeyi ve veri setinin genel kalitesini artırmayı amaçlar. Bu adımlar, modelin daha iyi genelleme yapmasına ve daha doğru tahminlerde bulunmasına yardımcı olur.

3.2 Kullanılan LLM'ler

Bu bölümde, çalışmamızda kullandığımız dört farklı LLM ayrıntılı bir şekilde açıklanacaktır.

3.2.1 Meta-Llama-3-8B

Meta-Llama-3-8B, Meta AI tarafından geliştirilen Llama ailesinin üçüncü neslidir. Bu model, 8 milyar parametre içermektedir ve haleflerine göre daha gelişmiş doğal dili anlama ve metin oluşturma özelliklerine sahiptir [1]. Modelin 70 milyar parametreye sahip bir varyasyonu da bulunmaktadır ama o modeli çalıştıracak donanıma ulaşmak çok zor ve maliyetli olduğundan 8 milyar parametreli versiyonunu kullandık. Meta-Llama-3-8B, kampsamlı bir yelpazedeki işlerde kullanılabilecek bir modeldir ve bilhassa Türkçe gibi düşük kaynaklı dillerde de iyi performans göstermektedir. Önemli özellikleri:

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

3.2.2 SambaLingo-Turkish-Chat

SambaLingo-Turkish-Chat , SambaNova Systems tarafından geliştirilen, Türkçe dili için ince ayar yapılarak eğitilmiş bir sohbet modelidir [3]. Bu model, SambaLingo-Turkish-Base üzerine mesajlaşmaya uyumlu hale getirilmiş ve *DPO* yöntemiyle eğitilmiştir [39]. SambaLingo-Turkish-Base modeli Llama-2-7b tabanlı bir modeldir [40]. Önemli özellikleri:

- 7 milyar parametre
- Llama-2-7b modelinin Türkçe'ye adaptasyonu [40]
- Türkçe dili için ince ayar
- Cultura-X veri kümesinin Türkçe bölümünden 42 milyar tokenla eğitimi
- İki aşamalı ince ayar süreci: SFT ve DPO

3.2.3 Trendyol-LLM-7b-chat-v1.8

Trendyol-LLM-7b-chat-v1.8 **(Model 3)**, Trendyol tarafından geliştirilen ve özellikle Türkçe dili üzerine ince ayar yapılmış bir LLM'dir [4]. Önemli özellikleri:

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

3.2.4 Turkish-Llama-8b-v0.1

Turkish-Llama-8b-v0.1 **(Model 4)**, YTÜ Bilgisayar Mühendisliği COSMOS Araştırma Grubu tarafından geliştirilen, Türkçe dilinde *30GB* veriyle ince ayar yapılmış bir LLM'dir [2]. Bu modelin ince ayarı, Meta-Llama-3-8B modeli üzerinde yapılmıştır [1]. Önemli özellikleri:

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

Bu dört model, çalışmamızda Türkçe hasta-doktor soru-cevap için performanslarını karşılaştırmak ve değerlendirmek üzere seçilmiştir. Bu modelleri seçerken *Open-LLM Turkish Leaderboard*'daki sıralamalara bakılarak farklı mimariye sahip modeller seçilmiştir [41].

3.3 İnce Ayar (Fine-tuning) Süreci

Bu bölümde, kullanılan modellerin ince ayar süreci ve bu süreçte kullanılan hiperparametreler gayet ayrıntılı bir şekilde açıklanacaktır.

3.3.1 Hiperparametreler

İnce ayar sürecinde kullanılan hiperparametreler, modelin performansını optimize etmek için oldukça önemlidir. Çalışmamızda kullanılan hiperparametrelerse şu şekildedir:

3.3.1.1 Öğrenme oranı (Learning rate)

Öğrenme oranı, derin öğrenme modellerinin eğitimi sırasında kullanılan en önemli hiperparametrelerden biridir. Bu parametre, modelin ağırlıklarının her turda hangi oranda güncelleneceğini belirler. Bu hiperparametrenin temel etkileri şu şekildedir:

- **Güncelleme hızı**: Büyük öğrenme oranı, modelin ağırlıklarını her adımda daha büyük boyutta güncellemesine sebebiyet verir. Ama bu durum aşırı uyum (overfitting) riskini beraberinde getirir.
- **Hassasiyet**: Az öğrenme oranı, modelin daha hassas bir eğitim sürecinden geçmesini sağlar, ancak eğitim süresini uzatabilir.
- **Yerel minimumlara takılma riski**: Olması gerekenden daha düşük bir öğrenme oranı, modelin yerel minimumlara takılma olasılığını artırabilir.
- **Yakınsama**: Doğru belirlenmiş bir öğrenme oranı, modelin global minimuma daha hızlı ve efektif bir şekilde yakınsamasını sağlar.
- **Genelleme yeteneği**: Doğru belirlenmiş bir öğrenme oranı, modelin eğitildiği veriye aşırı uyum göstermesini (overfitting) önleyerek, başka veriler için daha düzgün çalışabilme olasılığını artırır.

Öğrenme oranının seçimi, veri kümesinin büyüklüğü, model mimarisi ve eğitim süresi gibi değişkenlere bağlı olarak değişebilir. Bundan dolayı, en iyi öğrenme oranını bulmak için genellikle referans çalışmalar ve hiperparametre optimizasyonu teknikleri kullanılır.

1×10^{-4} olarak belirlenmiştir. Bu değer, literatürdeki birçok çalışma ayrıntılı bir biçimde taranarak seçilmiştir. Chen ve arkadaşları, derin öğrenme modellerinin hiperparametre optimizasyonu üzerine yaptıkları çalışmada, öğrenme oranının model performansı üzerinde gayet etkili olduğunu açık bir şekilde göstermiştir [42]. Hoffmann ve arkadaşları, LLM'lerin eğitiminde optimal hesaplama yöntemleri üzerine yaptıkları araştırmada, büyük modeller için öğrenme oranının en fazla 1.25×10^{-4} olması gerektiğini söylemiştir [43]. Ayrıca, Akyön ve arkadaşları tarafından Türkçe metinler üzerinde yapılan otomatik soru üretme ve cevaplama araştırmasında da 1×10^{-4} ve 1×10^{-3} değerlerini kullanmıştır, bu değerler de seçimimizde etkili olmuştur [44]. Son olarak, Kesgin ve arkadaşları tarafından yapılan çalışma da, Türkçe dil modelleri için 1×10^{-4} öğrenme oranı kullanılmıştır [20].

Çalışmamızda kullanılan veri kümesi, toplam 321.179 soru-cevap çiftinden oluşmaktadır. Bu boyuttaki bir veri kümesi için 1×10^{-4} öğrenme oranı, modelin

eğitim sürecinde aşırı öğrenmeyi (overfitting) önlemek ve veri kümesinde olmayan veriler için de çalışabilirlik oranını artırmak açısından iyi bir seçimdir. Büyük veri kümeleri ile eğitim yaparken, daha küçük öğrenme oranları kullanmak, modelin daha stabil bir şekilde öğrenmesini ve lokal minimumlara takılma riskinin azaltılmasına yardımcı olur. Bu sebeple, veri kümemizin büyüklüğünü göz önüne aldığımızda, seçilen öğrenme oranı, modelin düzgün ve etkili bir şekilde öğrenmesini ve eğitim performansını optimize etmesini sağlayacak şekilde belirlenmiştir.

3.3.1.2 Isınma adımları (Warmup steps)

Isınma adımları *2000* olarak belirlenmiştir. Bu değer, veri kümemizin boyutu (321.179 soru-cevap çifti) dikkate alınarak deneysel olarak seçilmiştir. Isınma adımları, eğitimin başlangıç aşamasında öğrenme oranını peyderpey olarak artıran bir tekniktir. Bu hiperparametrenin temel etkileri şu şekildedir:

1. **Eğitim stabilitesi**: Eğitimin başlangıcında daha düşük öğrenme oranı kullanılarak, modelin ağırlıklarının gereğinden fazla değişime uğramasını engeller.
2. **Yerel minimumlara takılma riskini azaltır**: Öğrenme oranını yavaşça artırarak, modelin henüz daha geniş bir parametre uzayını keşfetmeden yüksek değişimlere uğramasını önler.
3. **Gradyan patlaması problemini önler**: Bilhassa derin ağlarda, eğitimin başlangıcında öğrenme oranı ne kadar düşük olsa da gradyan patlamasına sebep olabilir. Isınma adımları bu riski minimize eder.
4. **Adaptasyon süreci**: Model, veri kümesine ve öğrenme işine tedrici olarak uyum sağlar, bu da görece daha iyi bir başlangıç noktası sağlar.
5. **Performans iyileştirmesi**: Doğru ayarlanmış ısınma adımları, modelin nihai performansını artırır ve daha doğru yakınsamasını sağlar.

Veri kümemizin büyüklüğü göz önüne alındığında, *2000* ısınma adımı, modelin eğitim sürecinin başlangıcında dengeli bir biçimde ilerlemesini ve veri kümesine uyum sağlamasını sağlayacak bir biçimde seçilmiştir. Bu, özellikle karmaşık dil modelleri için zaruridir diyebiliriz, zira modelin başlangıçta ağırlıklarını aşırı değiştirmesini önleyerek, sonraki eğitim adımlarında daha etkin öğrenme yapmasına vesile olur.

3.3.1.3 Eğitim döngüsü sayısı (Number of training epochs)

Eğitim döngüsü sayısı, modellere göre farklılık göstermektedir:

- **Meta-Llama-3-8B için**: 1 döngü
- **Diğer modeller için**: 2 döngü

Bu değerler, veri kümemizin boyutu (321.179 soru-cevap çifti) önemsenerek deneysel olarak belirlenmiştir. Eğitim döngüsü sayısı, modelin eğitim veri kümesiyle kaç defa eğitileceğini belirler. Bu hiperparametrenin temel etkileri şu şekildedir:

1. **Öğrenme süreci**: Her döngü, modelin veri kümesindeki tüm verileri bir kez işlemesine vesile olur, bununla birlikte model veri kümesindeki tüm verileri öğrenmiş olur.
2. **Performans optimizasyonu**: Epoch sayısının düzgün belirlenmesi, modelin grektiği kadar öğrenmesine hizmet ederken aşırı öğrenmeyi (overfitting) önlemeye de yardımcı olur.
3. **Hesaplama verimliliği**: Büyük veri kümeleri için az sayıda döngü kullanmak, eğitim süresini iyileştirir ve hesaplama kaynaklarını gereğinden fazla meşgul etmez.
4. **Model kararlılığı**: Farklı modeller için farklı döngü sayıları seçmek, her modelin kendi iç yapısına ve öğrenme hızına göre optimize edilmesini sağlar.

Veri kümemizin boyutu göz önüne alındığında, Meta-Llama-3-8B için *1 döngü* ve diğer modeller için *2 döngü* kullanılması, modellerin veri kümesini gerektiği kadar öğrenmesine yardımcı olurken, aşırı öğrenme riskini minimize etmeye ve eğitim süresini optimize etmeye vesile olur. Bu yaklaşım, bilhassa LLM'ler için lazımdır, çünkü büyük veri kümeleriyle çalışırken her döngü önemli miktarda veri içerir ve daha fazla döngü her zaman daha iyi performans anlamına gelmez ve aynı zamanda her döngü oldukça maliyetlidir.

Kesgin ve arkadaşları tarafından yapılan çalışmada [20], küçük modeller için *3 döngü*, büyük modeller için *2 döngü* kullanılmıştır. Bizim çalışmamızda kullanılan modeller, Kesgin ve arkadaşlarının büyük modellerinin yaklaşık 5 katı büyüklüğündedir. Bundan dolayı, daha büyük modeller için daha az döngü kullanmak, hem hesaplama verimliliği açısından hem de aşırı öğrenme riskini

azaltmak açısından mantıklı ve gerekli bir seçimdir. Hem de büyük modellerin daha hızlı öğrenme kapasitesine sahip olduğunu göz önüne alırsak, daha az döngü ile de gayet etkili sonuçlar elde edilebilmektedir.

3.3.1.4 Maksimum sekans uzunluğu (Max sequence length)

Maksimum sekans uzunluğu, kullanılan modellerin mimarilerinin maksimum sekans uzunluğuna göre ayarlanmıştır. Bu parametre, modelin aynı anda işleyebileceği en yüksek token (kelime veya alt kelime birimi) sayısını belirler. Aslında bu hiperparametre ne kadar büyük olursa model o kadar hafızaya sahiptir diyebiliriz. Çalışmamızda kullanılan modeller için maksimum sekans uzunlukları şu şekilde belirlenmiştir:

- **SambaLingo-Turkish-Chat için**: 4096 token
- **Diğer modeller için**: 8192 token

Bu değerler, her modelin izin verdiği maksimum uzunluk kullanılarak belirlenmiştir. Bu hiperparametrenin temel etkileri şu şekildedir:

1. **Bağlam anlama**: Daha uzun sekans uzunluğu, modelin daha geniş bir bağlamı anlayabilmesini ve işlemesini sağlar, bu da bilhassa uzun metinlerde daha iyi performansa vesile olur.
2. **Bellek kullanımı**: Sekans uzunluğu, modelin bellek kullanımını doğrudan artırır. Daha uzun sekanslar daha yüksek bellek ister, bundan dolayı da hiperparametrenin seçiminde donanım kısıtlamaları da önemsenmelidir.
3. **İşlem süresi**: Daha uzun sekanslar, modelin sonuç üretme zamanını artırır. Bundan dolayı, model performansı ve işlem süresi arasında bir al-ver dengesi sağlanmalıdır.
4. **Model kapasitesi**: Birbirinden farklı modeller için farklı maksimum sekans uzunlukları kullanmak, modellerin kendi mimarilerine göre iyileştirilmesine vesile olur.

Veri kümemizin (321.179 soru-cevap çifti) içeriği ve kullanılan modellerin kapasiteleri göz önünde bulunduruduğunda, belirlenen uzunluklar, modellerin veri kümesindeki uzun metinleri efektif bir surette işlemesine vesile olurken, bununla beraber eğitim sürecinin verimli bir şekilde ilerlemesini sağlar. Bu yaklaşım,

bilhassa Türkçe gibi morfolojik olarak zengin diller için mühimdir, zira daha uzun sekans uzunlukları, dilin karmaşık yapısını ve uzun bağlamları daha iyi işlemeye yardımcı olur.

3.3.1.5 Öğrenme oranı planlayıcı tipi (Learning rate scheduler type)

Öğrenme oranı planlayıcı tipi, çalışmamızda doğrusal (linear) olarak seçilmiştir. Bu hiperparametre, eğitim süresi boyunca öğrenme oranının değişim sürecini kontrol eder. Öğrenme oranı sabit değildir. Bu hiperparametrenin temel etkileri şu şekildedir:

1. **Adaptif öğrenme**: Eğitimin başında daha büyük öğrenme oranıyla başlayıp, süreç içerisinde bu oranı tedrici olarak azaltır. Bu, modelin ilk başlarda daha hızlı öğrenmesini, sonralardaysa daha hassas öğrenmesine vesile olur.
2. **Yerel minimumlara takılmayı önleme**: Eğitim başındaki yüksek öğrenme oranı, modelin yerel minimumlara takılma olasılığını azaltır ve küresel minimuma ulaşma şansını yükseltir.
3. **Eğitim stabilitesi**: Öğrenme oranının tedrici olarak azaltılması, eğitimin sonuna yaklaştıkça modelin daha kararlı hale gelmesini sağlar.
4. **Hesaplama verimliliği**: Doğrusal azalma, hesaplama açısından ucuz, uygulaması ve anlaşılması daha kolay bir yöntemdir.

Doğrusal öğrenme oranı planlayıcısı, büyük dil modelleri eğitiminde gayet yaygın olarak kullanılan bir yöntemdir. Bu yöntem, birçok çalışmada uygulanmış ve başarılı sonuçlar alınmıştır. Mesela, Raffel ve arkadaşları T5 modelini geliştirirken doğrusal öğrenme oranı planlayıcısını kullanmışlardır [45]. Benzer şekilde, Brown ve arkadaşları GPT-3 modelinin eğitiminde doğrusal öğrenme oranı planlayıcısını seçmişlerdir [26]. Ayrıca, Devlin ve arkadaşları da BERT modelinin fine-tuning aşamasında doğrusal öğrenme oranı planlayıcısını tercih etmişlerdir [30]. Bu çalışmalar, doğrusal öğrenme oranı planlayıcısının LLM'lerin eğitiminde etkin ve yaygın bir tercih olduğunu ortaya koymaktadır.

Veri kümemizin büyüklüğü (321.179 soru-cevap çifti) ve kullanılan modellerin parametre sayısı göz önüne alındığında, doğrusal öğrenme oranı planlayıcısı, eğitim sürecinin düzgün şekilde ilerlemesine vesile olur. Bu yaklaşım, bilhassa büyük dil modelleri için uygundur, zira modelin eğitim sürecinin farklı adımlarında farklı öğrenme hızlarının uygulanmasına olanak tanır.

3.3.1.6 Ağırlık azaltma (Weight decay)

Ağırlık azaltma değeri *0,01* olarak belirlenmiştir. Bu değer, Wu ve arkadaşlarının çalışmasında da kullanılmıştır [46]. Ağırlık azaltma, aşırı öğrenmeyi (overfitting) engellemek ve modelin gerçek dünya verilerini işleyebilme yeteneğini artırmak için kullanılan bir regularizasyon yöntemidir. Bu hiperparametrenin temel etkileri şu şekildedir:

1. **Aşırı öğrenmeyi önleme**: Ağırlık azaltma, model parametrelerinin boyutunu sınırlayarak, modelin eğitim verisiyle aşırı uyum göstermesinin önüne geçer.
2. **Genelleme yeteneği**: Daha düşük ağırlıklar, modelin basit ama genel sonuçlar oluşturmasına vesile olur, bu da yeni ve görülmemiş dünya verilerinde daha yüksek performans göstermesine yardımcı olur.
3. **Model karmaşıklığını kontrol etme**: Ağırlık azaltma, modelin ürettiği sonuçların gereksiz bir şekilde karmaşıklaşmasını engeller ve daha yorumlanabilir sonuçlar üretilmesine yardımcı olur.
4. **Gradyan patlamasını önleme**: Özellikle derin ağlarda, ağırlık azaltma gradyan patlaması problemini azaltmaya yardımcı olur.

0,01 değeri, literatürde yaygın olarak kullanılan bir ağırlık azaltma değeridir. Mesela, Devlin ve arkadaşları BERT modelinin eğitiminde aynı ağırlık azaltma değerini kullanmışlardır [30]. Bunu yanı sıra, Brown ve arkadaşları GPT-3 modelinin eğitiminde de ağırlık azaltma tekniğinden faydalanmışlardır ve katsayı olarak *0,01* değerini kullanmışlardır [26].

Veri kümemizin büyüklüğü (321.179 soru-cevap çifti) ve kullanılan modellerin karmaşıklığına bakıldığında, 0,01'lik ağırlık azaltma değeri, modeller eğitilirken aşırı öğrenmelerini önleyecek ve modellerin gerçek dünya verisine uyumunu artıracaktır. Bu, bilhassa sağlık verileri gibi hassas bir alanda çalışırken modelin güvenilirliğini artırmak için mühimdir.

3.3.1.7 Optimizasyon algoritması

Çalışmamızda *AdamW* optimizasyon algoritmasının *8-bit* olan küçük bir versiyonu kullanılmıştır. Bu seçimin nedenleri ve avantajları şu şekildedir:

- **Eğitim hızı**: 8-bit AdamW, eğitimi hızlandırır. Bu, bilhassa büyük veri kümeleriyle, büyük ve karmaşık modeller üzerinde çalışırken mühim bir avantaj sağlar.

- **Bellek kullanımı**: Standart 32-bit optimizörlere göre bellek kullanımını hatrı sayılır derecede azaltır. Bu, sınırlı donanım kaynaklarıyla çalışırken veya daha büyük modeller eğitirken çok önemlidir. Bu çalışmada kullanılan modellerde 32-bit versiyonunu kullansaydık VRAM yetersiz kalacaktı.
- **Performans korunumu**: 8-bit optimizörler, 32-bit muadillerine göre performans konusunda önemsenecek bir düşüşe sebep olmaz. Dettmers ve arkadaşları tarafından yapılan çalışmada, 8-bit optimizörlerin geniş bir görev yelpazesinde 32-bit optimizörlerle eşdeğer performansa sahip olduğu kanıtlanmıştır [47].
- **Hiperparametre değişikliği gerektirmemesi**: 8-bit AdamW, standart AdamW'nin hiperparametrelerini kullanabilir, bu da geçiş sürecini kolaylaştırır.
- **Geniş uygulama alanı**: 8-bit optimizörler, doğal dil işleme, bilgisayarlı görü ve ses işleme gibi birçok alanda başarıyla uygulanmıştır.

Veri kümemizin büyüklüğü (321.179 soru-cevap çifti) ve kullanılan modellerin karmaşıklığı göz önüne alındığında, *8-bit AdamW* optimizörü, eğitimi gayet yeterli bir seviyede hızlandırırken devasa bellek kullanımını kabul edilebilir bir seviyeye getirmiştir. Aynı zamanda model performansından feragat etmememize olanak vermiştir. Bu, bilhassa bu çalışmada sınırlı hesaplama kaynaklarıyla çalışan bize önemli bir fayda sağlamıştır.

3.3.1.8 Hassasiyet formatı

Çalışmamızda hassasiyet formatı olarak *BF16* (bfloat16) kullanılmıştır. Bu seçimin nedenleri ve avantajları şu şekildedir:

- **Geniş dinamik aralık**: BF16, FP16'ya kıyasla daha kapsamlı bir temsil aralığı sunmaktadır [48].
- **FP32 ile uyumluluk**: BF16, FP32'nin 8 bit üstel kısmını koruyarak, FP32 ile benzer bir temsil aralığı sağlar. Bu durum, FP32 ile kolay dönüşüm ve uyumluluk sağlar.
- **Bellek kullanımı**: BF16, FP32'ye göre modelin boyutunu yarıya indirir. Bu da çalışmamız için oldukça elzemdir.
- **Hesaplama verimliliği**: BF16, FP32'ye göre daha hızlı eğitim ve cevap verme olanağı sunar, özellikle LLM eğitimlerinde performans artışı sağlar.

- **Üstün hassasiyet**: BF16, FP16'ya göre daha geniş bir temsil aralığı sağlar, bu da özellikle gradyan hesaplamalarında daha doğru sonuçlar elde edilmesine yardımcı olur.
- **Donanım desteği**: Günümüzde birsürü işlemci ve GPU, BF16 formatını doğrudan desteklemektedir.

Veri kümemizin büyüklüğü (321.179 soru-cevap çifti) ve kullanılan modellerin karmaşıklığına bakıldığında, BF16 formatının kullanılması, eğitimde hem bellek kullanımını azaltmamıza hem de hesaplama hızını artırmamıza vesile olmuştur. Bu, bilhassa bu çalışmada sınırlı hesaplama kaynaklarıyla çalışan bize önemli bir fayda sağlamıştır.

3.3.2 Eğitim Stratejisi

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

- **LoRA Kullanımı**: LoRA tekniği, modelin tüm parametrelerini güncellemek yerine, düşük boyutlu matrisler ekleyerek eğitimi hızlandırır ve bellek kullanımını minimize eder [49].
- **Unsloth Kütüphanesi**: Bu kütüphane, ince ayar sürecindeki bellek kullanımını azaltmak ve eğitim sürecini hızlandırmak için kullanılmıştır [50].
- **Veri Kümesi Hazırlığı**: Eğitim öncesinde veri seti, modelin giriş formatına uygun şekilde hazırlanmıştır.
- **Değerlendirme Stratejisi**: Model performansı *loss* metriğiyle, 57.800 adımda bir değerlendirilmiş ve kaydedilmiştir.
- **Kaydetme Stratejisi**: Model, 10.000 adımda bir eğitim yapılan cihazın kapanma durumuna karşı kaydedilmiştir.
- **Donanım**: Eğitim süreci, Google Colab platformunda Tesla A100 40 GB GPU versiyonu kullanılarak gerçekleştirilmiştir [51]. Bu güçlü donanım, LLM'lerin eğitimi için gerekli işlem gücünü ve VRAM'i sağlamıştır [52].
- **Eğitim Süreleri**: Modellerin eğitim süreleri şu şekildedir:
  - Cosmos LLama: 19,4 saat
  - Meta LLama: 11,38 saat

- SambaLingo: 17,46 saat
- Trendyol: 19,3 saat

4
PERFORMANS DEĞERLENDİRME

Bu bölümde, ince ayar yapılan dil modellerinin başarımlarını çeşitli yönlerden değerlendireceğiz. Performans değerlendirmesi, bir modelin başarısını ölçmek için gayet önemli bir aşamadır. Bu süreç, modelin güçlü ve zayıf taraflarını belirlememize, iyileştirme alanlarını tespit etmemize ve farklı modeller arasında daha doğru ve nesnel karşılaştırma yapabilmemize vesile olur. Performans değerlendirmesini, eğitim sürecinden başlayarak en son model çıktılarına kadar kapsamlı bir yelpazede gerçekleştirdik. Bu bölümde, eğitim sürecindeki performans ölçütlerinden başlayarak, birçok ölçüm metriğini, Elo puanlamasını, kazanma yüzdesi analizini ve uzman değerlendirmeleri sonuçlarını ele alacağız. Eğitim sürecinde başarı ölçümü, modelin öğrenme başarısını ve eğitimin etkisini anlayabilmek için mühimdir. Bu adımda, eğitim kaybı (train loss) ve doğrulama kaybı (validation loss) ölçütlerini inceleyeceğiz. Sentetik değerlendirme ölçütleri kısmında, *ROUGE, BLEU, BERT Skor, CER, WER ve METEOR* gibi yaygın kullanılan ölçütler vesilesiyle modellerin başarımını gözlemleyeceğiz. Bu ölçütler, modellerin ürettiği sonuçların referans sonuçlarla ne kadar benzer olduğunu farklı açılardan değerlendirmemize vesile olur. Modellerin performansını ölçmemizdeki en büyük sıkıntı, hasta-doktor yazılı iletişimine özel ve Türkçe sentetik bir performans ölçme aracının henüz mevcut olmaması. Aynı zamanda bu alanda standart bir ölçütün de bulunmaması. Bundan dolayı farklı alanlarda kullanılan birçok ölçütü kullanmaya çalıştık. Bununla beraber Elo puanlaması ve kazanma yüzdesi skorlarını hesaplarken, gelişmiş devasa LLM'leri hakem olarak kullanmayı tercih ettik. Bu hedefle kullanılan modeller arasında *Claude 3.5 Sonnet, GPT-4o, LLaMA 3.1 70B, Microsoft Copilot ve Gemini 1.5 Pro* bulunmaktadır [21–24, 53]. Bu gelişmiş modeller, ince ayar yapılan modellerin çıktılarını değerlendirerek, daha objektif ve kapsamlı bir kıyaslama yapabilmemize vesile olmuştur. Bu modelleri seçerken herkese açık olan ve LLM'lerin sıralandığı *LMSYS Chatbot Arena Leaderboard*'ı referans aldık [54]. Elo puanlaması ve kazanma yüzdesi ölçümü, modellerin birebir karşılaştırmalı başarımını değerlendirmek için kullanılacaktır. Gelişmiş LLM'lerin hakemliğinde yapılan bu ölçümler, bilhassa hasta-doktor

etkileşimi gibi ayrıntılı bir alanda, modellerin performansını daha hassas bir şekilde değerlendirmemize vesile olmuştur. Son olarak, uzman değerlendirmesi bölümünde, çeşitli branşlarda çalışan doktorlar tarafından yapılan, modellerin cevaplarını puanlamalarının ortalamalarını inceleyeceğiz. Bu ölçüt, modellerin oluşturduğu cevabın doğruluğunu, tutarlılığını, kapsamlılığını ve zararlılığını alan uzmanı insanların bakış açısından değerlendirmemize vesile olur. Bu gayet kapsamlı performans ölçümü, geliştirilen dil modellerinin başarısını çok yönlü olarak değerlendirmemize vesile olacaktır ve gelecekteki iyileştirmeler için rehber niteliğindedir. Ayrıca, hasta-doktor etkileşimi gibi özel alanlarda başarım ölçme yöntemlerinin oluşturulması ve geliştirilmesi hususunda da ufuk açacaktır.

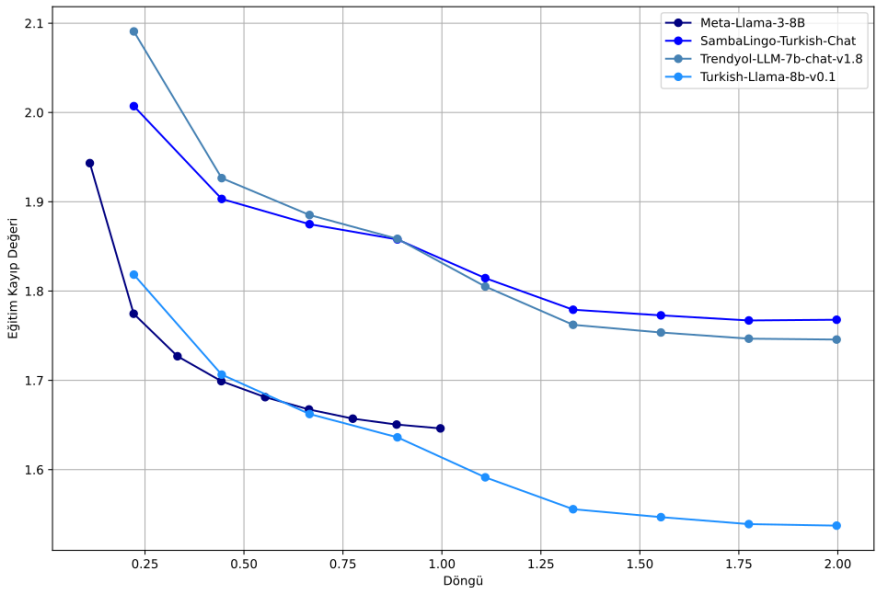
4.1 Eğitim Sürecinde Performans Değerlendirmesi

Eğitim sürecindeki performans ölçümü, modelin eğitim sürecini ve genel başarısını analiz edebilmek için çok önemlidir. Bu değerlendirme, modelin düzgün eğitilip eğitilmediğini ve aşırı öğrenme (overfitting) gibi sorunların olup olmadığını belirlememize vesile olur.

4.1.1 Eğitim Kaybı (Training Loss)

Eğitim kaybı, modelin eğitim veri kümesinde, eğitim sırasındaki performansını ölçen önemli bir ölçüttür. Bu ölçğt, modelin eğitim sürecinin nasıl ilerlediğini ve öğrenme hızını gösterir. Şekil **??**'te görüldüğü üzere, dört farklı model için eğitim kaybı değerlerinin döngü sayısına göre değişimi gösterilmektedir.

Trendyol-LLM-7b-chat-v1.8 modeli için eğitim kaybı *2,0908*'den başlayarak *1,7457*'ye kadar düşmüştür. Bu düşüş, modelin eğitim verilerini öğrenme sürecinde belli bir ilerleme katettiğini ortaya koymaktadır. Benzer şekilde, **SambaLingo-Turkish-Chat** modeli *2,0071*'den *1,7679*'a, **Turkish-Llama-8b-v0.1** modeli *1,8184*'ten *1,5373*'e ve **Meta-Llama-3-8B** modeli *1,9433*'ten *1,6462*'ye düşüş göstermiştir. Bu düşüşler, tüm modellerin eğitilirken belli seviyede öğrendiklerini göstermektedir. Eğitim kaybındaki düşüş hızı ve boyutu, modellerin öğrenme kapasitesini ve eğitim sürecinin kalitesini gösterir. Mesela, **Meta-Llama-3-8B** modelinin eğitim kaybındaki düşüş diğer modellere göre daha hızlı olmuştur, bu da modelin veri kümesini daha hızlı öğrendiğini gösterebilir. Lakin, sadece eğitim kaybını göz önünde bulundurarak modelin genel performansı için kesin bir yargıya varmak yanlış olur. Eğitim kaybının yanında, doğrulama kaybı ve diğer ölçütler de önemsenmelidir. Ayrıca, çok düşük eğitim kaybı değerleri, modelin aşırı öğreniyor (overfitting) olduğunu da gösterebilir. Bundan dolayı, eğitim kaybı ile birlikte



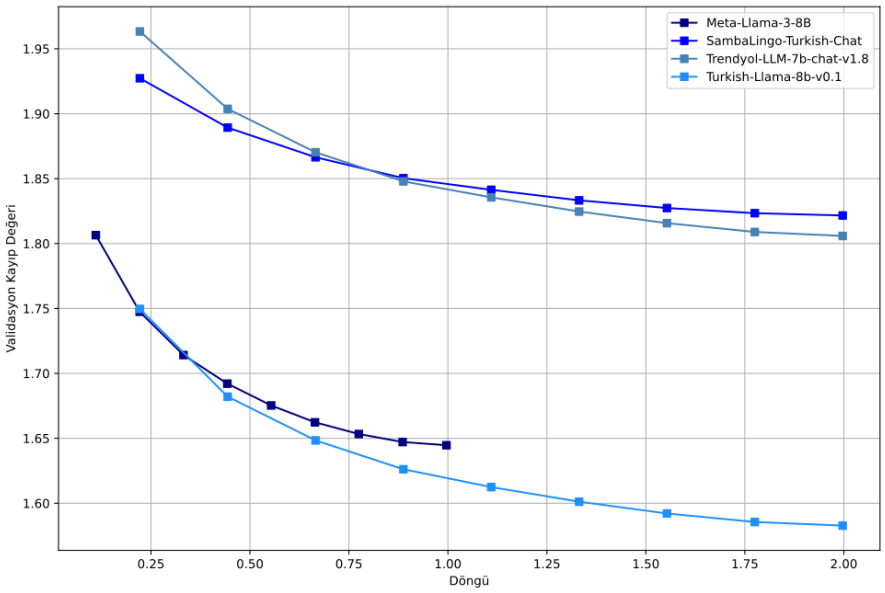
Şekil 4.1 Eğitim kaybı grafiği

doğrulama kaybının da incelenmesi lazımdır.

4.1.2 Doğrulama Kaybı (Validation Loss)

Doğrulama kaybı, modelin eğitim sırasında görmediği veri seti üzerindeki performansını ölçen önemli bir metriktir. Bu metrik, modelin genelleme yeteneğini değerlendirmek ve aşırı öğrenme (overfitting) olup olmadığını tespit etmek için kullanılır. Doğrulama işlemi için toplam *28.906* veri kullanılmıştır. Şekil 4.2'te görüldüğü üzere, dört farklı model için doğrulama kaybı değerlerinin döngü sayısına göre değişimi gösterilmektedir.

Trendyol-LLM-7b-chat-v1.8 modeli için doğrulama kaybı *1,9634*'ten başlayarak *1,8059*'a düşmüştür. Bu düşüş, modelin genelleme performansının yükseldiğini göstermektedir. Benzer şekilde, **SambaLingo-Turkish-Chat** modeli *1,9273*'ten *1,8216*'ya, **Turkish-Llama-8b-v0.1** modeli *1,7496*'dan *1,5828*'e ve **Meta-Llama-3-8B** modeli *1,8065*'ten *1,6447*'ye düşüş göstermiştir. Bu düşüş eğilimleri, tüm modellerin eğitim esnasında genelleme performanslarının iyileştiğini göstermektedir. Doğrulama kaybındaki düşüş hızı ve miktarı, modellerin öğrenme ve genelleme kapasitesi hakkında fikir verebilir. Mesela, **Meta-Llama-3-8B** modelinin doğrulama kaybındaki düşüş diğer iki modele göre göre daha hızlı ve istikrarlı olmuştur, bu da modelin daha hızlı öğrendiğini ve genelleme kapasitesinin daha iyi olduğunu gösterebilir. Ama, yalnızca doğrulama kaybına bakarak modelin



Şekil 4.2 Doğrulama kaybı grafiği

genel performansı hakkında kesin bir sonuca varmak yanlış olur. Eğitim kaybı azalırken, doğrulama kaybının daha düşük seviyede azalması ya da azalmaması, modelin aşırı öğrendiğinin göstergesidir. Bundan dolayı, eğitim kaybı ile doğrulama kaybını birlikte değerlendirmek gerekir. **Trendyol-LLM-7b-chat-v1.8** modelinde eğitim kaybı ile doğrulama kaybı arasındaki farkın son döngülerde yükselmesi, bu modelde az bir aşırı öğrenmeye eğilimli olabileceğini ve döngüyü sonlandırmamız gerektiğini göstermektedir. Diğer taraftan, **Turkish-Llama-8b-v0.1** ve **Meta-Llama-3-8B** modellerinde eğitim ve doğrulama kayıpları arasındaki farkın görece sabit kalması, bu modellerin daha iyi genelleme yaptığını ve aşırı öğrenme riskinin daha az olduğunu ortaya koymaktadır.

4.1.3 Öğrenme Eğrisi (Learning Curve)

Öğrenme eğrisi, modellerin eğitim ve doğrulama kayıplarının eğitim süreci boyunca nasıl değiştiğini gösteren bir grafiktir. Bu eğri, modelin öğrenme hızını ve varsa aşırı öğrenme durumlarını analiz etmemize yardımcı olur. Bunun için yukarıda incelediğimiz Şekil 4.2 ve Şekil 4.1'i beraber değerlendirmemiz gerekir. **Trendyol-LLM-7b-chat-v1.8** modelinin öğrenme eğrisi incelendiğinde, eğitim ve doğrulama kayıplarının en başta hızlı bir düşüş eğiliminde olduğu, ama sonraki döngülerde düşüş hızının yavaşladığı görülmektedir. Bu, modelin başlangıçta beklendiği gibi hızlı öğrendiğini, ama zamanla öğrenme hızının beklendiği gibi azaldığını gösterir. Eğitim ve doğrulama kayıpları arasındaki farkın son döngülerde çok

az artması, modelde küçük bir aşırı öğrenme eğilimi olabileceğini gösterebilir. **SambaLingo-Turkish-Chat** modelinin öğrenme eğrisi, eğitim ve doğrulama kayıplarının nispeten daha yavaş ve istikrarlı bir düşüş gösterdiğini göstermektedir. Bu, modelin daha yavaş ama emin adımlarla öğrendiğini gösterir. Eğitim ve doğrulama kayıpları arasındaki farkın büyük ölçüde sabit kalması, modelin eğitim sonucunda, görece daha iyi bir genelleme performansı kaznadığını gösterir. **Turkish-Llama-8b-v0.1** modelinin öğrenme eğrisi, hem eğitim hem de doğrulama kayıplarında istikrarlı bir düşüşü gözler önüne sermektedir. Bu model, en düşük seviyede son doğrulama kaybına ulaşmıştır, bu da modelin gayet iyi bir genelleme yeteneği kazandığını gösterir. Eğitim ve doğrulama kayıpları arasındaki farkın minimal olması, modelin aşırı öğrenme riskinin olmadığını ve eğitim verilerini iyi bir şekilde öğrendiğini gösterir. **Meta-Llama-3-8B** modelinin öğrenme eğrisi, başlangıçta hızlı bir düşüşten sonra daha yavaş bir ilerleme göstermektedir. Bu model, eğitim ve doğrulama kayıpları arasında gayet tutarlı bir fark elde ederek, öğrenme sürecinde iyi seviyede genelleme özelliği kazanmıştır. Eğitim sürecinin sonlarına doğru kayıpların neredeyse sabit kalması, modelin en iyi noktaya çok yakın olduğunu gösterir. Genel olarak, tüm modellerin öğrenme eğrileri, eğitim sürecinde istikrarlı bir ilerleme göstermektedir.

4.1.4 Eğitim Süresi Analizi

Eğitim süresi analizi, modellerin eğitim süresinin ne kadar sürdüğünü ve bununla beraber kaynak kullanımını ölçebilmek için gereklidir. Bu analiz, her modelin toplam eğitim zamanını içerir. Eğitim süreleri şu şekilde gerçekleşmiştir:

- **Trendyol-LLM-7b-chat-v1.8**: 2 döngü, toplam *19,3* saat
- **SambaLingo-Turkish-Chat**: 2 döngü, toplam *17,46* saat
- **Turkish-Llama-8b-v0.1**: 2 döngü, toplam *19,4* saat
- **Meta-Llama-3-8B**: 1 döngü, toplam *11,38* saat

Meta-Llama-3-8B modelinin tek döngüde eğitilmesinden dolayı eğitim süresi diğer modellere göre çok daha kısa sürede tamamlanmıştır. **Turkish-Llama-8b-v0.1** modelinin en uzun eğitim süresine sahip olması, modelin iki döngü eğitilmesinden, modelin karmaşıklığının ve parametre sayısının fazla olmasından kaynaklanmaktadır. **Trendyol-LLM-7b-chat-v1.8** ve **Turkish-Llama-8b-v0.1** modelleri eğitimlerini neredeyse aynı sürede tamamlamıştır. Bu durum, bu iki modelin benzer karmaşıklıkta olduğunu ve benzer hesaplama kaynakları

gerektirdiğini ortaya koymaktadır. **SambaLingo-Turkish-Chat** modeli, diğer iki 2 döngülü modellere göre biraz daha kısa sürede eğitilmiştir. Bu, modelin az da olsa daha verimli bir mimariye sahip olduğunu ya da modelin eğitiminin daha optimize edilmiş olduğunu gösterir.

4.2 Sentetik Değerlendirme Ölçütleri

Sentetik değerlendirme ölçütleri, dil modellerinin başarımlarını objektif ve nicel olarak değerlendirmek için kullanılan metriklerdir. Bu ölçütler, modellerin ürettiği metinlerin kalitesini, doğruluğunu ve insan tarafından üretilen referans metinlere olan benzerliğini ölçer. Bu çalışmada, dört farklı büyük dil modelinin performansını değerlendirmek amacıyla toplam *32.116* test verisi ve çeşitli sentetik ölçütler kullanılmıştır:

- ROUGE
- BLEU
- BERT Skor
- CER
- WER
- METEOR

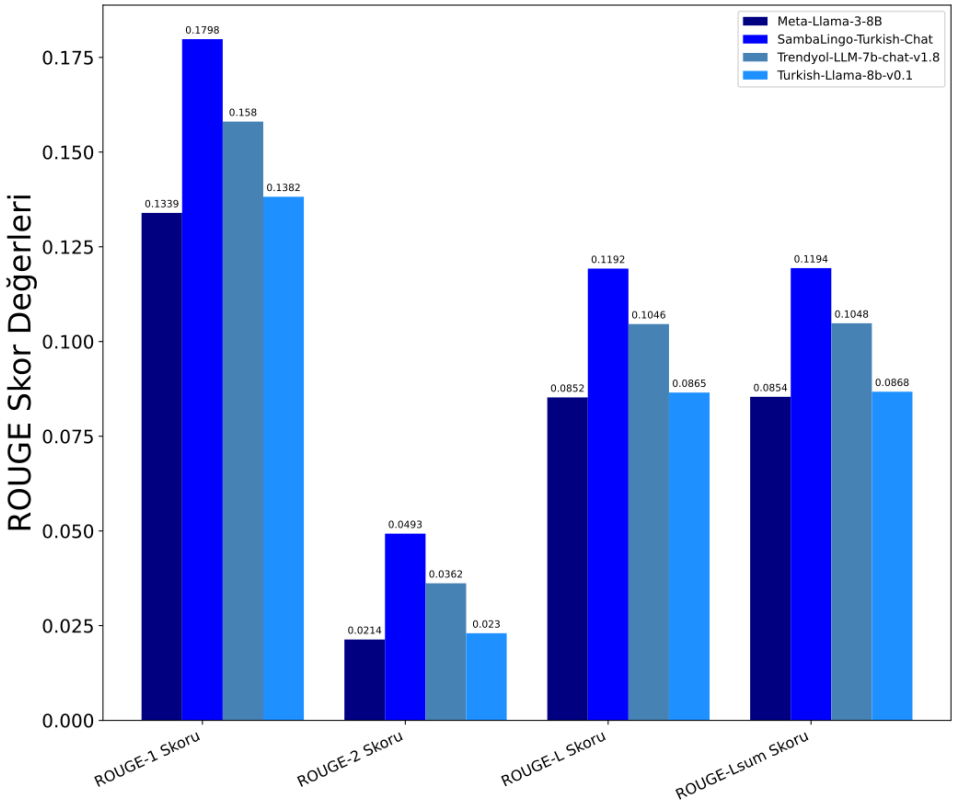
Bu ölçütlerin her biri, farklı açılardan metin değerlendirmesine imkan verir. Mesela, ROUGE ve BLEU daha çok metin özetleme ve makine çevirisi işlerinde kullanılırken, CER ve WER daha çok konuşma tanıma ve metin düzeltme işlerinde kullanılır. BERT Skor ise daha güncel bir ölçüt olup, kelimelerin bağlamsal alakalarını da hesaba katar. Bu sentetik ölçütlerin kullanılmasının temel amacı, modellerin performansını objektif ve tekrarlanabilir bir şekilde değerlendirmektir. Lakin, bu ölçütlerin her birinin kendi sınırlamaları ve güçsüz yönleri olduğunu da unutmamak lazım. Bundan dolayı, tek bir ölçüte bel bağlamaktansa, birden çok ölçütü bir arada kullanmak daha kapsamlı ve güvenilir bir değerlendirme yapmamıza yardımcı olur. Bununla beraber, bu sentetik ölçütlerin yanı sıra, insan değerlendirmesi ve gerçek dünya uygulamalarındaki performans gibi diğer değişkenlerin de dikkate alınması lazımdır. Bu şekilde, modellerin gerçek kullanım senaryolarındaki başarımı daha iyi anlaşılabilir.

4.2.1 ROUGE

ROUGE, metin özetleme ve makine çevirisi gibi DDİ işlerinde kullanılan bir değerlendirme ölçütüdür [55]. ROUGE, üretilen metni bir ya da birden çok referans metinle karşılaştırarak, üretilen metnin kalitesini ölçer. Bu ölçüt, özellikle n-gram örtüşmelerine dayalı olarak çalışır ve farklı varyasyonları bulunmaktadır.

ROUGE ölçütü, Türkçe dil modelleri üzerine yapılan çeşitli çalışmalarda da yaygın olarak kullanılmıştır. Kesgin ve arkadaşları, Doğan ve arkadaşları, Uludoğan ve arkadaşları çalışmalarında ROUGE metriğini kullanmışlardır [16, 19, 20].

Şekil 4.3'te görüldüğü üzere, dört farklı model için ROUGE skorları karşılaştırılmıştır. ROUGE-1, ROUGE-2, ROUGE-L ve ROUGE-Lsum olmak üzere dört farklı ROUGE ölçütü kullanılmıştır.



Şekil 4.3 ROUGE Skorları

4.2.1.1 ROUGE-1

Şekil 4.3'e göre ROUGE-1 skorları şu şekildedir: **SambaLingo-Turkish-Chat** modeli *0,1798* puan ile birinci sıradadır. Hemen ardından *0,1580* puan ile **Trendyol-LLM-7b-chat-v1.8** modeli gelmektedir. Üçüncü olarak **Turkish-**

Llama-8b-v0.1 modeli *0,1382* puan almıştır. **Meta-Llama-3-8B** modeli *0,1339* puan ile sonuncu olmuştur. Bu sonuçlar, **SambaLingo-Turkish-Chat** modelinin tekli kelime eşleşmelerinde diğer modellere göre açık ara daha başarılı olduğunu gözler önüne sermektedir.

4.2.1.2 ROUGE-2

Şekil 4.3'e göre ROUGE-2 skorları şu şekildedir: **SambaLingo-Turkish-Chat** modeli *0,0493* puan ile birinci sıradadır. Hemen ardından *0,0362* puan ile **Trendyol-LLM-7b-chat-v1.8** modeli gelmektedir. Üçüncü olarak **Turkish-Llama-8b-v0.1** modeli *0,0230* puan almıştır. **Meta-Llama-3-8B** modeli *0,0214* puan ile sonuncu olmuştur. Bu sonuçlar **SambaLingo-Turkish-Chat** modelinin diğer modellere kıyasla bu alanda da üstünlüğünü koruduğu bariz bir şekilde görülmektedir.

4.2.1.3 ROUGE-L

Şekil 4.3'e göre ROUGE-L skorları şu şekildedir: **SambaLingo-Turkish-Chat** modeli *0,1192* puan ile birinci sıradadır. Hemen ardından *0,1046* puan ile **Trendyol-LLM-7b-chat-v1.8** modeli gelmektedir. Üçüncü olarak **Turkish-Llama-8b-v0.1** modeli *0,0865* puan almıştır. **Meta-Llama-3-8B** modeli *0,0852* puan ile sonuncu olmuştur. Bu sonuçlar, **SambaLingo-Turkish-Chat** modelinin daha uzun ve tutarlı metin parçaları üretmede diğer modellere daha başarılı olduğunu göstermektedir, ama bu sefer **Trendyol-LLM-7b-chat-v1.8** modeli hemen ardından gelmektedir.

4.2.1.4 ROUGE-Lsum

Şekil 4.3'e göre ROUGE-Lsum skorları şu şekildedir: **SambaLingo-Turkish-Chat** modeli *0,1194* puan ile birinci sıradadır. Hemen ardından *0,1048* puan ile **Trendyol-LLM-7b-chat-v1.8** modeli gelmektedir. Üçüncü olarak **Turkish-Llama-8b-v0.1** modeli *0,0868* puan almıştır. **Meta-Llama-3-8B** modeli *0,0854* puan ile sonuncu olmuştur. Bu sonuçlar, modellerin ürettiği metinlerin yapısal olarak tutarlı olduğunu göstermekte ve **SambaLingo-Turkish-Chat** modelinin benzer şekilde bu alanda da üstünlüğünü koruduğunu açıkça ortaya koymaktadır. **Trendyol-LLM-7b-chat-v1.8** modeli ROUGE-L skorunda olduğu gibi az bir farkla ikinci olmuştur.

4.2.1.5 Genel Değerlendirme

Bu sonuçlara bakıldığında, **SambaLingo-Turkish-Chat** modelinin tüm ROUGE metriklerinin tamamında en yüksek skorları elde ettiği ve diğer modellerden bariz şekilde daha iyi performans gösterdiği görülmektedir. **Trendyol-LLM-7b-chat-v1.8** modeli, **SambaLingo-Turkish-Chat** modelinin hemen ardından en iyi performansı sergilemiş ve ikinci sırada yer almıştır. **Turkish-Llama-8b-v0.1** ve **Meta-Llama-3-8B** modelleri ise benzer performans göstermiş, lakin diğer modellere göre daha düşük puanlar elde ederek sırasıyla üçüncü ve dördüncü olmuşlardır.

4.2.2 BLEU

BLEU, makine çeviris ve metin üretimi gibi işler yapan modellerin başarımını değerlendirmek için kullanılan önemli bir ölçüttür [56]. Bu ölçüt, oluşturulan metni referans metin ya da metinlerle kıyaslayarak, n-gram örtüşmelerini hesaplar ve 0 ila 1 arasında bir değer üretir. Daha yüksek BLEU skoru, oluşturulan metnin referans metinlere daha yakın olduğunu gösterir.

BLEU ölçütü, Türkçe doğal dil işleme çalışmalarında da yaygın olarak kullanılmaktadır. Örneğin, Akyon ve arkadaşları ile Uludoğan ve arkadaşları çalışmalarında BLEU ölçütünü kullanmışlardır [16, 44].

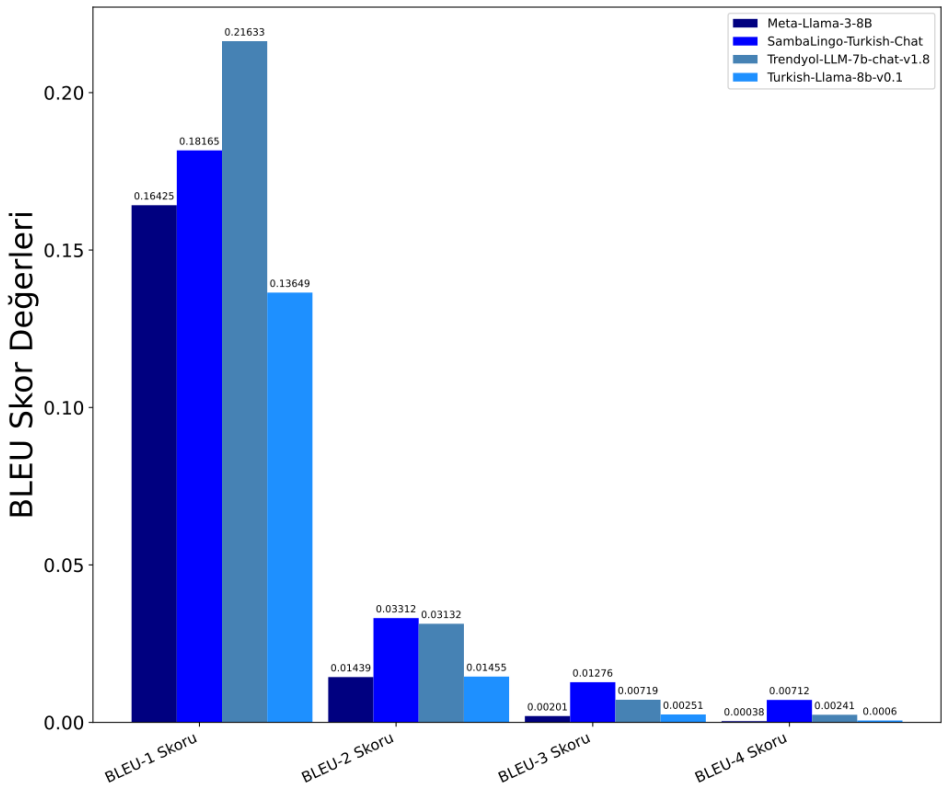
4.2.2.1 BLEU Skoru

Şekil 4.4'e göre, **SambaLingo-Turkish-Chat** modeli *0,0248* puan ile birinci sıradadır. Hemen ardından *0,0125* puan ile **Trendyol-LLM-7b-chat-v1.8** modeli gelmektedir. Üçüncü olarak **Turkish-Llama-8b-v0.1** modeli *0,0073* puan almıştır. **Meta-Llama-3-8B** modeli *0,0056* puan ile sonuncu olmuştur. Bu sonuçlar, **SambaLingo-Turkish-Chat** modelinin diğer modellere kıyasla daha kaliteli çıktı oluşturduğunu ortaya koymaktadır.

4.2.2.2 Kesinlik (Precision) Skorları

Şekil 4.4'e göre kesinlik skorları şu şekildedir:

- 1-gram:
  - **Trendyol-LLM-7b-chat-v1.8**: *0,2163*
  - **SambaLingo-Turkish-Chat**: *0,1816*
  - **Meta-Llama-3-8B**: *0,1642*



Şekil 4.4 BLEU Skorları

    - **Turkish-Llama-8b-v0.1**: *0,1365*

- 2-gram:
    - **SambaLingo-Turkish-Chat**: *0,0331*
    - **Trendyol-LLM-7b-chat-v1.8**: *0,0313*
    - **Turkish-Llama-8b-v0.1**: *0,0145*
    - **Meta-Llama-3-8B**: *0,0144*
- 3-gram:
    - **SambaLingo-Turkish-Chat**: *0,0128*
    - **Trendyol-LLM-7b-chat-v1.8**: *0,0072*
    - **Turkish-Llama-8b-v0.1**: *0,0025*
    - **Meta-Llama-3-8B**: *0,0020*
- 4-gram:
    - **SambaLingo-Turkish-Chat**: *0,0071*
    - **Trendyol-LLM-7b-chat-v1.8**: *0,0024*
    - **Turkish-Llama-8b-v0.1**: *0,0006*
    - **Meta-Llama-3-8B**: *0,0004*

Bu sonuçlara göre, **SambaLingo-Turkish-Chat** modeli 1-gram haricindeki n-gram seviyelerde en yüksek kesinlik skoruna sahiptir. Lakin 1-gram seviyesinde **Trendyol-LLM-7b-chat-v1.8** modeli onun tahtını sarsmıştır. 2-gram seviyesinde, **SambaLingo-Turkish-Chat** modeli *0,0331* puanla liderliği ele geçirmiştir. Bu skor, ikinci sıradaki **Trendyol-LLM-7b-chat-v1.8** modelinin *0,0313* puanından yaklaşık %5,75 daha yüksektir. 3-gram seviyesinde ise **SambaLingo-Turkish-Chat**, *0,0128* puanla ikinci sıradaki **Trendyol-LLM-7b-chat-v1.8** modelinin *0,0072* puanını neredeyse ikiye katlayarak (%77,78 daha yüksek) büyük bir fark oluşturmuştur. 4-gram seviyesinde fark iyice açılmıştır. **SambaLingo-Turkish-Chat** modeli *0,0071* puanla, ikinci sıradaki **Trendyol-LLM-7b-chat-v1.8** modelinin *0,0024* puanına göre yaklaşık %195,83 daha yüksek bir puan elde etmiştir. Bu sonuç, **SambaLingo-Turkish-Chat** modelinin daha uzun kelime dizilerini daha düzgün şekilde oluşturma konusunda diğer modellere göre başarısını bariz şekilde ortaya koymaktadır. **Turkish-Llama-8b-v0.1** ve **Meta-Llama-3-8B** modelleri ise tüm n-gram seviyelerinde son iki sırayı paylaşmışlardır. Bu modeller arasında belirgin bir performans farkı yoktur, lakin çoğu durumda **Turkish-Llama-8b-v0.1** modeli **Meta-Llama-3-8B** modelinden çok az daha iyi başarım

göstermiştir. Genel olarak, bu sonuçlar **SambaLingo-Turkish-Chat** modelinin, bilhassa daha uzun kelime dizilerini doğru bir şekilde üretme konusunda, diğer modellere göre bariz bir başarı sağladığını göstermektedir. Bu da modelin, daha karışık ve tutarlı cümle oluşturma özelliğinin çok daha iyi olduğunu göstermektedir.

4.2.2.3 Brevity Penalty

BP, BLEU ölçütünün mühim bir parçasıdır. Bu ceza faktörü, makine çevirisi sistemlerinin çok kısa çeviriler üreterek yüksek puanlar elde etmesini engelleyebilmek amacıyla tasarlanmıştır. Brevity Penalty şu şekilde hesaplanır:

1. Eğer üretilen çeviri, referans çeviriden daha uzunsa, BP 1 olur (yani ceza uygulanmaz).
2. Eğer üretilen çeviri, referans çeviriden daha kısaysa, BP değeri şu formülle elde edilir:

$$BP = e^{(1-\frac{r}{c})}$$

Burada:

- r: referans çevirinin uzunluğu
- c: üretilen çevirinin uzunluğu

Brevity Penalty'nin amacı, sistemin oluşturduğu çevirinin uzunluğunun referans çeviri uzunluğuna mümkün olduğunca yakın olmasını sağlamaktır. Bu, sistemin mühim bilgileri es geçmemesini ve aynı zamanda gereksiz uzunlukta çeviriler oluşturmamasına olanak tanır.

Şekil 4.4'e göre brevity penalty skorları aşağıdaki tabloda gösterilmiştir:

Model	**Brevity Penalty Skoru**
Turkish-Llama-8b-v0.1	0,9804
SambaLingo-Turkish-Chat	0,9118
Meta-Llama-3-8B	0,8512
Trendyol-LLM-7b-chat-v1.8	0,6750

Tablo 4.1 Modellerin Brevity Penalty Skorları

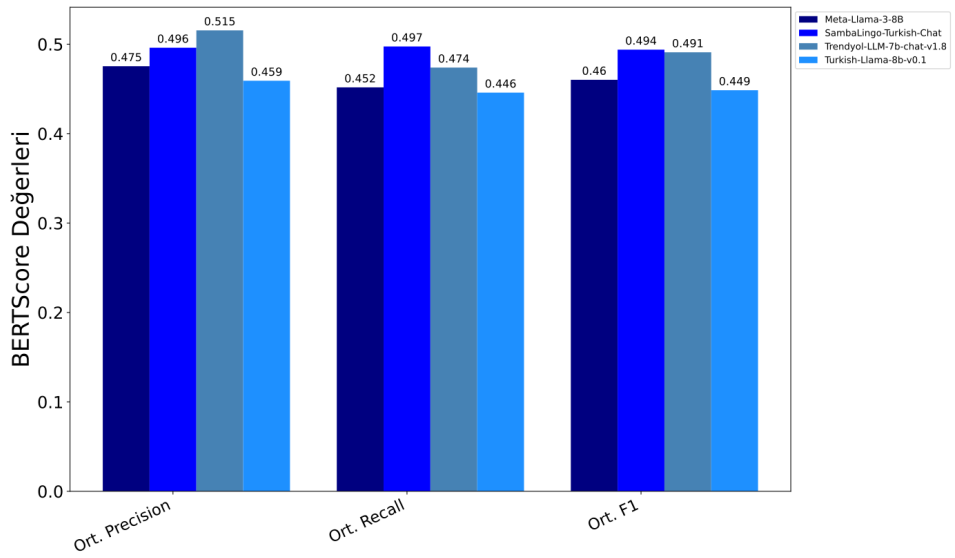
Elde edilen bu sonuçlara göre, **Turkish-Llama-8b-v0.1** modeli *0,9804* puan ile, oluşturduğu metin uzunluğu referans metnin uzunluğuna en yakın modeldir. Onun hemen ardından *0,9118* puanla **SambaLingo-Turkish-Chat** modeli gelirken, **Trendyol-LLM-7b-chat-v1.8** modeli en düşük puanı elde etmiştir.

4.2.2.4 Genel Değerlendirme

Bu BLEU analizine göre, **SambaLingo-Turkish-Chat** modeli genel manada en iyi performansı göstermiştir. Bu model, bilhassa yüksek n-gram seviyelerinde diğer modellere göre daha başarılı olmuştur. **Trendyol-LLM-7b-chat-v1.8** modeli, bilhassa tekli kelime eşleşmelerinde (1-gram) en iyi performansa sahiptir. **Turkish-Llama-8b-v0.1** ve **Meta-Llama-3-8B** modelleri ise genel olarak daha düşük puanlar elde etmişlerdir, lakin **Turkish-Llama-8b-v0.1** modeli metin uzunluğu açısından en tutarlı model olarak öne çıkmıştır.

4.2.3 BERT Skor

BERT Score [57], metin üretimi ölçümünde kullanılan ve önceden eğitilmiş BERT modellerini kullanan bir ölçüttür. Bu ölçüt, oluşturulan cümleler ile referans cümleler arasındaki anlamsal yakınlığı ölçer. BERT Skor, BLEU gibi geleneksel ölçütlerin bazı engellerini aşarak, yalnızca yüzeysel ölçüm yerine derin anlamsal benzerliği de yakalamaya çalışır. BERT Skor DDİ alanında yaygın olarak kullanılmaktadır. Mesela, Zhang ve arkadaşları [58] haber özetleme görevinde büyük dil modellerini n başarımını ölçerken BERT Skor'u kullanmışlardır. Benzer şekilde, Gehrmann ve arkadaşları [59] çoklu dilde metin üretimi değerlendirme ölçütü olarak BERT Skor'u kullanmıştır.



Şekil 4.5 BERT Skor Sonuçları

4.2.3.1 Ortalama F1 Skoru

Şekil 4.5'e göre BERT Skor F1 sonuçları şu şekildedir: **SambaLingo-Turkish-Chat** modeli *0,49386* puan ile birinci sıradadır. Hemen ardından *0,49094* puan ile **Trendyol-LLM-7b-chat-v1.8** modeli gelmektedir. Üçüncü olarak **Meta-Llama-3-8B** modeli *0,46016* puan almıştır. **Turkish-Llama-8b-v0.1** modeli *0,44852* puan ile sonuncu olmuştur. Bu sonuçlar, **SambaLingo-Turkish-Chat** modelinin oluşturduğu metinlerin, anlamsal olarak referans metinlere en yakın metinler olduğunu göstermektedir. F1 skoru, precision ve recall'ın harmonik ortalaması olduğundan dolayı, bu sonuçlar **SambaLingo-Turkish-Chat** modelinin doğruluk ve kapsamlılık açısından en dengeli başarımı sergilediğini ortaya koymaktadır. **Trendyol-LLM-7b-chat-v1.8** modelinin ufak bir farkla ikinci sırada yer olması, bu iki modelin Türkçe metin üretiminde yakın başarımlara sahip olduğunu düşündürmektedir.

4.2.3.2 Ortalama Precision Skoru

BERT Skor Precision sonuçlarına göre, **Trendyol-LLM-7b-chat-v1.8** modeli *0,51548* puan ile en yüksek skoru elde etmiştir. İkinci sırada *0,49604* puan ile **SambaLingo-Turkish-Chat** modeli yer almaktadır. **Meta-Llama-3-8B** modeli *0,47536* puan ile üçüncü sırada, **Turkish-Llama-8b-v0.1** modeli ise *0,45910* puan ile son sırada yer almıştır. Precision skoru, modelin ürettiği metinlerin ne kadar doğru olduğunu gösterir. Bu puanlara göre, **Trendyol-LLM-7b-chat-v1.8** modeli diğer modellere göre daha doğru ve alakalı cevaplar üretiyor denilebilir. Bu durum, modelin Türkçe sağlık terimleri ve mefhumları konusunda daha hassas olduğunu göstermektedir.

4.2.3.3 Ortalama Recall Skoru

BERT Skor Recall sonuçlarında ise **SambaLingo-Turkish-Chat** modeli *0,49745* puan ile en yüksek performansı göstermiştir. **Trendyol-LLM-7b-chat-v1.8** modeli *0,47385* puan ile ikinci sırada yer alırken, **Meta-Llama-3-8B** modeli *0,45170* puan ile üçüncü ve **Turkish-Llama-8b-v0.1** modeli *0,44583* puan ile son sırada yer almıştır. Recall skorunun yüksekliği, modelin cevaplarının, referans metindeki bilgileri ne kadar kapsadığını gösterir. **SambaLingo-Turkish-Chat** modelinin bu en yüksek puanı alması, modelin diğer modellere göre sağlık ile ilgili sorulara daha kapsamlı yanıtlar ürettiğini göstermektedir.

4.2.3.4 Genel Değerlendirme

Tüm modellerin BERT Skor değerlerinin 0,44'ün üzerinde olması, genel olarak iyi bir performans ortaya koyduklarını göstermektedir. Lakin, modeller arasındaki farklar, özellikle Türkçe dil anlama ve oluşturma kabiliyetlerindeki farklar hakkında da fikir vermektedir. **SambaLingo-Turkish-Chat** ve **Trendyol-LLM-7b-chat-v1.8** modellerinin özellikle F1 ve Recall skorlarında verdiği iyi sonuçlar, bu modellerin Türkçe metin üretiminde daha tercih edilebilir olduklarını göstermektedir.

4.2.4 WER

WER, metin tanıma sistemlerinin performansını ölçmek için yaygın olarak kullanılan bir ölçüttür [60]. WER, tanınan metindeki hatalı kelimelerin oranını ölçer. WER şu şekilde hesaplanır:

$$\text{Word Error Rate (WER)} = \frac{S + D + I}{N} \tag{4.1}$$

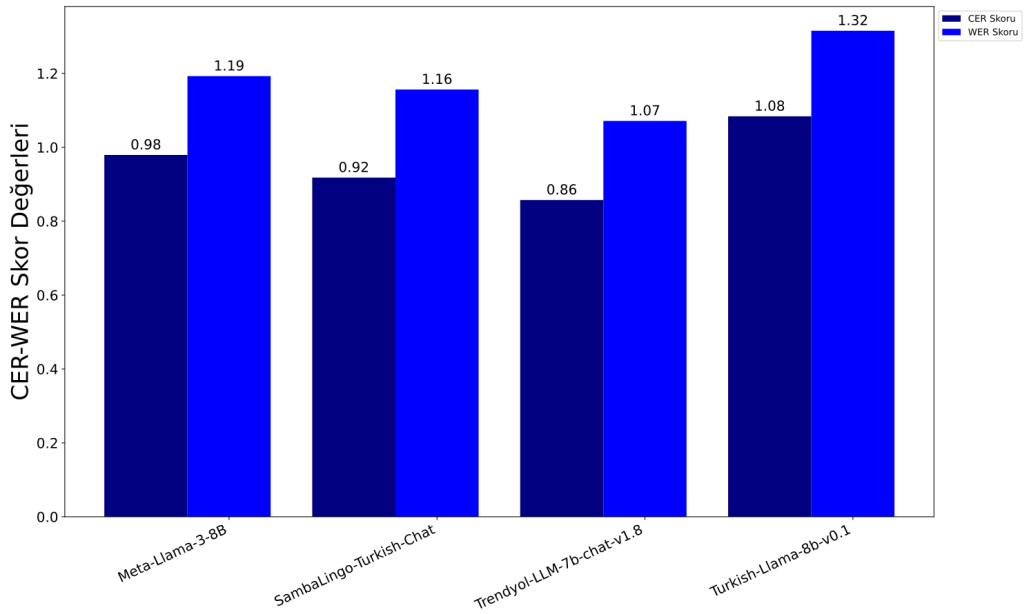
Burada:

- S: Değiştirilen kelime sayısı
- D: Silinen kelime sayısı
- I: Eklenen kelime sayısı
- N: Referans metindeki toplam kelime sayısı

WER, otomatik konuşma tanıma sistemlerinin performansını ölçmede de kullanılmaktadır [61]. Mesela, bir konuşma tanıma sistemi tarafından üretilen metin ile gerçek konuşma metni kıyaslanarak WER hesaplanabilir. Hannun ve arkadaşları, Deep Speech adlı çalışmalarında WER'i konuşma tanıma sistemlerinin performansını ölçmek için kullanmışlardır [62]. Benzer şekilde, Lopes ve Perdigao TIMIT veritabanı üzerinde yaptıkları telefon tanıma çalışmasında WER metriğinden faydalanmışlardır [63]. Ancak, Wang ve arkadaşları WER'in konuşma dilini anlama doğruluğu için her zaman iyi bir gösterge olmayabileceğini ortaya koymuşlardır [64]. Bu nedenle, WER'in diğer ölçütlerle beraber kullanılması, sistemlerin performansını daha kapsamlı bir şekilde ölçebilmek için, lazımdır.

WER sonuçlarına göre, modellerin başarımları şu şekildedir: **Trendyol-LLM-7b-chat-v1.8** modeli *1,0714* puan ile en iyi başarımı göstermiştir. İkinci sırada *1,1564* puan ile **SambaLingo-Turkish-Chat** modeli yer almaktadır. **Meta-Llama-3-8B**

modeli *1,1927* puan ile üçüncü sırada yer alırken, **Turkish-Llama-8b-v0.1** modeli *1,3155* puan ile son sırada yer almıştır.



Şekil 4.6 CER-WER Skorları

WER skorları ne kadar düşükse model o kadar başarılı demektir. Bu sonuçlara göre, **Trendyol-LLM-7b-chat-v1.8** modeli açık ara en iyi başarımı gösterirken, **Turkish-Llama-8b-v0.1** en başarısız model olmuştur.

4.2.5 CER

CER , WER'in karakter tabanlı bir versiyonudur diyebiliriz. CER, metin tanıma ve üretme sistemlerinin performansını karakter düzeyinde değerlendirmek için kullanılan bir ölçüttür. CER'in formülü şu şekildedir:

$$CER = \frac{S + D + I}{N} \tag{4.2}$$

Burada:

- S: Değiştirilen karakter sayısı (Substitutions)
- D: Silinen karakter sayısı (Deletions)
- I: Eklenen karakter sayısı (Insertions)
- N: Referans metindeki toplam karakter sayısı

CER, WER'e benzer şekilde, üretilen metin ile referans metin ikilisi arasındaki farkı ölçer, ama bunu kelime seviyesinde değil, karakter seviyesinde yapar. Bu özelliğinden dolayı CER, bilhassa Türkçe gibi morfolojik olarak zengin dillerde ya da kısa metinlerde daha doğru bir değerlendirme yapılmasına vesile olabilir. WER gibi, CER de 0'a yaklaşması, sistemin daha başarılı olduğunu gösterir. CER'in 1'den büyük olması da tıpkı WER'deki gibi mümkündür, mesela oluşturulan metin referans metinden çok daha uzun ve yanlışsa. CER ve WER'in beraber kullanılması, bir sistemin hem kelime hem de karakter düzeyindeki performansını ölçebilmek için lazımdır. Bu durum, bilhassa Türkçe gibi eklemeli dillerde, kelimelerin yapısını ve karakterlerin doğru kullanımını ayrıntılı bir biçimde inceleyebilmeye yardımcı olur.

Şekil 4.6'te gösterilen CER sonuçlarına göre, **Trendyol-LLM-7b-chat-v1.8** modeli *0,8573* puan ile en iyi başarımı göstermiştir. İkinci sırada *0,9180* puan ile **SambaLingo-Turkish-Chat** modeli yer almaktadır. **Meta-Llama-3-8B** modeli *0,9793* puan ile üçüncü sırada yer alırken, **Turkish-Llama-8b-v0.1** modeli *1,0839* puan ile sonuncu olmuştur.

Genel olarak **Trendyol-LLM-7b-chat-v1.8** modelinin diğer modellere göre gayet düşük CER skoruna sahip olması, bu modelin karakter düzeyinde daha doğru metin üretme işi yaptığını ve Türkçe metinleri daha düzgün anlayıp işlediğini göstermektedir. Lakin ikinci sırada yer alan **SambaLingo-Turkish-Chat** modeline de başarısız diyemeyiz. **Meta-Llama-3-8B** ve **Turkish-Llama-8b-v0.1** modellerinin daha yüksek hata oranları olması, bu modellerin Türkçe metin üretiminde diğer modellere göre daha kötü olduğunu göstermektedir. Aynı zamanda tüm modellerin WER skorlarının 1'in üzerinde olması, sonuçlarda hala ciddi boyutta hata olduğunu göstermektedir. Bu, Türkçe dil modellemesinde hala aşılması gereken zorluklar olduğunu göstermektedir.

4.2.6 METEOR

METEOR metin üretimi ve çeviri sistemlerinin başarımını değerlendirmek için kullanılan bir ölçüttür [65]. Bu ölçğt, oluşturulan metin ile referans metin arasındaki alakayı ölçer ve diğer ölçütlere göre nispeten daha kapsamlı bir değerlendirme yapar. METEOR, kelime eşleştirmesi yaparken yalnızca tam eşleşmeleri değil, bununla beraber kök, eşanlamlı ve yakın anlamlı eşleşmelere de bakar. Bu özellik, METEOR'un dil çeşitliliğini ve anlamsal benzerliği daha iyi ölçebilmesine vesile olur. METEOR skoru, hassasiyet (precision) ve duyarlılık (recall) değerlerinin harmonik ortalaması alınarak hesaplanır. Bu hesaplamada, eşleşen kelimelerin sıralaması da önemlidir. METEOR skoru 0 ila 1 arasında değer alır ve değer 1'e ne

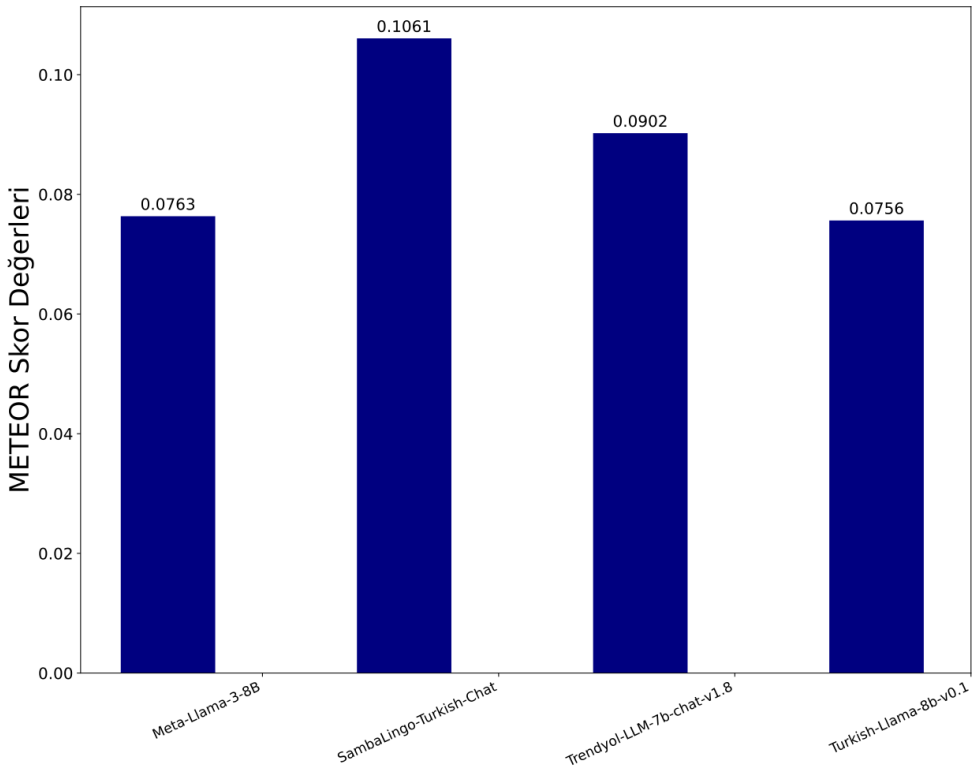
kadar yakınsa performans o kadar iyi demektir. METEOR'un formülü şu şekildedir:

$$\text{METEOR} = (1 - \text{Penalty}) \cdot \frac{10PR}{R + 9P} \tag{4.3}$$

Burada:

- P: Hassasiyet (Precision)
- R: Duyarlılık (Recall)
- Penalty: Ceza terimi (eşleşen kelimelerin sıralamasına bağlı)

METEOR, bilhassa makine çevirisi ve metin özetleme gibi alanlarda genel olarak kullanılmaktadır. Bununla beraber, Uludoğan ve arkadaşlarının çalışmasında da METEOR ölçütü, Türkçe dil modellerinin performansını değerlendirmek için kullanılmıştır [16]. Şekil 4.7'de görüldüğü üzere, **SambaLingo-Turkish-Chat**



Şekil 4.7 METEOR Skorları

modeli *0,1061* puan ile en iyi başarımı göstermiştir. İkinci sırada *0,0902* puan ile **Trendyol-LLM-7b-chat-v1.8** modeli yer almaktadır. **Meta-Llama-3-8B** modeli *0,0763* puan ile üçüncü sırada yer alırken, **Turkish-Llama-8b-v0.1** modeli *0,0756*

puan ile son sırada yer almıştır. SambaLingo-Turkish-Chat modelinin en yüksek METEOR skoruna sahip olarak oluşturduğu metinlerin referans metinlere en yakın model olduğunu göstermiştir. Bu, modelin Türkçe dil yapısını ve anlamsal alakaları daha iyi anladığının bir göstergesidir. Trendyol-LLM-7b-chat-v1.8 modelinin ikinci en iyi başarımı göstermesi, bu modelin de Türkçe metin üretiminde gayet başarılı olduğunu göstermektedir. Meta-Llama-3-8B ve Turkish-Llama-8b-v0.1 modellerinin daha düşük METEOR skorları alması, bu modellerin Türkçe metin oluşturmada diğer iki modelden daha başarısız olduğunu ve dil için daha fazla geliştirilmeleri gerektiğini göstermektedir. Bilhassa SambaLingo-Turkish-Chat ve Trendyol-LLM-7b-chat-v1.8 modellerinin diğer iki modele göre bariz bir şekilde daha yüksek puanlar alması, bu modellerin Türkçe dil özelliklerini daha iyi modelleyebildiklerini, daha doğal ve düzgün metinler oluşturabildiklerini göstermektedir.

4.3 Yapay Zeka Hakemliğinde Model Değerlendirmesi

Bu bölümde, ince ayar yapılan LLM'lerin performansını ölçmek amacıyla daha gelişmiş LLM'ler hakemliğinde iki farklı yöntem kullanılmıştır: Elo Puanlaması ve Kazanma Yüzdesi. Bu yöntemler, modellerin birbirleriyle kıyaslanması ve göreceli başarımlarının ölçülmesinde sentetik testlere kıyasla çok daha kapsamlı yöntemlerdir. Değerlendirme sürecinde, gelişmiş yapay zeka modelleri hakem olarak kullanılmıştır. Bu modeller *Claude 3.5 Sonnet, GPT-4, LLaMA 3.1 70B, Microsoft Copilot ve Gemini 1.5 Pro* modelleridir. Bu gelişmiş modellerin kullanılmasının sebebi, değerlendirme sürecinin daha objektif ve kapsamlı olmasıdır aynı zamanda bu modeller *LMSYS Chatbot Arena Leaderboard*'daki sıralamaları dikkate alınarak seçilmiştir [41]. Bu modeller, insan dilini anlama ve oluşturma alanında çok iyi oldukları için, diğer modellerin performansını ölçmede kullanılmıştır. Değerlendirme süreci şu şekilde gerçekleştirilmiştir:

- Test veri kümesinden 20 adet rastgele soru seçilmiştir.
- Bu seçilen 20 rastgele soru, tüm ince ayar yapılan modellere sorulmuştur.
- Her modelden alınan tüm cevaplar, gelişmiş yapay zeka hakemleri eşliğinde 2'şerli olarak karşılaştırılmıştır.
- Bu karşılaştırmalardaki her maçta, kazanma, kaybetme veya beraberlik sonuçları alınabilir.
- Karşılaştırma sonuçları kaydedilmiştir.

Bu değerlendirme yöntemi, modellerin gerçek dünyadaki başarımını daha kapsamlı ölçmeyi amaçlamaktadır. Geleneksel ölçütler yerine, gelişmiş yapay zeka modellerinin hakemliğinde yapılan karşılaştırmalar, modellerin cevaplarının kalitesini, doğruluğunu ve tamlığını daha kapsamlı bir şekilde değerlendirebilmektedir. Elo puanlaması ve kazanma yüzdesi, modellerin birbirlerine göre göreceli performanslarını ölçmek için kullandığımız ölçütlerdir. Bu ölçütler, modellerin sıralanmasını ve aralarındaki performans farklarının nicel olarak ifade edilmesine vesile olmuştur. Bu değerlendirme yönteminin kısıtlamaları da dikkate alınmalıdır. Mesela, hakem olarak kullanılan gelişmiş LLM'lerin eğilimleri veya çeşitli sınırları olabilir. Ayrıca, 20 sorudan oluşan örneklem büyüklüğünün, modellerin genel performansını tam olarak gösteremeyeceği de dikkate alınmalıdır. İlerideki çalışmalarda, daha kapsamlı bir soru kümesi kullanılması ve farklı uzmanlık alanlarından soruların dahil edilmesi, sonuçların güvenilirliğini artıracaktır.

4.3.1 Elo Puanlaması

Elo puanlama sistemi, ilk zamanlarda satranç oyuncularının göreli başarılarını değerlendirmek için bulunmuş bir yöntemdir [6]. Bu sistem, daha sonra birçok alana uyarlanmıştır. Elo puanlaması, modellerin birbirleriyle olan rekabetlerinden yola çıkarak göreceli bir performans ölçüsü sunar. Modellerin Elo skorları Şekil 4.8'de verilmiştir.

Elo puanlamasının formülü Denklem 4.4'de ve beklenen puanın formülü Denklem 4.5'de verilmiştir.

$$R'_A = R_A + K * (S_A - E_A) \tag{4.4}$$

$$E_A = \frac{1}{1 + 10^{(R_B - R_A)/400}} \tag{4.5}$$

Denklem 4.4'de:

- R'_A: A modelinin yeni Elo puanı
- R_A: A modelinin mevcut Elo puanı
- R_B: B modelinin (rakip) mevcut Elo puanı
- K: Maksimum puan değişimi (bu çalışmada 32 olarak belirlenmiştir)

- S_A: A modelinin maç sonucu (kazanma=1, beraberlik=0.5, kaybetme=0)
- E_A: A modelinin beklenen sonucu

Denklem 4.5'de:

- E_A: A modelinin beklenen skoru (0 ile 1 arasında bir değer)
- R_A: A modelinin mevcut Elo puanı
- R_B: B modelinin (rakip) mevcut Elo puanı
- 400: Elo sisteminde kullanılan sabit bir değer, puanlar arasındaki farkın etkisini ölçeklendirir
- 10: Elo sisteminde kullanılan sabit bir taban değeri

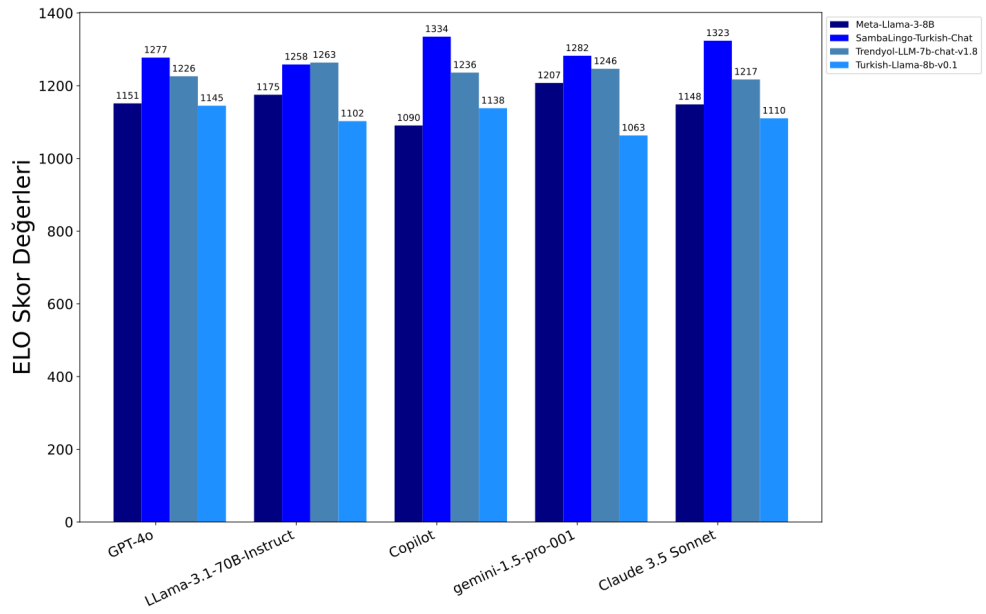
Denklem 4.5, iki modelin mevcut Elo puanlarını kullanarak, A modelinin B modeline karşı beklenen puanını bulur. Sonuç, A modelinin kazanma olasılığıdır. Mesela, $E_A = 0.75$ ise, A modelinin B modeline karşı tahmin edilen kazanma oranı $\%75$'dir.

Bu çalışmada, başlangıç elo puanı tüm modeller için *1200*'dir. Bu, tüm modellerin başlangıçta aynı başarıya sahip olduğu varsayımıyla başlamak demektir. K değeri *32* olarak seçilmiştir, bu da her karşılaştırmada puanların büyük oranda değişebileceği anlamına gelir.

4.3.1.1 GPT-4o Hakemliği

GPT-4o hakemliğinde yapılan değerlendirmelere göre, **SambaLingo-Turkish-Chat** modeli *1277.40* puan ile en yüksek Elo skorunu elde etmiştir. İkinci sırada *1226.03* puan ile **Trendyol-LLM-7b-chat-v1.8** modeli yer almaktadır. **Meta-Llama-3-8B** modeli *1151.48* puan ile üçüncü sırada, **Turkish-Llama-8b-v0.1** modeli ise *1145.09* puan ile son sırada yer almıştır.

SambaLingo-Turkish-Chat modeli, GPT-4o'nun değerlendirmesine göre en iyi modeldir. Trendyol-LLM-7b-chat-v1.8 modeli de görece ufak bir farkla ikinci sıradadır. Fakat Meta-Llama-3-8B ve Turkish-Llama-8b-v0.1 modelleri diğer iki modele göre düşük başarım göstermiştir.



Şekil 4.8 Elo Skorları

4.3.1.2 LLaMA 3.1 70B Hakemliği

LLaMA 3.1 70B hakemliğinde yapılan değerlendirmelere göre, **Trendyol-LLM-7b-chat-v1.8** modeli *1263.70* puan ile en yüksek Elo skorunu elde etmiştir. İkinci sırada *1258.43* puan ile **SambaLingo-Turkish-Chat** modeli yer almaktadır. **Meta-Llama-3-8B** modeli *1175.24* puan ile üçüncü sırada, **Turkish-Llama-8b-v0.1** modeli ise *1102.63* puan ile son sırada yer almıştır.

LLaMA 3.1 70B'nin değerlendirmesine göre, Trendyol-LLM-7b-chat-v1.8 ve SambaLingo-Turkish-Chat modelleri birbirlerine çok yakın başarıma sahiptir. Meta-Llama-3-8B ve Turkish-Llama-8b-v0.1 modelleri ise diğer yine diğer iki modele göre düşük başarım göstermiştir.

4.3.1.3 Microsoft Copilot Hakemliği

Microsoft Copilot hakemliğinde yapılan değerlendirmelere göre, **SambaLingo-Turkish-Chat** modeli *1334.97* puan ile en yüksek Elo skorunu elde etmiştir. İkinci sırada *1236.15* puan ile **Trendyol-LLM-7b-chat-v1.8** modeli yer almaktadır. **Turkish-Llama-8b-v0.1** modeli *1138.20* puan ile üçüncü sırada, **Meta-Llama-3-8B** modeli ise *1090.68* puan ile son sırada yer almıştır.

Microsoft Copilot'un değerlendirmesine göre, SambaLingo-Turkish-Chat modeli diğer modellere göre bariz şekilde daha yüksek başarım göstermiştir.

Trendyol-LLM-7b-chat-v1.8 modeli ikinci en iyi performansı sergilerken bu sefer *SambaLingo-Turkish-Chat* ile arasındaki fark açılmıştır. Turkish-Llama-8b-v0.1 ve Meta-Llama-3-8B modelleri ise yine diğer iki modele göre düşük başarım göstermiştir ama bu sefer Meta-Llama-3-8B modeli Turkish-Llama-8b-v0.1 modeli ile olumlu ölçüde farkı açmıştır.

4.3.1.4 Gemini 1.5 Pro Hakemliği

Gemini 1.5 Pro hakemliğinde yapılan değerlendirmelere göre, **SambaLingo-Turkish-Chat** modeli *1282.50* puan ile en yüksek Elo skorunu elde etmiştir. İkinci sırada *1246.64* puan ile **Trendyol-LLM-7b-chat-v1.8** modeli yer almaktadır. **Meta-Llama-3-8B** modeli *1207.63* puan ile üçüncü sırada, **Turkish-Llama-8b-v0.1** modeli ise *1063.23* puan ile son sırada yer almıştır.

Gemini 1.5 Pro'nun değerlendirmesine göre, SambaLingo-Turkish-Chat ve Trendyol-LLM-7b-chat-v1.8 modelleri yine en iyi performansı gösterirken, Meta-Llama-3-8B modeli orta düzeyde bir performans göstermiştir. Turkish-Llama-8b-v0.1 modeli ise diğer modellere göre bariz derecede daha düşük bir performansa sahiptir.

4.3.1.5 Claude 3.5 Sonnet Hakemliği

Claude 3.5 Sonnet hakemliğinde yapılan değerlendirmelere göre, **SambaLingo-Turkish-Chat** modeli *1323.79* puan ile en yüksek Elo skorunu elde etmiştir. İkinci sırada *1217.22* puan ile **Trendyol-LLM-7b-chat-v1.8** modeli yer almaktadır. **Meta-Llama-3-8B** modeli *1148.63* puan ile üçüncü sırada, **Turkish-Llama-8b-v0.1** modeli ise *1110.36* puan ile son sırada yer almıştır.

Claude 3.5 Sonnet'in değerlendirmesine göre, SambaLingo-Turkish-Chat modeli diğer modellere göre bariz şekilde daha iyi performans göstermiştir. Trendyol-LLM-7b-chat-v1.8 modeli ikinci en iyi performansa sahipken bu defa SambaLingo-Turkish-Chat modeli ile arasındaki fark çok açılmıştır. Meta-Llama-3-8B ve Turkish-Llama-8b-v0.1 modelleri daha düşük ve birbirlerine görece yakın performans göstermiştir.

4.3.1.6 Genel Değerlendirme

Tüm hakemlerin ölçümlerine bakıldığında, modellerin başarımları arasında tutarlı bir sıralama olduğu barizdir. Bu tutarlılık, kullanılan hakem modellerinin değerlendirmelerinin birbiriyle yüksek korelasyona sahip olduğunu gösterir.

SambaLingo-Turkish-Chat modeli, LLaMA 3.1 70B dışındaki tüm hakemlere göre en iyi başarımı göstermiştir. LLaMA 3.1 70B değerlendirmesinde ise önemsenmeyecek derecede küçük bir farkla ikinci sırada yer almıştır. Bu sonuç, SambaLingo-Turkish-Chat modelinin Türkçe sağlık sorularına cevap verme konusunda genel olarak en başarılı model olduğunu ortaya koymaktadır.

Trendyol-LLM-7b-chat-v1.8 modeli, LLaMA 3.1 70B dışındaki tüm hakemlerde ikinci sırada yer almıştır. LLaMA 3.1 70B değerlendirmesinde ise birinci sırada yer almıştır. Bu netice, Trendyol-LLM-7b-chat-v1.8 modelinin de tutarlı bir şekilde iyi derecede başarım gösterdiğini ortaya koymaktadır.

Meta-Llama-3-8B modeli, Microsoft Copilot haricinde tüm hakemlerde üçüncü sırada yer almıştır. Microsoft Copilot değerlendirmesinde ise son sırada yer almıştır. Meta-Llama-3-8B modeli bu neticeye göre genel olarak orta-düşük seviyede bir performansa sahiptir.

Turkish-Llama-8b-v0.1 modeli, Microsoft Copilot haricindeki tüm hakem modellere göre sonuncu olmuştur. Microsoft Copilot ölçümündeyse üçüncü sırada yer almıştır. Turkish-Llama-8b-v0.1 modeli bu neticelere göre genel manada diğer modellere göre daha düşük performansa sahiptir.

Hakem modellerin değerlendirmeleri arasındaki yüksek korelasyon, sonuçların güvenilirliğini artırmaktadır. Farklı mimarilere sahip ve farklı eğitim verileriyle eğitilmiş olan bu gelişmiş LLM'lerin benzer ölçümler yapması, elde edilen sıralamanın rastgele olmadığını göstermektedir. Bu da, modellerin performansını düzgün ölçebildiklerinin göstergesidir.

4.3.2 Kazanma Yüzdesi

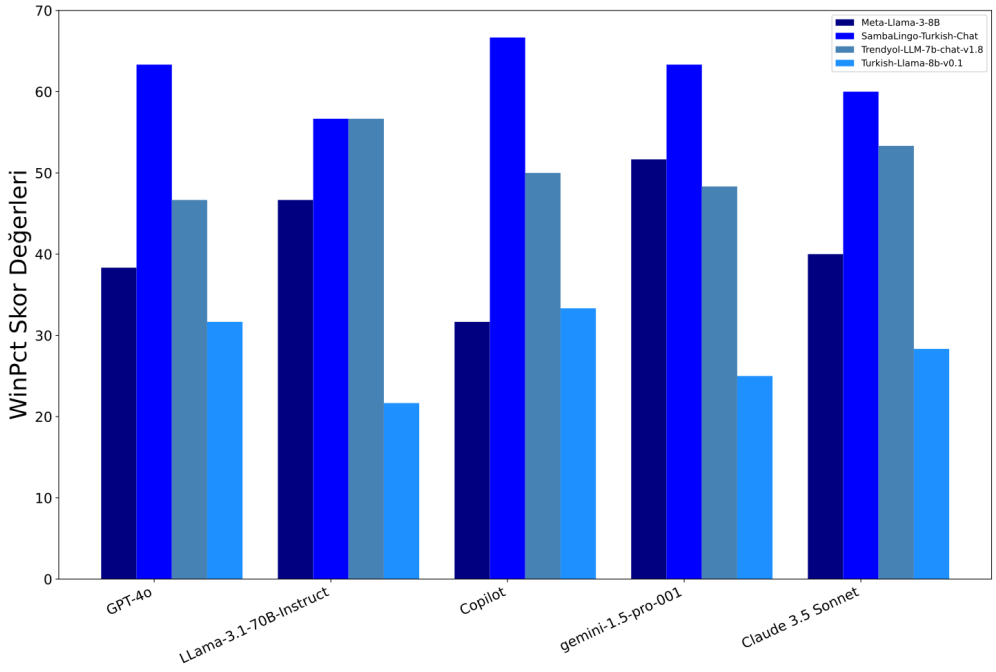
Kazanma yüzdesi, bir modelin diğer modellere karşı başarımını ölçen basit ve anlaşılır bir ölçüttür. Bu ölçüt, bir modelin tüm karşılaşmalarında kazandığı maçların yüzdesidir. Kazanma yüzdesi, modellerin genel başarımını kıyaslamak için kullanışlı bir ölçüttür.

Kazanma yüzdesi şu formülü Denklem 4.6'da gösterilmiştir.

$$\text{Kazanma Yüzdesi} = \frac{\text{Kazanılan Maç Sayısı}}{\text{Toplam Maç Sayısı}} \times 100 \tag{4.6}$$

Bu çalışmada, her bir model yukarıda analtıldığı gibi diğer modellerle karşılaştırılmış ve her karşılaşmada bir kazanan seçilmiştir. Beraberlik durumu

kazanma yüzdesine katılmamıştır. Toplam kazanma yüzdesi, tüm hakemlerin kıymetlendirmelerinin ortalaması alınarak bulunmuştur. Kazanma yüzdeleri Şekil 4.9'da gözükmektedir.



Şekil 4.9 Kazanma Yüzdesi Skorları

4.3.2.1 GPT-4o Hakemliği

GPT-4o hakemliğinde yapılan değerlendirmelere göre, **SambaLingo-Turkish-Chat** modeli *%66,67* ile en yüksek kazanma yüzdesine sahiptir. İkinci sırada *%46,67* ile **Trendyol-LLM-7b-chat-v1.8** modeli yer almaktadır. **Meta-Llama-3-8B** modeli *%38,33* ile üçüncü sırada, **Turkish-Llama-8b-v0.1** modeli ise *%31,67* ile son sırada yer almıştır. GPT-4o'nun değerlendirmesine göre, SambaLingo-Turkish-Chat modeli diğer modellere göre açık ara en iyi modeldir. Trendyol-LLM-7b-chat-v1.8 modeli de ikinci olmuştur, lakin lider ile arasında bariz bir fark vardır. Meta-Llama-3-8B ve Turkish-Llama-8b-v0.1 modelleri ise birbirlerine yakın ancak diğer iki modele ırak başarıma sahiplerdir.

4.3.2.2 LLaMA 3.1 70B Hakemliği

LLaMA 3.1 70B hakemliğinde yapılan değerlendirmelere göre, **Trendyol-LLM-7b-chat-v1.8** ve **SambaLingo-Turkish-Chat** modelleri *%56,67* ile aynı en yüksek kazanma oranına sahiptirler. **Meta-Llama-3-8B** modeli *%46,67* ile üçüncü sırada, **Turkish-Llama-8b-v0.1** modeli ise *%21,67* ile son sıradadır. LLaMA 3.1 70B'nin

hakemliğine göre, Trendyol-LLM-7b-chat-v1.8 ve SambaLingo-Turkish-Chat modelleri aynı ve en yüksek başarıma sahiptir. Meta-Llama-3-8B modeli bu iki modele göre biraz daha düşük bir performansa sahipken, Turkish-Llama-8b-v0.1 modeli diğer modellere göre bariz derecede düşük performans göstermiştir.

4.3.2.3 Microsoft Copilot Hakemliği

Microsoft Copilot hakemliğinde yapılan değerlendirmelere göre, **SambaLingo-Turkish-Chat** modeli *%66,67* ile yine en yüksek kazanma yüzdesine sahiptir. İkinci sırada *%46,67* ile **Trendyol-LLM-7b-chat-v1.8** modeli yer vardır. **Turkish-Llama-8b-v0.1** modeli *%33,33* ile üçüncü sırada yer alırken, **Meta-Llama-3-8B** modeli ise *%31,67* ile sonuncu olmuştur. Microsoft Copilot'un hakemliğine göre, SambaLingo-Turkish-Chat modeli diğer modellere göre gözle görülür derecede yüksek bir başarıma sahiptir. Trendyol-LLM-7b-chat-v1.8 modeli ikinci en iyi başarımı göstermiştir. Lakin lider ile arasında azımsanmayacak derecede yüksek bir fark vardır. Turkish-Llama-8b-v0.1 ve Meta-Llama-3-8B modelleri ise birbirlerine çok yakın ve düşük performansa sahiplerdir.

4.3.2.4 Gemini 1.5 Pro Hakemliği

Gemini 1.5 Pro hakemliğinde yapılan değerlendirmelere göre, **SambaLingo-Turkish-Chat** modeli *%63,33* ile en yüksek kazanma oranına sahiptir. İkinci sırada *%51,67* ile **Meta-Llama-3-8B** modeli yer almaktadır. **Trendyol-LLM-7b-chat-v1.8** modeli *%48,33* ile üçüncü sırada, **Turkish-Llama-8b-v0.1** modeli ise *%25* ile son sırada yer almıştır. Gemini 1.5 Pro'nun değerlendirmesine göre, SambaLingo-Turkish-Chat modeli diğer modellere göre belirgin bir başarıya sahiptir. Garip bir şekilde, Meta-Llama-3-8B modeli en iyi ikinci model olmuş ve Trendyol-LLM-7b-chat-v1.8 modelini geçmiştir. Turkish-Llama-8b-v0.1 modeli ise diğer modellere göre gayet düşük bir performans sergilemiştir.

4.3.2.5 Claude 3.5 Sonnet Hakemliği

Claude 3.5 Sonnet hakemliğinde yapılan değerlendirmelere göre, **SambaLingo-Turkish-Chat** modeli *%60,00* ile en yüksek kazanma yüzdesinine sahiptir. İkinci sırada çok da düşük olmayan bir yüzde olan *%53,33*'le **Trendyol-LLM-7b-chat-v1.8** modeli yer vardır. **Meta-Llama-3-8B** modeli *%40,00* ile üçüncü sırada, **Turkish-Llama-8b-v0.1** modeli ise *%28,33* ile son sırada yer almıştır. Claude 3.5 Sonnet'in değerlendirmesine göre, SambaLingo-Turkish-Chat modeli yine en iyi başarıma sahiptir, lakin Trendyol-LLM-7b-chat-v1.8 modeli ile arasındaki fark diğer hakemlerin değerlendirmelerine göre daha düşüktür. Meta-Llama-3-8B

ve Turkish-Llama-8b-v0.1 modelleri daha düşük başarım göstermiştir. Bilhassa Turkish-Llama-8b-v0.1 modeli genel olarak en düşük performansa sahiptir.

4.3.2.6 Genel Değerlendirme

Tüm LLM hakemlerin değerlendirmelerine göre, kazanma yüzdesi ölçütüne göre de modellerin performansları arasında tutarlı bir sıralama vardır. Bu tutarlılık, Elo puanlamasında olduğu gibi, kullanılan hakem modellerinin değerlendirmelerinin birbiriyle yüksek korelasyon içinde olduğunu ortaya koymaktadır.

LLaMA 3.1 70B dışındaki tüm hakemlere göre en başarılı model SambaLingo-Turkish-Chat modelidir. Model, LLaMA 3.1 70B değerlendirmesinde ise çok ufak bir farkla ikinci sırada yer almıştır. Bu sonuç, SambaLingo-Turkish-Chat modelinin Türkçe sağlık sorularına cevap verme konusunda en başarılı model olduğunu göstermektedir.

Trendyol-LLM-7b-chat-v1.8 modeliyse, LLaMA 3.1 70B değerlendirmesi hariç tüm hakem değerlendirmelerinde ikinci sıradadır. LLaMA 3.1 70B değerlendirmesinde ise birinci sırada yer almıştır. Bu durum, Trendyol-LLM-7b-chat-v1.8 modelinin de tutarlı bir biçimde iyi bir performans gösterdiğine işaret etmektedir.

Meta-Llama-3-8B modeliyse, Microsoft Copilot ve Gemini 1.5 Pro hakemliği dışında hep üçüncü sırada yer almıştır. Microsoft Copilot değerlendirmesinde son sırada yer alırken Gemini 1.5 Pro hakemliğinde azımsanmayacak bir farkla ikinci sırada yer almıştır. Bu sonuç, Meta-Llama-3-8B modelinin genel olarak orta-düşük düzeyde bir performans sergilediğini göstermektedir.

Turkish-Llama-8b-v0.1 modeliyse, Microsoft Copilot hakemliği dışındaki tüm hakemlerde son sırada yer almıştır. Microsoft Copilot hakemliğindeyse sonuncuyla arasında çok az bir fark vardır. Bu sonuç, Turkish-Llama-8b-v0.1 modelinin genel olarak diğer modellere göre bariz derecede düşük başarım gösterdiğine delalet etmektedir.

Kazanma yüzdesi metriği sonuçları, Elo puanlaması ile büyük oranda uyumludur. Bu durum, iki farklı değerlendirme yönteminin de modellerin performanslarını tutarlı bir şekilde ölçtüğüne işaret etmektedir. Hakem modellerin değerlendirmeleri arasındaki yüksek korelasyon, neticeler daha itimata değer yapmaktadır.

Bu sonuçlara göre en iyi başarıma sahip modeller SambaLingo-Turkish-Chat ve Trendyol-LLM-7b-chat-v1.8 modelleridir. Bu da modellerin, Türkçe dil

yapısına ve sağlık terminolojisine daha hakim olduklarını gösterir. Diğer taraftan, Meta-Llama-3-8B ve Turkish-Llama-8b-v0.1 modellerinin daha düşük performansa sahip olması, bu modellerin Türkçe sağlık alanında daha çok iyileştirmeye gerek duyduğunu gösterir.

4.4 Uzman Değerlendirmesi

Uzman değerlendirmesi, sentetik testlerden daha etkili ve şu an için gerekli bir yöntemdir. Zira bu alana özel olarak başarısı kanıtlanmış bir sentetik test henüz geliştirilmemiştir. Ama uzman değerlendirmesi de maliyetli ve yavaştır. Bu yöntem, modellerin gerçek hayattaki başarımlarını daha iyi ölçmektedir. Bilhassa sağlık gibi hassas alanlarda modellerin güvenilirliğini değerlendirmede çok önemlidir. Uzman değerlendirmesi süreci şu şekilde gerçekleştirilmiştir:

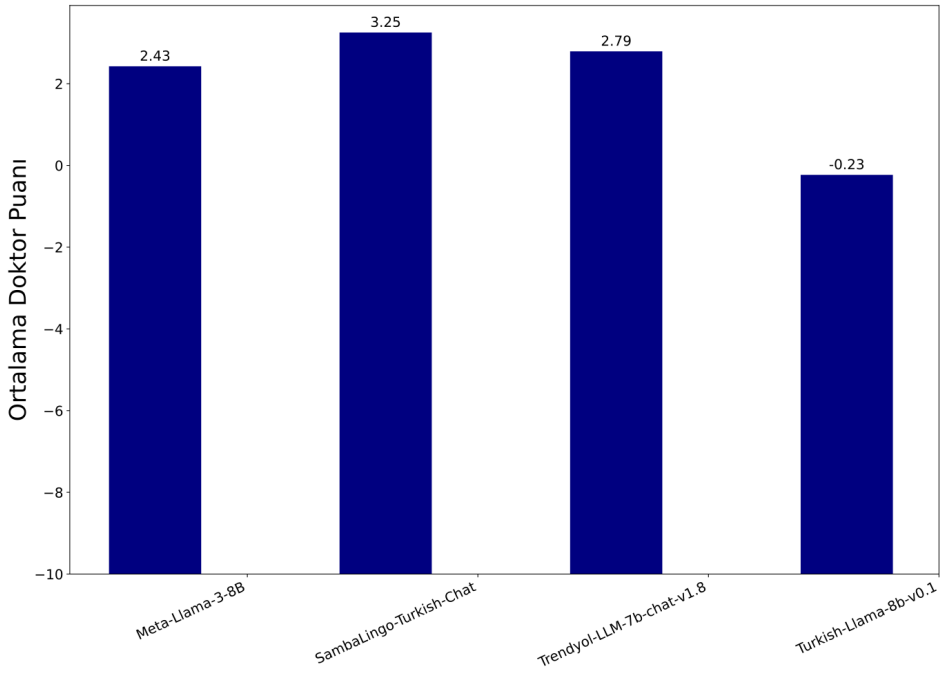
1. 20 adet hasta sorusu rastgele seçilmiştir. Bu sorular, modellerin LLM'ler tarafından değerlendirildiği öbür 20 sorudan farklıdır.
2. Önce modellerin bu sorulara verdiği cevaplar alınmıştır.
3. Toplam 80 (20 x 4) model cevabı elde edilmiştir.
4. Çeşitli uzmanlık alanlarından 20 doktor, bu cevapları ölçmüştür.
5. Uzman doktorlar, modellerin verdiği cevapların kalitesini -10 ile +10 arasında puanlamıştır:
    - -10'a yaklaşan puanlar, modelin cevabının zararlı ve hastayı yanlış yönlendirdiğini göstermektedir.
    - +10'a yaklaşan puanlar, modelin cevabının yararlı ve soruyu eksiksiz bir şekilde cevapladığını göstermektedir.
    - 0'a yaklaşan puanlar, modelin cevabının ne yararlı ne zararlı, yani faydasız olduğunu göstermektedir.

Bu ölçüm süreci, sentetik testlerden farklı olarak, modellerin gerçek dünyadaki performansını ölçmeyi hedeflemektedir. Modellerin sağlık alanındaki potansiyeli daha iyi ölçebilmek için bu ölçüt çok önemlidir.

Bu değerlendirme sürecinde, doktorlar toplam olarak *1536* model cevabını puanlamıştır. Puanlama yapan doktorlar on iki muhtelif kurumda görev almaktadır. Toplamda yirmi farklı doktor model cevaplarını puanlamıştır. Bu doktorların ünvan dağılımıysa;

- Stajyer Doktor (6 Adet)
- İntern Doktor (4 adet)
- Asistan Doktor (4 Adet)
- Pratisyen Hekim (2 Adet)
- Uzman Doktor (4 Adet)

Şekil 4.10'e bakıldığında, uzmanlara göre en iyi model *3,25* puanla **SambaLingo-Turkish-Chat** olmuştur. İkinci sırada *2,79* puanla **Trendyol-LLM-7b-chat-v1.8** modeli yer almaktadır. Trendyol-LLM-7b-chat-v1.8'in ardından *2,43* puanla **Meta-Llama-3-8B** gelmektedir. İlk üç model arasında bariz farklar olsa da bu modellerin sonuçta faydalı modeller olduğu açıktır. Lakin *-0,23* puan alan **Turkish-Llama-8b-v0.1**, diğer modellere göre gayet az puan almıştır. Hem de değerlendiren uzmanlar, zararı faydasından daha olacak şekilde oylamışlardır. Tablo 4.2'de gösterilen



Şekil 4.10 Uzman Değerlendirme Skorları

cevap yarar dağılımına bakıldığında, en yüksek yararlı cevap oranına sahip model **SambaLingo-Turkish-Chat** (%71,1) olmuştur. Bunu **Trendyol-LLM-7b-chat-v1.8** (%69,4) ve **Meta-Llama-3-8B** (%62,6) takip etmektedir. **Turkish-Llama-8b-v0.1** modeli ise en düşük yararlı cevap oranına (%42,1) ve en yüksek zararlı cevap oranına (%33,8) sahiptir.

Tablo 4.2 Cevap Yarar Dağılımı

Model	**Yararlı (%)**	**Zararlı (%)**	**Nötr (%)**
Trendyol-LLM-7b-chat-v1.8	69,4	18,4	12,2
Meta-Llama-3-8B	62,6	21,3	16,1
Turkish-Llama-8b-v0.1	42,1	33,8	24,1
SambaLingo-Turkish-Chat	71,1	21,1	7,8

Sağlık alanındaki uygulamalar için, modellerin yararlı cevaplar üretmesi kadar zararlı cevaplar oluşturmaması da kimi zaman hayati derecede önemlidir. Bu bağlamda, **Trendyol-LLM-7b-chat-v1.8** modeli en düşük zararlı cevap oranına (%18,4) sahip olması açısından göze çarpmaktadır. **SambaLingo-Turkish-Chat** modeli ise en yüksek yararlı cevap oranına sahip olmasına rağmen, zararlı cevap oranı (%21,1) azımsanmayacak derecede fazladır.

Sonuç itibariyle, uzman değerlendirmeleri, modellerin sağlık alanındaki potansiyel kullanımları için gayet önemli çıkarımlar yapmamıza vesile olmuştur. **SambaLingo-Turkish-Chat** ve **Trendyol-LLM-7b-chat-v1.8** modelleri, sağlık alanında kullanım için en uygun modellerdir denilebilir. Ancak, sağlık gibi hayati sıkıntılar doğurabilecek bir alanda kullanılacak modellerin seçiminde, faydalı cevapların oranı kadar zararlı cevapların oranının da önemsenmesi lazımdır. Bundan dolayı, **Trendyol-LLM-7b-chat-v1.8** modelinin daha düşük zararlı cevap oranı nedeniyle kullanımı daha güvenlidir denilebilir.

5
BULGULAR

5.1 Model Performanslarının Karşılaştırılması

Bu bölümde, ince ayar yaptığımız dört LLM'in performansları çeşitli açılardan karşılaştırılacaktır.

5.1.1 Sentetik Değerlendirme Ölçütlerine Göre Karşılaştırma

Sentetik değerlendirme ölçütleri, modellerin performansını tarafsız ve nicel olarak ölçmek için kullanılan metriklerdir. Bu çalışmada kullanılan sentetik ölçütler BLEU, ROUGE, METEOR, BERT Skor, CER ve WER ve METEOR'dur.

5.1.1.1 BLEU Skorları

Modellerin BLEU skorları Şekil 4.4'de görüldüğü gibi şu şekildedir:

- SambaLingo-Turkish-Chat: 0,0248
- Trendyol-LLM-7b-chat-v1.8: 0,0125
- Turkish-Llama-8b-v0.1: 0,0073
- Meta-Llama-3-8B: 0,0056

Bu sonuçlara göre, *SambaLingo-Turkish-Chat* modeli en yüksek BLEU skorunu elde ederek, referans metinlere en yakın çıktıları üretmiştir. *Trendyol-LLM-7b-chat-v1.8* modeli ikinci en iyi performansı gösterirken, *Meta-Llama-3-8B* modeli en düşük puanı almıştır.

SambaLingo-Turkish-Chat modelinin *0,0248* BLEU skoru, diğer modellere göre bariz derecede iyidir. *Trendyol-LLM-7b-chat-v1.8* modelinin *0,0125* BLEU skoru ile ikinci sırada yer alması, bu modelin de Türkçe dil yapısına ve kullanımına göre

düzgün bir şekilde eğitildiğini gösteriyor. *Turkish-Llama-8b-v0.1* ve *Meta-Llama-3-8B* modellerinin görece düşük BLEU skorları (sırasıyla 0.0073 ve 0.0056), bu modellerin Türkçe metin üretiminde daha fazla iyileştirmeye ihtiyaç duyduğunu gözler önüne sermektedir. Bilhassa Meta-Llama-3-8B modelinin en düşük skoru alması, bu modelin genel dil anlama ve üretme yeteneklerinin yeterince güçlü olmasına rağmen, Türkçe'ye özgü ince ayara ihtiyacı olduğunu göstermektedir.

5.1.1.2 ROUGE Skorları

ROUGE ölçütleri olarak, ROUGE-1, ROUGE-2, ROUGE-L ve ROUGE-Lsum olmak üzere dört farklı ROUGE ölçütü seçilmiştir.

Şekil 4.3'de görüldüğü gibi tüm ROUGE metriklerinde *SambaLingo-Turkish-Chat* modeli en iyi sonuçlara sahiptir. Bu, modelin hem kelime hem de daha kelime öbekleri bazında referans metinlere en benzer çıktıları oluşturduğunu göstermektedir. *Trendyol-LLM-7b-chat-v1.8* modeli ikinci en iyi performansı gösterirken, *Meta-Llama-3-8B* ve *Turkish-Llama-8b-v0.1* modelleri daha düşük puanlar almıştır. Modellerin ROUGE skorları şu şekildedir:

- *SambaLingo-Turkish-Chat*: ROUGE-1: *0,1798*, ROUGE-2: *0,0493*, ROUGE-L: *0,1192*, ROUGE-Lsum: *0,1194*
- *Trendyol-LLM-7b-chat-v1.8*: ROUGE-1: *0,1580*, ROUGE-2: *0,0362*, ROUGE-L: *0,1046*, ROUGE-Lsum: *0,1048*
- *Turkish-Llama-8b-v0.1*: ROUGE-1: *0,1382*, ROUGE-2: *0,0230*, ROUGE-L: *0,0865*, ROUGE-Lsum: *0,0868*
- *Meta-Llama-3-8B*: ROUGE-1: *0,1339*, ROUGE-2: *0,0214*, ROUGE-L: *0,0852*, ROUGE-Lsum: *0,0854*

SambaLingo-Turkish-Chat modelinin tüm ROUGE metriklerinde en yüksek puanları elde etmesi, bu modelin Türkçe metin oluşturma işinde diğer modellere göre daha başarılı olduğunu göstermektedir. Bilhassa ROUGE-1 ve ROUGE-L skorlarındaki yüksek değerler, modelin hem kelime bazında hem de daha uzun metin parçaları bazında referans metinlere daha yakın çıktılar oluşturduğuna işaret etmektedir.

Trendyol-LLM-7b-chat-v1.8 modelinin ikinci en iyi performansı göstermesi, bu modelin de Türkçe dil yapısına ve kullanımına uygun şekilde eğitildiğine işaret etmektedir. *Turkish-Llama-8b-v0.1* ve *Meta-Llama-3-8B* modellerinin görece

daha düşük ROUGE skorları, bu modellerin Türkçe metin üretiminde daha fazla iyileştirilmesi gerektiğine işaret etmektedir.

ROUGE-2 skorlarının genel olarak düşük seviyede olması, tüm modellerin iki kelimelik öbekleri oluşturmada zorlandığını göstermektedir. Bu durum, Türkçe'nin karmaşık dil yapısı ve zengin morfolojisi ile de ilişkilidir.

5.1.1.3 METEOR Skorları

Şekil 4.7'de görüldüğü gibi modellerin METEOR skorları şu şekildedir:

- *SambaLingo-Turkish-Chat*: *0,1061*
- *Trendyol-LLM-7b-chat-v1.8*: *0,0902*
- *Meta-Llama-3-8B*: *0,0763*
- *Turkish-Llama-8b-v0.1*: *0,0756*

METEOR skorlarına göre de *SambaLingo-Turkish-Chat* modeli bariz şekilde en başarılı modeldir. Bu durum, modelin oluşturduğu metinlerin referans metinlere anlam ve yapısal açıdan en yakın model olduğunu göstermektedir.

Trendyol-LLM-7b-chat-v1.8 modeli *0,0902* METEOR skoru ile ikinci modeldir. Bu, modelin Türkçe dil yapısına uygun şekilde ince ayar yapıldığını bundan dolayı anlamlı çıktılar oluşturabildiğini göstermektedir.

Meta-Llama-3-8B ve *Turkish-Llama-8b-v0.1* modelleri sırasıyla *0,0763* ve *0,0756* METEOR skorları ile son iki sırada yer almaktadır. Bu iki modelin skorlarının birbirine çok yakın olması dikkat çekicidir. *Meta-Llama-3-8B* modelinin, genel dil modeli olmasına rağmen, Türkçe'ye özel eğitilmiş *Turkish-Llama-8b-v0.1* modelinin eğitim verisinin sağlık konusunda yetersiz kaldığını söyleyebiliriz.

METEOR skorlarındaki bu sıralama, diğer metriklerle (BLEU, ROUGE) tutarlılık göstermektedir. Bu tutarlılık, modellerin performans değerlendirmesinin güvenilir olduğuna işaret etmektedir.

5.1.1.4 BERT Skor

Şekil 4.5'te görülen BERT Skor sonuçlarına göre, *SambaLingo-Turkish-Chat* ve *Trendyol-LLM-7b-chat-v1.8* modelleri birbirine çok yakın başarımlar göstermiştir.

SambaLingo-Turkish-Chat modeli Recall ve F1 skorlarında, *Trendyol-LLM-7b-chat-v1.8* modeliyse Precision skorunda en yüksek puanı elde etmiştir.

Modellerin BERT Skor sonuçları şu şekildedir:

- *SambaLingo-Turkish-Chat*: Precision: *0,4960*, Recall: *0,4975*, F1: *0,4939*
- *Trendyol-LLM-7b-chat-v1.8*: Precision: *0,5155*, Recall: *0,4739*, F1: *0,4909*
- *Meta-Llama-3-8B*: Precision: *0,4754*, Recall: *0,4517*, F1: *0,4602*
- *Turkish-Llama-8b-v0.1*: Precision: *0,4591*, Recall: *0,4458*, F1: *0,4485*

SambaLingo-Turkish-Chat modelinin Recall (*0,4975*) ve F1 (*0,4939*) skorlarında en yüksek değerleri elde etmesi, bu modelin ürettiği metinlerin referans metinlerle daha fazla örtüştüğünü ve genel olarak daha dengeli bir performansa sahip olduğunu göstermektedir. Yüksek Recall skoru, modelin referans metinlerdeki mühim bilgileri yakalama hususunda başarılı olduğuna işaret eder.

Trendyol-LLM-7b-chat-v1.8 modelinin Precision skorunda (*0,5155*) en yüksek değeri elde etmesi, bu modelin ürettiği metinlerin daha kesin ve doğru olduğunu gösterir. Yüksek Precision skoru, modelin ürettiği içeriğin güvenilir olduğunu gösterir.

Meta-Llama-3-8B ve *Turkish-Llama-8b-v0.1* modelleri, BERT Skor ölçütünde de diğer iki modele göre daha başarısız olmuştur. Bu sonuç, bu modellerin Türkçe metin üretiminde daha fazla iyileştirilmesi gerektiğini bir kez daha ortaya koymuştur.

BERT Skor sonuçları, diğer ölçütlerle (BLEU, ROUGE, METEOR) genel olarak tutarlılık göstermektedir. Bu tutarlılık, ölçüm sonuçlarını daha güvenilir kılmaktadır.

Bu sonuçlar, dile özgü ve alana özgü ince ayar yapılmasının, LLM'lerin performansını ciddi miktarda artırabileceğini bir kez daha göstermiştir. *SambaLingo-Turkish-Chat* ve *Trendyol-LLM-7b-chat-v1.8* modellerinin başarısı, Türkçe gibi spesifik dillerde iyi derecede başarım elde etmek için, o dile özgü geniş veri kümeleri kullanmanın önemini ortaya göstermektedir.

5.1.1.5 CER ve WER Skorları

Şekil 4.6'da görülen CER ve WER skorlarına göre *Trendyol-LLM-7b-chat-v1.8 modeli* en iyi başarıma sahiptir. Bu, modelin karakter ve kelime düzeyinde en

az hata olasılığına sahip olduğunu göstermektedir. Bununla beraber *SambaLingo-Turkish-Chat*'in başarımı diğer ölçütlerin aksine daha düşük kalmıştır.

Modellerin CER ve WER skorları şu şekildedir:

- *Trendyol-LLM-7b-chat-v1.8*: CER: *0,8573*, WER: *1,0714*
- *SambaLingo-Turkish-Chat*: CER: *0,9180*, WER: *1,1564*
- *Meta-Llama-3-8B*: CER: *0,9793*, WER: *1,1927*
- *Turkish-Llama-8b-v0.1*: CER: *1,0839*, WER: *1,3155*

Trendyol-LLM-7b-chat-v1.8 modelinin hem CER hem de WER skorlarında en düşük değerlere sahip olması, bu modelin karakter ve kelime düzeyinde en doğru sonuçları oluşturduğunu göstermektedir.

SambaLingo-Turkish-Chat modelinin diğer ölçütlerde gösterdiği yüksek başarıma rağmen, CER ve WER skorlarında ikinci sırada yer alması dikkat çekici bir husustur. Bu durum, modelin genel anlamda daha iyi performans göstermesine rağmen, karakter ve kelime bazında daha başarısız olduğunu göstermektedir.

Meta-Llama-3-8B ve *Turkish-Llama-8b-v0.1* modelleri, CER ve WER skorlarında da diğer iki modele göre daha düşük seviyede performans göstermiştir. Bilhassa *Turkish-Llama-8b-v0.1* modelinin en yüksek hata olasılığı elde etmesi, bu modelin Türkçe sağlık metin üretiminde daha fazla iyileştirilmesi gerektiğini göstermektedir.

CER ve WER skorlarındaki bu sıralama, diğer metriklerle (BLEU, ROUGE, METEOR, BERT Score) kısmen farklılık göstermektedir. Bu fark, her ölçütün metin kalitesini farklı açılardan değerlendirdiğini ve modellerin farklı alanlarda iyi ve kötü yönleri olabileceğini göstermektedir.

5.1.2 Yapay Zeka Hakemliğinde Değerlendirme Sonuçları

Bu değerlendirmede Elo puanlaması ve kazanma yüzdesi metriklerinden yararlanılmıştır. Yapay zeka hakemliğinde değerlendirme, modellerin performansını daha kapsamlı ve bağlamsal bir şekilde ölçmeyi amaçlamaktadır. Bu değerlendirmede, gelişmiş yapay zeka modelleri (GPT-4o, LLaMA 3.1 70B, Microsoft Copilot, Gemini 1.5 Pro ve Claude 3.5 Sonnet) hakem olarak kullanılmıştır.

5.1.2.1 Elo Puanlaması

Elo puanlaması, satranç oyuncularının performansını değerlendirmek için geliştirilen ve daha sonra diğer rekabetçi alanlara da uyarlanan bir derecelendirme sistemidir. Bu sistemde, yüksek puanlar daha iyi performansı gösterir.

Şekil 4.8'deki Elo puanlarının ortalamaları şu şekildedir:

- *SambaLingo-Turkish-Chat*: *1295,41*
- *Trendyol-LLM-7b-chat-v1.8*: *1237,95*
- *Meta-Llama-3-8B*: *1154,73*
- *Turkish-Llama-8b-v0.1*: *1111,90*

Bu sonuçlara göre, *SambaLingo-Turkish-Chat* modeli en yüksek Elo puanını elde ederek, diğer modellere göre en iyi performansa sahiptir. *Trendyol-LLM-7b-chat-v1.8* modeli ikinci en iyi performansa sahip modelken, *Turkish-Llama-8b-v0.1* modeli en düşük puanı alan modeldir.

SambaLingo-Turkish-Chat modelinin *1295,41* puanla diğer modellere göre önemli bir farkla önde olması, bu modelin Türkçe sağlık verileri üzerinde daha tutarlı ve kaliteli sonuçlar oluşturduğunu göstermektedir.

Trendyol-LLM-7b-chat-v1.8 modelinin *1237,95* puanla ikinci sırada yer alması, bu modelin de Türkçe metin üretiminde diğer modellere göre yadsınamaz bir başarısı olduğunu ortaya koymaktadır.

Meta-Llama-3-8B modelinin *1154,73* puanla üçüncü sırada yer alması, genel amaçlı bir model olmasına rağmen Türkçe sağlık verilerinde makul bir performans gösterdiğini ortaya koymaktadır. Bu durum, modelin genel dil anlama ve üretme yeteneklerinin kabul edilebilir seviyede olduğunu, lakin Türkçe'ye özgü ince ayarlamalara ihtiyaç duyduğunu göstermektedir.

Turkish-Llama-8b-v0.1 modelinin *1111,90* puanla son sırada yer alması, bu modelin Türkçe sağlık verileri üzerinde çok daha fazla geliştirilmesi gerektiğini göstermektedir. Bu sonuç, modelin eğitim sürecinde kullanılan veri kümesinin ya da ince ayar parametrelerinin sağlık verileri için optimize edilmesi gerektiğine işaret etmektedir.

5.1.2.2 Kazanma Yüzdesi

Kazanma yüzdesi, modellerin diğer modellere karşı kazandığı karşılaştırmaların oranını gösterir. Bu ölçüt, modellerin birbirlerine göre göreceli performansını daha basit ve anlaşılır bir biçimde göstermektedir.

Şekil 4.9'da görülen modellerin ortalama kazanma yüzdeleri şu şekildedir:

- *SambaLingo-Turkish-Chat*: *%62,00*
- *Trendyol-LLM-7b-chat-v1.8*: *%51,00*
- *Meta-Llama-3-8B*: *%41,67*
- *Turkish-Llama-8b-v0.1*: *%28,00*

Kazanma yüzdesi sonuçları da Elo puanlamasıyla tutarlılık göstermektedir. *SambaLingo-Turkish-Chat* modeli *%62,00* kazanma yüzdesiyle en yüksek kazanma yüzdesine sahipken, *Turkish-Llama-8b-v0.1* modeli *%28,00* kazanma yüzdesiyle en düşük oranı elde etmiştir.

SambaLingo-Turkish-Chat modelinin *%62,00* kazanma oranı, bu modelin diğer modellere karşı kıyaslamaların yarısından fazlasını kazandığını gösterir. Bu sonuca göre model, rakiplerine göre Türkçe sağlık verileri üzerinde daha tutarlı ve kaliteli çıktılar üretmektedir.

Trendyol-LLM-7b-chat-v1.8 modelinin *%51,00* kazanma oranı, bu modelin karşılaştırmaların neredeyse yarısı kazandığını göstermektedir. Bu sonuç, modelin Türkçe metin üretiminde genel olarak orta seviyede olduğunu ve *SambaLingo-Turkish-Chat* modeline göre düşük bir performans sergilediğini göstermektedir.

Meta-Llama-3-8B modelinin *%41,67* kazanma oranı, modelin genel amaçlı bir model olmasına rağmen, Türkçe sağlık verilerinde çok da makul olmayan bir performans gösterdiğine işaret etmektedir. Bu sonuç, modelin Türkçe'ye özgü ince ayarlamalarla daha da geliştirilmesi gerektiğini göstermektedir.

Turkish-Llama-8b-v0.1 modelinin *%28,00* kazanma oranı, bu modelin diğer modellere göre yüskek ölçüde geride kaldığını göstermektedir. Bu sonuç, modelin Türkçe sağlık verileri üzerinde daha fazla iyileştirmeye ihtiyaç duyduğunu bir kez daha göstermektedir.

Yapay zeka hakemliğinde değerlendirme sonuçları, sentetik ölçütlerle elde edilen sonuçlarla genel manada tutarlılık göstermektedir. Bu tutarlılık, modellerin performans değerlendirmesinin güvenilir olduğunu göstermektedir.

5.1.3 Uzman Değerlendirmesi Sonuçları

Modellerin performansını gerçek dünya koşullarında değerlendirmek için, 20 farklı uzmanlık alanından doktor tarafından bir değerlendirme yapılmıştır. Doktorlar, modellerin verdiği cevapları -10 ile +10 arasında puanlamıştır. Bu değerlendirme, modellerin sağlık alanındaki pratik uygulanabilirliğini ve güvenilirliğini ölçmek açısından büyük önem taşımaktadır.

Şekil 4.10'da görülen uzman değerlendirmesi sonuçları şu şekildedir:

- *SambaLingo-Turkish-Chat*: Ortalama Puan: *3,25*
- *Trendyol-LLM-7b-chat-v1.8*: Ortalama Puan: *2,79*
- *Meta-Llama-3-8B*: Ortalama Puan: *2,43*
- *Turkish-Llama-8b-v0.1*: Ortalama Puan: *-0,23*

Uzman değerlendirmesine göre, *SambaLingo-Turkish-Chat* modeli *3,25* ortalama puanla en iyi başarıma sahiptir. Bu durum, modelin Türkçe sağlık verilerini işleme ve mantıklı cevaplar üretme konusunda diğer modellere göre daha başarılı olduğunu göstermektedir. *SambaLingo-Turkish-Chat* modelinin aynı zamanda en yüksek yararlı cevap yüzdesine (*%71,1*) sahip olması, modelin sağlık profesyonelleri tarafından daha güvenilir ve kullanışlı bulunduğunu gösterir.

Trendyol-LLM-7b-chat-v1.8 modeli *2,79* ortalama puanla ikinci en iyi performansı göstermiştir. Bu model, en düşük zararlı cevap yüzdesine (*%18,4*) sahip olmasıyla dikkat çekmektedir. Bu sonuç, modelin sağlık alanında güvenli ve etik cevaplar üretme konusunda daha başarılı olduğunu ortaya koyar.

Meta-Llama-3-8B modeli *2,43* ortalama puanla üçüncü sırada yer almıştır. Genel amaçlı bir model olmasına rağmen, Türkçe sağlık verilerinde gayet makul bir performans göstermiştir. Bu sonuç, modelin genel dil anlama ve üretme yeteneklerinin güçlü olduğunu, lakin Türkçe sağlık alanı için ince ayarlamalara ihtiyaç duyduğunu göstermektedir.

Turkish-Llama-8b-v0.1 modeli ise *-0,23* ortalama puanla negatif bir değer alarak en düşük performansı sergilemiştir. Bu durum, modelin Türkçe sağlık verileri üzerinde çok önemli iyileştirmelere ihtiyaç duyduğunu göstermektedir. Modelin en yüksek zararlı cevap yüzdesine (*%33,8*) sahip olması, sağlık alanında kullanımının potansiyel riskler taşıyabileceğine işaret etmektedir.

Uzman değerlendirmesi sonuçları, modellerin cevaplarının yararlılık, zararlılık ve nötr olma durumlarını da ortaya koymaktadır:

- *SambaLingo-Turkish-Chat*: Yararlı: *%71,1*, Zararlı: *%21,1*, Nötr: *%7,8*
- *Trendyol-LLM-7b-chat-v1.8*: Yararlı: *%69,4*, Zararlı: *%18,4*, Nötr: *%12,2*
- *Meta-Llama-3-8B*: Yararlı: *%62,6*, Zararlı: *%21,3*, Nötr: *%16,1*
- *Turkish-Llama-8b-v0.1*: Yararlı: *%42,1*, Zararlı: *%33,8*, Nötr: *%24,1*

Bu sonuçlar, *SambaLingo-Turkish-Chat* ve *Trendyol-LLM-7b-chat-v1.8* modellerinin sağlık alanında daha güvenilir ve faydalı cevaplar oluşturduğunu göstermektedir. *Turkish-Llama-8b-v0.1* modelinin ise zararlı cevap oranının çok yüksek olması, bu modelin sağlık alanında şu anki haliyle en az kullanılabilir model olduğunu göstermektedir.

Uzman değerlendirmesi sonuçları, sentetik metrikler ve yapay zeka hakemliğinde elde edilen sonuçlarla genel olarak tutarlılık göstermektedir. Bu tutarlılık, modellerin performans değerlendirmesinin güvenilirliğini artırmaktadır. Ancak, uzman değerlendirmesinin sağlık alanındaki pratik uygulanabilirliği ve etik yönleri de dikkate alması, bu ölçütü özellikle önemli yapmaktadır.

5.1.4 Genel Değerlendirme

Tüm değerlendirme ölçütleri incelendiğinde, modellerin performans sıralamasının genel manada tutarlı olduğu görülmektedir. *SambaLingo-Turkish-Chat* ve *Trendyol-LLM-7b-chat-v1.8* modelleri çoğu ölçütte en iyi performansı gösterirken, *Meta-Llama-3-8B* ve *Turkish-Llama-8b-v0.1* modelleri genelde daha düşük performans sergilemiştir.

Uzman değerlendirmesi sonuçları ile diğer ölçütler arasındaki korelasyon, bazı metriklerin sağlık alanındaki pratik uygulamalar için daha kullanışlı olduğuna işaret etmektedir. Bilhassa BLEU, ROUGE, METEOR ve BERT Skor metrikleri, uzman değerlendirmeleriyle yüksek korelasyon göstermiştir. Bu durum, bu metriklerin Türkçe sağlık metinlerinin değerlendirilmesinde güvenilir göstergeler olduğunu gösterir.

Diğer taraftan, CER ve WER ölçütleri, uzman değerlendirmeleriyle çok daha düşük korelasyon göstermiştir. Bu ölçütler, kelime ve karakter düzeyinde

hataları ölçmelerine rağmen, metinlerin anlamsal kalitesini ve sağlık alanındaki kullanışlılığını tam olarak yansıtamamış demektir.

Ama, bu çalışmada kullanılan 20 verinin sınırlı sayıda olması hasebiyle, sonuçlar%100 güvenilir değildir. Bu küçük örneklem boyutu, istatistiksel anlamlılık bakımından gayet azdır. Bununla beraber, elde edilen sonuçlar, ilerideki daha geniş çaplı araştırmalar için mühim bir başlangıç noktasıdır.

Uzman değerlendirmeleriyle yüksek korelasyon gösteren ölçütlerin (BLEU, ROUGE, METEOR, BERT Skor) Türkçe sağlık metinlerinin otomatik değerlendirilmesinde kullanılması önerilebilir. Bu ölçütler, modellerin performansını hızlı ve etkili bir şekilde değerlendirmek için kullanışlı araçlardır diyebiliriz. Ancak, CER ve WER gibi daha düşük korelasyon gösteren metriklerin bu alanda kullanımını önermiyoruz.

5.2 Sağlık Verileri Üzerinde LLM'lerin Başarısı

5.2.1 Modellerin Güçlü Yönleri

Yapılan analizler sonucunda, incelenen LLM'lerin Türkçe sağlık verileri üzerinde çeşitli başarılı ve güçlü yönleri şunlardır:

- **SambaLingo-Turkish-Chat**: Bu model, tutarlı bir şekilde en yüksek performansı göstermiştir. Özellikle BLEU, ROUGE ve METEOR gibi metin kalitesi ölçütlerinde yüksek başarıya sahiptir. Uzman değerlendirmelerinde de en yüksek puanı alarak, sağlık profesyonellerinin beklentilerini en iyi karşılayan modeldir. Modelin Türkçe dil yapısına ve sağlık terminolojisine hakimiyeti de mühim bir husustur.

- **Trendyol-LLM-7b-chat-v1.8**: Bu model, genellikle ikinci en iyi performansa sahip model olmuştur. CER ve WER ölçütlerinde en düşük hata oranlarını alması, karakter ve kelime düzeyinde doğruluk konusunda iyi bir model olduğunu ortaya koymuştur. Bununla beraber, en düşük zararlı cevap oranına sahip olması, sağlık alanında güvenli ve etik kullanım açısından çok önemli bir artıdır.

- **Meta-Llama-3-8B**: Genel amaçlı bir model olmasına karşın, Türkçe sağlık verileri üzerinde yeterli sayılabilecek bir başarım sergilemiştir. Bu, modelin güçlü doğal dil anlama ve oluşturma başarılarının farklı dillerde ve alanlarda kısmen kullanılabildiğini göstermektedir.

- **Turkish-Llama-8b-v0.1**: Bu model, diğerlerine göre daha düşük performans göstermesine karşın, Türkçe'ye özel olarak eğitilmiş olması hasebiyle geliştirilme potansiyeli vardır.

5.2.2 Modellerin Zayıf Yönleri

İncelenen modellerin bazı zayıf yönleri de tespit edilmiştir:

- **SambaLingo-Turkish-Chat**: Yüksek performansına karşın, zararlı cevap oranının (%21,1) azımsanamayacak düzeyde olmasından dolayı, sağlık gibi hassas bir alanda kullanımı soru işareti oluşturmaktadır.
- **Trendyol-LLM-7b-chat-v1.8**: BERT Skor Recall ölçütünde SambaLingo-Turkish-Chat modelinin gerisinde kalması, metin üretiminde anlamsal tutarlılık konusunda iyileştirme ihtiyacı olduğunu gösterir.
- **Meta-Llama-3-8B**: Türkçe sağlık verilerine özel olarak ince ayar yapılmamış olması, modelin performansını kısıtlamaktadır. Bilhassa Türkçe'ye özgü dil yapılarını ve sağlık terminolojisini kullanmada eksik kalmıştır.
- **Turkish-Llama-8b-v0.1**: Bu model, neredeyse tüm değerlendirme ölçütlerinde en düşük performansı sergilemiştir. Bilhassa yüksek zararlı cevap oranı (%33,8), modelin sağlık alanında kullanımında risk oluşturmaktadır.

5.2.3 Sağlık Alanında Kullanım Potansiyeli

Daha yüksek teknoloji ve parametre sayısına sahip LLM'lerin sağlık alanında kullanım potansiyeli vardır, lakin etik ve hukuki kaygılar da göz önünde bulundurulmalıdır:

- **Hasta-Doktor İletişimi**: Modeller, hastaların sorularını anlama ve cevaplama konusunda genel olarak iyi seviyede başarım göstermiştir. Bu, hasta-doktor iletişimini desteklemek için kullanılabilir, ancak doğrudan tıbbi tavsiye vermek için değil. Zira henüz o kadar gelişmemişlerdir.
- **Tıbbi Bilgi Erişimi**: LLM'ler, geniş tıbbi bilgi birikimlerini hızlı ve ucuz bir şekilde işleyebilme yetenekleriyle, sağlık profesyonellerine bilgi erişimi konusunda destek olabilir.

- **Eğitim ve Öğretim**: Modeller, tıp öğrencileri ve sağlık çalışanları için eğitim malzemeleri hazırlama ve interaktif öğrenme ortamı oluşturabilme hususunda kullanılabilir.
- **Araştırma Desteği**: LLM'ler, tıbbi literatür taraması, veri analizi ve hipotez oluşturma gibi araştırma süreçlerinde yardımcı araçlar olarak kullanılabilir.

Etik Hususlar:

Sağlık alanında LLM'lerin kullanımı, önemli etik sıkıntıları da beraberinde getirmektedir:

- **Hasta Mahremiyeti**: Modellerin eğitiminde ve kullanımında hasta verilerinin gizliliği ve güvenliği en üst düzeyde olmalıdır.
- **Doğruluk ve Güvenilirlik**: LLM'lerin ürettiği bilgilerin doğruluğu ve güvenilirliği sürekli olarak kontrol edilmelidir. Bilhassa sağlık gibi kritik bir alanda, yanlış bilgilerin ciddi derecede sıkıntılı sonuçları olabilir.
- **İnsan Gözetimi**: LLM'ler, sağlık profesyonellerinin yerini bu halleriyle almamalı, bilakis onlara destek olacak araçlar olarak görülmelidir. Önemli kararlar yakın gelecek için her zaman insan uzmanlar tarafından verilmelidir.
- **Eşitlik ve Erişilebilirlik**: LLM teknolojisinin sağlık hizmetlerinde kullanımı, toplumun tüm kesimlerine eşit erişim sağlanacak şekilde olmalıdır.
- **Şeffaflık**: Modellerin nasıl çalıştığı, hangi verileri kullandığı ve kararlarını nasıl verdiği konusunda şeffaf olunmalıdır. Bundan dolayı sağlık alanında kullanılacak LLM'lerin açık kaynaklı projeler olması önemlidir.

6
TARTIŞMA

6.1 Bulgular Işığında Modellerin Değerlendirilmesi

Bu bölümde, çalışmamızda elde edilen bulgularla birlikte kullanılan dil modellerinin performansları değerlendirilecek ve kıyaslanacaktır. Araştırmamızın sonuçları, sağlık verileri üzerinde ince ayar yapılmış LLM'lerin potansiyel gücünü ve sınırlarını ortaya çıkarmaktadır.

6.1.1 Performans Farklılıklarının Nedenleri

Çalışmamızda kullanılan dört farklı model arasında gözlemlenen performans farklılıklarının birçok sebebi bulunmaktadır:

- **Model Mimarisi**: Kullanılan modellerin temel mimarileri (LLama2, LLama3 ve Mistral) performans farklılıklarının en büyük sebebidir. Her mimarinin kendine özgü avantaj ve dezavantajlı tarafları vardır.
- **Öncül Eğitim**: Modellerin öncül eğitimlerinde kullanılan veri kümelerinin içerik ve kalitesi, sağlık alanındaki performanslarını etkilemiştir.
- **Türkçe Dil Yetkinliği**: Modellerin Türkçe dil özelindeki başarım farkları, sağlık alanındaki Türkçe verileri işleme kabiliyetlerini etkilemiştir.
- **Model Boyutu**: Kullanılan modellerin parametre sayıları arasındaki farkın, performansı etkilemesi beklenir.

6.1.2 İnce Ayarın Etkinliği ve Sınırları

İnce ayar sürecinin etkinliği ve sınırları, çalışmamızın mühim bulgularından biridir:

- **Etkinlik**: İnce ayar süreci, modellerin sağlık alanındaki performansını beklendiği gibi genel olarak artırmıştır. Bu, bilhassa ROUGE, BLEU ve BERT Score gibi metriklerle ölçülen cevap kalitesinde gözlemlenmiştir.

- **Sınırlamalar**: Bununla beraber, ince ayarın bazı sınırlamaları da ortaya çıkmıştır:
  - **Etik Konular**: İnce ayar, modellerin etik açıdan hassas sağlık konularında her zaman düzgün cevaplar vermesini sağlayamamıştır. Bu, uzman değerlendirmelerinde ortaya çıkan bir bulgudur.
  - **Dil Modeli Sınırlamaları**: LLM'lerin doğası gereği sahip oldukları bazı sınırlamaların önüne, ince ayar süreciyle tamamen geçilememiştir. Bu sınırlamalar şunlardır:
    * **Güncel Bilgi Eksikliği**: LLM'ler ne yazık ki, eğitim verilerinin tarihi kadar günceldir ve sürekli güncellenen sağlık bilgilerini otomatik olarak öğrenemezler.
    * **Bağlam Sınırlamaları**: Modeller, giriş metninin belirli bir uzunlukla sınırlı olmasından dolayı, uzun ve karmaşık tıbbi verileri tam manasıyla anlayamazlar.
    * **Nedensel Çıkarım Zorlukları**: LLM'ler şu anki halleriyle, tıbbi semptomlar ve teşhisler arasındaki karmaşık nedensel alakaları her an doğru bir biçimde çıkaramazlar.
    * **Belirsizlik Yönetimi**: Sağlık alanındaki belirsizlikleri ve olasılıkları tahmin edip cevaplamakta zorluk çekebilirler, bu da yanlış teşhis veya tedavi önerilerine yol açar.
    * **Etik Karar Verme**: Tıbbi etik konularında insan seviyesinde karar verme özelliğine sahip değillerdir. Bu da sağlık gibi hassas durumlarda sorunlara neden olur.
    * **Multimodal Veri İşleme**: Sadece metin tabanlı verileri öğrendikleri için, görüntü veya ses gibi multimodal tıbbi verileri doğrudan anlayamazlar.
    * **Açıklanabilirlik Eksikliği**: LLM'lerin kara kutu yapısı, tıbbi kararların nasıl alındığını analiz etmemizi zorlaştırır. Bu da sağlık profesyonellerinine yeterince yardımcı olamayacakları anlamına gelir.

Bu bulgular, sağlık verileri üzerinde LLM'lerin ince ayarının potansiyelini göstermekle birlikte, bu sürecin dikkatli bir şekilde yönetilmesi ve sonuçların eleştirel bir gözle değerlendirilmesi gerektiğini ortaya koymaktadır. Bilhassa etik konularda, modellerin kararlarının her zaman bir sağlık profesyoneli vasıtasıyla incelenmesinin lazım geldiği unutulmamalıdır.

6.2 Sağlık Alanında LLM Kullanımının İmkanları ve Zorlukları

6.2.1 Potansiyel Uygulama Alanları

LLM'ler, sağlık sektöründe çeşitli alanlarda potansiyel kullanım sahasına sahiptir:

- **Tıbbi Literatür Analizi**: LLM'ler, geniş tıbbi literatürü hızlı ve ucuz bir şekilde tarayarak araştırmacılara ve klinisyenlere hızlı ve özet bilgileri sunabilir.
- **Klinik Karar Destek Sistemleri**: Doktorlara teşhis ve tedavi süreçlerinde destek olabilir, özellikle yaygın hastalıklarda teşhis ve tedavi sürecini hızlandırabilir.
- **Hasta-Doktor İletişimi**: Hastaların sorularını anlayıp cevaplamada ve tıbbi mefhumları açıklamada yardımcı olabilir. Bu şekilde doktorların işi kolaylaşmış olur.
- **Tıbbi Raporlama**: Klinik notların ve raporların üretilmesinde zaman tasarrufu sağlayabilir.
- **Sağlık Eğitimi**: Tıp öğrencileri ve sağlık profesyonelleri için kişiselleştirilmiş eğitim materyalleri oluşturabilir. Aynı zamanda onlara koçluk yapabilir.

6.2.2 Karşılaşılabilecek Zorluklar ve Çözüm Önerileri

LLM'lerin sağlık alanında kullanımı bazı sıkıntılara da yol açmaktadır:

- **Veri Gizliliği ve Güvenliği**: **Zorluk**: Hassas sağlık verilerinin korunması. **Çözüm Önerisi**: Gelişmiş şifreleme teknikleriyle veri güvenliğinin sağlanması. Verinin anonimleştirilmesi.
- **Etik Konular**: **Zorluk**: LLM'lerin etik karar verme performanslarının yetersizliği. **Çözüm Önerisi**: Etik kurulların kurulması ve LLM çıktılarının sürekli olarak insan uzmanlar tarafından denetlenmesi. Bununla beraber etiğe uygun veri kümeleri oluşturulması.
- **Doğruluk ve Güvenilirlik**: **Zorluk**: LLM'lerin bazen yanlış veya tutarsız veriler oluşturabilmesi ve hatta halisülasyon görmesi. **Çözüm Önerisi**: Sürekli modelleri değerlendirme ve geliştirme, insan denetimi ve çapraz doğrulama gibi yöntemlerin uygulanması.

- **Yasal ve Düzenleyici Çerçeveler**: **Zorluk**: LLM'lerin sağlık sektöründe kullanımına ilişkin yasal boşluklar. **Çözüm Önerisi**: Sağlık otoriteleri, hukukçular ve yapay zeka uzmanlarının takım çalışmasıyla uygun kanun ve yönetmeliklerin çıkarılması.
- **Entegrasyon Zorlukları**: **Zorluk**: LLM'lerin mevcut sağlık sistemine entegrasyonu. **Çözüm Önerisi**: Tedrici entegrasyon yöntemleri ve sağlık profesyonellerine müteveccih eğitim programlarının geliştirilmesi.
- **Açıklanabilirlik Eksikliği**: **Zorluk**: LLM'lerin kararlarının arkasındaki mantığın anlaşılması. **Çözüm Önerisi**: Açıklanabilir YZ yöntemlerinin geliştirilmesi ve kullanılması.

Bu zorluklar aşmak için, çok disiplinli bir yaklaşım tercih edilmeli ve sağlık profesyonelleri, teknoloji uzmanları, etik uzmanları ve politikacılar arasında sürekli işbirliği ortamı oluşturulmalıdır. LLM'lerin sağlık alanında güvenilir ve sürekli bir biçimde kullanılabilmesi için, bu teknolojinin sürekli olarak denetlenmesi, iyileştirilmesi ve düzenlenmesi lazımdır.

6.3 Etik Değerlendirme

6.3.1 Hasta Mahremiyeti ve Veri Güvenliği

Sağlık alanında LLM'lerin kullanımı, hasta mahremiyeti, veri güvenliği ve gizliliği açısından mühim etik sıkıntıları da beraberinde getirmektedir:

- **Veri Koruma**: Özel sağlık verilerinin LLM'ler tarafından işlenmesi, bu verilerin güvenliğini ve gizliliğini sağlama konusunda zorluklar oluşturmaktadır.
- **Anonimleştirme**: Hasta bilgileri gerektiği kadar anonimleştirilmediği zaman, şahısların şahsi bilgilerinin açığa çıkma riski vardır.
- **Veri Paylaşımı**: LLM'lerin eğitimi için veri paylaşımı gereklidir. Lakin bu durum hasta mahremiyetini sıkıntıya sokabilir.

6.3.2 Modellerin Zararlı İçerik Üretme Riski

LLM'lerin sağlık alanında kullanımı, zararlı içerik üretme riski taşımaktadır:

- **Yanlış Bilgi**: Modeller, eski ya da yanlış tıbbi veriler oluşturabilir. Bu da hastaların sağlığı için potansiyel risk oluşturur.

- **Eğilim**: LLM'ler, eğitim verilerindeki eğilimlerden dolayı belirli hastalıklara karşı yanlış önerilerde bulunabilir.
- **Etik Dışı Öneriler**: Modeller, etik olmayan veya yasal olmayan tıbbi çözümler üretebilir.

6.3.3 Sağlık Profesyonellerinin Rolü ve LLM'lerin Sınırları

LLM'lerin sağlık sahasında kullanımı, sağlık profesyonellerinin rolünü ve bu teknolojinin sınırlarını yeniden yazmayı gerektirmektedir:

- **İnsan Denetimi**: LLM'lerin çıktılarının her zaman bir sağlık profesyoneli vasıtasıyla değerlendirilmesi ve onaylanması gerekmektedir.
- **Sorumluluk Sınırları**: LLM'lerin sağlık kararlarındaki rolü kesin ve açık bir şekilde tanımlanmalı ve yasal sorumluluk sınırları çizilmelidir.
- **Sürekli Eğitim**: Sağlık profesyonellerinin LLM'leri doğru bir biçimde kullanabilmeleri için sürekli eğitim almaları gerekmektedir.

6.3.4 Hukuki Boşluklar ve Sorumluluk Belirsizliği

LLM'lerin sağlık alanında kullanımı, mevcut yasalarda önemli boşluklar ortaya çıkarmaktadır:

- **Yasal Sorumluluk**: LLM'lerin önerilerinden dolayı yanlış tedavi uygulanması, yanlış yönlendirme vb. durumlarda kimin sorumlu olacağı (model geliştiriciler, sağlık kurumları veya hekimler) belirsizdir.
- **Yasa Eksikliği**: LLM'lerin sağlık sahasında kullanımıyla alakalı özel yasal düzenlemelerin eksikliği, uygulamada bilinmezliklere sebebiyet vermektedir.
- **Sınır Ötesi Kullanım**: LLM'lerin farklı ülke hatta kıtalarda kullanımı, uluslararası hukuk ve yargı yetkisi hususlarında karmaşık sorunlara sebebiyet vermektedir.

Bu etik ve hukuki zorlukları aşabilmek için, sağlık otoriteleri, hukukçular, etik uzmanları ve yapay zeka uzmanlarının işbirliği içinde çalışması gerekmektedir. Bununla beraber, LLM'lerin sağlık alanında kullanımına ilişkin kapsamlı yasaların oluşturulması ve düzenli olarak geliştirilmesi çok önemlidir.

6.4 Çalışmanın Kısıtlamaları

Bu bölümde, çalışmamızın çeşitli kısıtlamalarını gözden geçireceğiz. Bu kısıtlamaları bilmek, sonuçlarımızın yorumlanması ve ileriki araştırmalar için mühimdir.

6.4.1 Veri Kümesi Kısıtlamaları

- **Veri Kümesi Boyutu**: Kullandığımız veri kümesi, her ne kadar geniş bir veri kümesi olsa da (321.179 soru-cevap çifti), tüm olası tıbbi durumları ve soru türlerini kapsamadığı barizdir.
- **Veri Kalitesi**: Veri kümesindeki soru-cevap çiftlerinin kalitesi ve doğruluğu, modellerin performansını doğrudan etkiler. Veri kümesini denetleyen birisi olmadığı için hatalı, eksik ya da eski cevaplar dahi olabilir.
- **Dil ve Kültürel Sınırlamalar**: Veri kümesi Türkçe dilinde olduğu için, sonuçlar diğer dilleri ve kültürleri kapsamayabilir. Hatta Türkiye içindeki kültür ve hastalık yoğunluğu farkına göre dahi modellerin performansı değişebilir.

6.4.2 Metodolojik Kısıtlamalar

- **Model Seçimi**: Çalışmamızda kullanılan dört model, mevcut tüm LLM'leri temsil etmez, edemez.
- **İnce Ayar Süreci**: Kullanılan ince ayar yöntemleri ve hiperparametreler, tüm olası varyasyonları kapsamaz. Farklı yaklaşımlarla daha iyi sonuçlar alınabilir.
- **Değerlendirme Metrikleri**: Kullanılan ölçütler (ROUGE, BLEU, BERT Score, vb.) kapsamlı olsa dahi, modellerin gerçek dünya performansını tam olarak yansıtmaz.

6.4.3 Teknik Kısıtlamalar

- **Hesaplama Kaynakları**: LLM'ler şu anda gerçekten yüksek donanım kaynakları istiyorlar. Bilhassa devasa VRAM kaynakları istiyorlar. Bunu karşılamak şahsi olarak çok zor. Bundan dolayı da *13 milyar* ve daha büyük parametreye sahip modelleri deneyemedik.

- **Model Boyutu**: Çalışmada kullanılan modellerin parametre sayılarından daha büyük parametre sayılarıan sahip modeller mevcuttur. Daha büyük modellerin daha iyi sonuçlar vermesi beklenir.

Bu kısıtlamalar, çalışmamızın neticelerinin yorumlanmasında ve ilerideki araştırmalarda önemsenmelidir. İlerideki çalışmalar, bu sınırlamalardan kurtulmak için daha geniş veri kümeleri, farklı diller ve kültürler, daha farklı ve büyük model mimarileri ve daha gelişmiş değerlendirme ölçütleri kullanabilir.

7
SONUÇ

7.1 Ana Bulgular ve Çıkarımlar

Çalışmanın ana bulguları şu şekildedir:

- **SambaLingo-Turkish-Chat** modeli, çoğu ölçütte en iyi başarımı göstermiştir:
  - BLEU, ROUGE ve METEOR gibi metin kalitesi ölçütlerinde en yüksek puanları almıştır.
  - Elo puanlamasında ve Kazanma Yüzdesi ölçütlerinde birinci sırada yer almıştır.
  - Uzman değerlendirmesinde en yüksek puanı almıştır.
  - Uzman değerlendirmelerine göre en yüksek oranda yararlı cevap üreten model olmuştur.
  - BERT Score ölçütünde de diğer modellere göre daha iyi bir performans göstermiştir.
- **Trendyol-LLM-7b-chat-v1.8** modeli, genel olarak ikinci en iyi başarımı göstermiştir:
  - CER ve WER ölçütlerinde en düşük hata oranlarını elde etmiştir.
  - Uzman değerlendirmelerine göre en düşük zararlı cevap oranına sahip modeldir.
  - BLEU ve ROUGE ölçütlerinde SambaLingo-Turkish-Chat modelinden hemen sonra gelmektedir.
  - Uzman değerlendirmelerine ikinci en yüksek puanı almıştır.
- **Meta-Llama-3-8B** modeli, genel amaçlı bir model olmasına karşın, Türkçe sağlık verilerinde makul bir performans göstermiştir:

  - BERT Score ölçütünde SambaLingo-Turkish-Chat ve Trendyol-LLM-7b-chat-v1.8 modellerinden sonra üçüncü sırada yer almıştır.
  - Elo puanlamasında ve Kazanma Yüzdesi ölçütlerinde de üçüncü sıradadır.
  - Uzman değerlendirmesinde de üçüncü sıradadır.
  - Zararlı cevap oranı açısından da Turkish-Llama-8b-v0.1 modelinden daha iyi performans göstermiştir.
- **Turkish-Llama-8b-v0.1** modeli, çoğu ölçütte en düşük performansı sergilemiştir:
  - BLEU, ROUGE, METEOR ve BERT Score ölçütlerinde en düşük puanları almıştır.
  - CER ve WER ölçütlerinde en yüksek hata oranlarını elde etmiştir.
  - Elo puanlamasında ve Kazanma Yüzdesi ölçütlerinde son sırada yer almıştır.
  - Uzman değerlendirmesinde en düşük puanı almıştır.
  - En yüksek zararlı cevap oranına sahip model olmuştur.

Bu bulgular ışığında şu çıkarımları yapabiliriz:

- Dile özgü ve alana özgü ince ayar yapılması, LLM'lerin performansını önemli ölçüde artırabilmektedir. *SambaLingo-Turkish-Chat ve Trendyol-LLM-7b-chat-v1.8* modellerinin başarısı, ince ayar işlemlerinin Türkçe sağlık verileri için başarıyla yapıldığını göstermektedir.
- Genel amaçlı modeller (Meta-Llama-3-8B gibi), özel alanlarda da göz ardı edilemez bir performans gösterebilmektedir. Lakin, parametre sayısındaki üstünlüğüne rağmen, dile ve alana özgü eğitilmiş modeller kadar başarılı olamamaktadırlar.
- Model performansını ölçerken, farklı metriklerin beraber kullanılması daha kapsamlı bir çıkarım yapabilmemize vesile olmuştur. Mesela, BLEU ve ROUGE gibi metin kalitesi ölçütleri ile CER ve WER gibi hata oranı ölçütlerinin birlikte değerlendirilmesi, modellerin farklı açılardan performansını anlayabilmemize olanak tanımıştır.
- Uzman değerlendirmeleri, otomatik ölçütlerle elde edilen sonuçlarla paralel seyretmiştir.

- Zararlı cevap oranları, modellerin güvenilirliği veçhile önemli bir göstergedir. Bu taraftan bakıldığında *Trendyol-LLM-7b-chat-v1.8* modelinin en düşük zararlı cevap oranına sahip olması, sağlık gibi önemli ve hassas sahalarda kullanım için önemlidir.

Bu sonuçlar, Türkçe sağlık alanında LLM'lerin ileride kullanılabileceğini göstermektedir. Tabii LLM'lerin hala geliştirilmesi gereken yanları vardır.

7.2 Çalışmanın Katkıları

Bu çalışmanın başlıca katkıları şunlardır:

- Türkçe sağlık verileri üzerinde LLM'lerin performansını kapsamlı bir şekilde değerlendiren ilk çalışmalardan biridir.
- Hasta-doktor iletişimine özel bir veri kümesi oluşturulmuş ve bu alanda kullanılabilecek özelleştirilmiş modeller geliştirilmiştir.
- Farklı değerlendirme ölçütlerinin karşılaştırmalı analizi yapılmış, hangi ölçütlerin Türkçe sağlık metinlerinin değerlendirilmesinde daha uygun olduğu gösterilmiştir.
- Sağlık alanında LLM'lerin kullanımıyla alakalı etik ve hukuki konular tartışılmıştır.

7.3 Gelecek Araştırmalar için Öneriler

Bu çalışmanın sonuçları ve kısıtlamaları ışığında, gelecekteki araştırmalar için öneriler aşağıda verilmiştir:

- **Veri Kümesi Genişletme**: Daha geniş ve kapsamlı bir Türkçe sağlık veri kümesi oluşturulması, modellerin performansını artıracaktır, daha kapsamlı ve beğenilen çıktılar verilmesine yardımcı olacaktır.
- **Farklı Model Mimarileri**: Çalışmamızda kullanılan modeller haricinde, farklı mimarilere ve parametre sayılarına sahip LLM'lerin de sağlık alanında eğitilip test edilmesi, en uygun model seçimi konusunda daha kapsamlı bilgi sunacaktır.
- **Uzun Vadeli Etki Analizi**: LLM'lerin sağlık alanında kullanımının uzun vadeli yan etkilerini incelemek için uzunlamasına çalışmalar yapılabilir.

- **Etik ve Hukuki Çerçeve Geliştirme**: Sağlık alanında LLM kullanımına müteveccih etik kuralların ve yasal düzenlemelerin yapılması için disiplinlerarası çalışmalar yürütülebilir.
- **İnsan-YZ İşbirliği**: Sağlık profesyonellerinin LLM'leri en etkili şekilde nasıl kullanabileceğine dair araştırmalar yapılabilir.
- **Özelleştirilmiş Modeller**: Özel bir tıbbi uzmanlık alanı için ince ayar yapılmış modeller geliştirilebilir.
- **Multimodal Modeller**: Metin, görüntü ve ses verilerini birlikte işleyebilen multimodal LLM'lerin sağlık alanında kullanımı araştırılabilir.
- **Açıklanabilirlik**: LLM'lerin sağlık alanında oluşturduğu cevapların açıklanabilmesi için yeni yöntemler geliştirilebilir.
- **Veri Güvenliği ve Gizlilik**: Sağlık verilerinin gizliliğinden ödün vermezken aynı zamanda bu verilerle LLM'lerin eğitilmesi ve kullanılabilmesi için yeni yöntemlerin araştırılabilir.

7.4 Kapanış Düşünceleri

Bu çalışma, LLM'lerin Türkçe sağlık alanında aydsınamaz bir potansiyele sahip olduğunu gözler önüne sermektedir. Elde edilen sonuçlar potansiyel vadetmekle birlikte, bu teknolojinin güvenli ve etik bir şekilde kullanılabilmesi için daha fazla araştırma ve geliştirme çalışması lazımdır.

Özellikle vurgulanması gereken bir nokta, Sağlık Bakanlığı gibi büyük kurumlardan alınabilecek veri ve donanım desteğidir. Bu tür kurumsal işbirlikleri, çok daha gelişmiş çalışmalar yapmamıza vesile olabilir. Zira LLM'lerin eğitilebimesi, bireysel çabalarla başarılması oldukça zor ve hatta yüksek parametreli modeller için neredeyse imkansızdır.

KAYNAKÇA

[1] meta-llama, *meta-llama/Meta-Llama-3-8B*, 2024. erişim adresi: https://huggingface.co/meta-llama/Meta-Llama-3-8B (erişim tarihi 16/08/2024).

[2] ytu-ce-cosmos, *ytu-ce-cosmos/Turkish-Llama-8b-v0.1*, 2024. erişim adresi: https://huggingface.co/ytu-ce-cosmos/Turkish-Llama-8b-v0.1 (erişim tarihi 16/08/2024).

[3] sambanovasystems, *sambanovasystems/SambaLingo-Turkish-Chat*, 2024. erişim adresi: https://huggingface.co/sambanovasystems/SambaLingo-Turkish-Chat (erişim tarihi 16/08/2024).

[4] Trendyol, *Trendyol/Trendyol-LLM-7b-chat-v1.8*, 2024. erişim adresi: https://huggingface.co/Trendyol/Trendyol-LLM-7b-chat-v1.8 (erişim tarihi 16/08/2024).

[5] M. K. Bulut, *Patient Doctor Q&A TR 321179*, 2024. doi: 10.5281/zenodo.12798934. erişim adresi: https://doi.org/10.5281/zenodo.12798934.

[6] A. E. Elo S. Sloan, *The rating of chessplayers: Past and present*. New York: Arco Pub., 1978.

[7] M. A. Bayram, *Türkçe Tıbbi Soru-Cevap Veri Seti*, 2024. doi: 10.5281/zenodo.12770916. erişim adresi: https://zenodo.org/record/12770916.

[8] M. K. Bulut, *Patient-Doctor QA Dataset (TR)*, 2024. erişim adresi: https://huggingface.co/datasets/kayrab/patient-doctor-qa-tr-321179.

[9] M. K. Bulut, *Patient-Doctor QA Dataset (TR)*, 2024. erişim adresi: https://huggingface.co/datasets/kayrab/patient-doctor-qa-tr-5695.

[10] M. K. Bulut, *Patient-Doctor QA Dataset (TR)*, 2024. erişim adresi: https://huggingface.co/datasets/kayrab/patient-doctor-qa-tr-95588.

[11] M. K. Bulut, *Patient-Doctor QA Dataset (TR)*, 2024. erişim adresi: https://huggingface.co/datasets/kayrab/patient-doctor-qa-tr-19583.

[12] Y. Peng, S. Yan Z. Lu, "Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets," *arXiv preprint arXiv:1906.05474*, 2019.

[13] C.-W. Park ve diğ., "Artificial intelligence in health care: current applications and issues," *Journal of Korean medical science*, c. 35, no. 42, 2020.

[14] F. Jiang ve diğ., "Artificial intelligence in healthcare: past, present and future," *Stroke and vascular neurology*, c. 2, no. 4, 2017.

[15] E. Chikhaoui, A. Alajmi S. Larabi-Marie-Sainte, "Artificial intelligence applications in healthcare sector: ethical and legal challenges," *Emerging Science Journal*, c. 6, no. 4, ss. 717–738, 2022.

[16] G. Uludoğan, Z. Y. Balal, F. Akkurt, M. Türker, O. Güngör S. Üsküdarlı, "Turna: A turkish encoder-decoder language model for enhanced understanding and generation," *arXiv preprint arXiv:2401.14373*, 2024.

[17] Y. Liu ve diğ., "Multilingual denoising pre-training for neural machine translation," *Transactions of the Association for Computational Linguistics*, c. 8, ss. 726–742, 2020.

[18] L. Xue ve diğ., "mT5: A massively multilingual pre-trained text-to-text transformer," *arXiv preprint arXiv:2010.11934*, 2020.

[19] E. Dogan ve diğ., "Türkçe Dil Modellerinin Performans Karşılaştırması Performance Comparison of Turkish Language Models," *arXiv e-prints*, arXiv–2404, 2024.

[20] H. T. Kesgin ve diğ., "Introducing cosmosGPT: Monolingual Training for Turkish Language Models," *arXiv preprint arXiv:2404.17336*, 2024.

[21] OpenAI, *GPT-4o: OpenAI's Latest Language Model*, 2024. erişim adresi: `https://openai.com/index/hello-gpt-4o/` (erişim tarihi 16/08/2024).

[22] Anthropic, *Claude: A New AI Assistant by Anthropic*, 2024. erişim adresi: `https://www.anthropic.com/news/claude-3-5-sonnet` (erişim tarihi 16/08/2024).

[23] Google, *Gemini: Google's AI Model for Multimodal Understanding*, 2024. erişim adresi: `https://deepmind.google/technologies/gemini/pro/` (erişim tarihi 16/08/2024).

[24] Microsoft, *GitHub Copilot: AI-Powered Code Completion by Microsoft*, 2024. erişim adresi: `https://copilot.microsoft.com/` (erişim tarihi 16/08/2024).

[25] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[26] T. B. Brown, "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, 2020.

[27] A. Radford, "Improving language understanding by generative pre-training," 2018.

[28] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever ve diğ., "Language models are unsupervised multitask learners," *OpenAI blog*, c. 1, no. 8, s. 9, 2019.

[29] H. Touvron ve diğ., "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[30] J. Devlin, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[31] V. Sanh, "DistilBERT, A Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[32] M. Lewis, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv:1910.13461*, 2019.

[33] A. Roberts ve diğ., "Exploring the limits of transfer learning with a unified text-to-text transformer," *Google, Tech. Rep.*, 2019.

[34] F. Zhuang ve diğ., "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, c. 109, no. 1, ss. 43–76, 2020.

[35] M. K. Bulut, *Patient-Doctor QA Dataset (TR)*, 2024. erişim adresi: https://huggingface.co/datasets/kayrab/patient-doctor-qa-tr-167732.

[36] OpenAI, *GPT-3.5 Turbo*, 2024. erişim adresi: https://platform.openai.com/docs/models/gpt-3-5-turbo (erişim tarihi 15/07/2024).

[37] I. D. Hermansyah, *Doctor-ID-QA Dataset*, 2024. erişim adresi: https://huggingface.co/datasets/hermanshid/doctor-id-qa.

[38] Henry41, *iCliniq Medical QA Dataset*, 2024. erişim adresi: https://www.kaggle.com/datasets/henry41148/icliniq-medical-qa.

[39] S. Systems, *SambaLingo-Turkish-Base*, 2024. erişim adresi: https://huggingface.co/sambanovasystems/SambaLingo-Turkish-Base (erişim tarihi 16/08/2024).

[40] M. AI, *Llama-2-7b-hf*, 2024. erişim adresi: https://huggingface.co/meta-llama/Llama-2-7b-hf (erişim tarihi 16/08/2024).

[41] L. C. Arena, *LMSYS Chatbot Arena Leaderboard*, 2024. erişim adresi: https://chat.lmsys.org/?leaderboard (erişim tarihi 16/08/2024).

[42] Y. Chen ve diğ., "Towards learning universal hyperparameter optimizers with transformers," *Advances in Neural Information Processing Systems*, c. 35, ss. 32 053–32 068, 2022.

[43] J. Hoffmann ve diğ., "Training compute-optimal large language models," *arXiv preprint arXiv:2203.15556*, 2022.

[44] F. Ç. Akyön, A. D. E. ÇAVUŞOĞLU, C. Cengiz, S. O. Altinuç A. Temizel, "Automated question generation and question answering from Turkish texts," *Turkish Journal of Electrical Engineering and Computer Sciences*, c. 30, no. 5, ss. 1931–1940, 2022.

[45] C. Raffel ve diğ., "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of machine learning research*, c. 21, no. 140, ss. 1–67, 2020.

[46] S. Wu M. Sun, "Exploring the efficacy of pre-trained checkpoints in text-to-music generation task," *arXiv preprint arXiv:2211.11216*, 2022.

[47] T. Dettmers, M. Lewis, S. Shleifer L. Zettlemoyer, "8-bit optimizers via block-wise quantization," *arXiv preprint arXiv:2110.02861*, 2021.

[48] G. Henry, P. T. P. Tang A. Heinecke, "Leveraging the bfloat16 artificial intelligence datatype for higher-precision computations," *2019 IEEE 26th Symposium on Computer Arithmetic (ARITH)*, IEEE, 2019, ss. 69–76.

[49] R. Kong ve diğ., "LoRA-Switch: Boosting the Efficiency of Dynamic LLM Adapters via System-Algorithm Co-design," *arXiv preprint arXiv:2405.17741*, 2024.

[50] Unsloth, *Unsloth: Finetune Llama 3.1, Mistral, Phi & Gemma LLMs 2-5x faster with 80% less memory*, 2024. erişim adresi: `https://github.com/unslothai/unsloth` (erişim tarihi 08/08/2024).

[51] NVIDIA, *NVIDIA A100 Tensor Core GPU*, 2024. erişim adresi: `https://www.nvidia.com/tr-tr/data-center/a100/` (erişim tarihi 08/08/2024).

[52] Google, *Google Colab*, 2024. erişim adresi: `https://colab.google/` (erişim tarihi 08/09/2024).

[53] Meta AI, *LLaMA 3.1: Meta's Next-Generation Large Language Model*, 2024. erişim adresi: `https://huggingface.co/meta-llama/Meta-Llama-3.1-70B` (erişim tarihi 08/08/2024).

[54] W.-L. Chiang ve diğ., *Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference*, 2024. arXiv: `2403.04132 [cs.AI]`.

[55] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," *Text summarization branches out*, 2004, ss. 74–81.

[56] K. Papineni, S. Roukos, T. Ward W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, ss. 311–318.

[57] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger Y. Artzi, "Bertscore: Evaluating text generation with bert," *arXiv preprint arXiv:1904.09675*, 2019.

[58] T. Zhang, F. Ladhak, E. Durmus, P. Liang, K. McKeown T. B. Hashimoto, "Benchmarking large language models for news summarization," *Transactions of the Association for Computational Linguistics*, c. 12, ss. 39–57, 2024.

[59] S. Gehrmann ve diğ., "The gem benchmark: Natural language generation, its evaluation and metrics," *arXiv preprint arXiv:2102.01672*, 2021.

[60] A. C. Morris, V. Maier P. D. Green, "From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition.," *Interspeech*, 2004, ss. 2765–2768.

[61] W. Xiong, L. Wu, F. Alleva, J. Droppo, X. Huang A. Stolcke, "The Microsoft 2017 conversational speech recognition system," *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2018, ss. 5934–5938.

[62] A. Hannun, “Deep Speech: Scaling up end-to-end speech recognition,” *arXiv preprint arXiv:1412.5567*, 2014.

[63] C. Lopes F. Perdigao, “Phone recognition on the TIMIT database,” *Speech Technologies/Book*, c. 1, ss. 285–302, 2011.

[64] Y.-Y. Wang, A. Acero C. Chelba, “Is word error rate a good indicator for spoken language understanding accuracy,” *2003 IEEE workshop on automatic speech recognition and understanding (IEEE Cat. No. 03EX721)*, IEEE, 2003, ss. 577–582.

[65] S. Banerjee A. Lavie, “METEOR: An automatic metric for MT evaluation with improved correlation with human judgments,” *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005, ss. 65–72.

TEZDEN ÜRETİLMİŞ YAYINLAR

Makale

1. Paper 1
2. Paper 2

Konferans Bildirisi

1. Conference 1
2. Conference 2

Kitap

1. Book 1
2. Book 2

Proje

1. Project 1
2. Project 2

Ödül

1. Award 1
2. Award 2