T.C.
YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

SAĞLIK ALANINDA LLM'LER

Muhammed Kayra BULUT

YÜKSEK LISANS TEZİ
Bilgisayar Bilimleri Anabilim Dalı
Bilgisayar Bilimleri Programı

Danışman
Prof. Dr. Banu DİRİ

Mayıs, 2024

T.C.
YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

SAĞLIK ALANINDA LLM'LER

Muhammed Kayra BULUT tarafından hazırlanan tez çalışması 01.05.2024 tarihinde aşağıdaki jüri tarafından Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Bilimleri Anabilim Dalı Bilgisayar Bilimleri 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. Hamza Osman İLHAN, Üye
Yildiz Technical University _____

Danışmanım Prof. Dr. Banu DİRİ sorumluluğunda tarafımca hazırlanan Sağlık Alanında LLM'ler 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

Dedicated to my family
and my best friend

TEŞEKKÜR

Projemin her aşamasında benden desteklerini esirgemeyen Banu Diri ve Himmet Toprak Kesgin'e çok teşekkür ederim. Akademik yolculuğumda bana rehber oldukları, bilgi ve tecrübelerini bana aktardıkları için minnettarım. Bu çalışmanın her sayfasında onların katkıları vardır. Banu Hanım ve Himmet Bey'in profesyonellikleri, bilgileri ve destekleri olmasaydı bu tez projem bu kadar başarılı olamazdı.

Muhammed Kayra BULUT

İÇİNDEKİLER

SİMGE LİSTESİ

Ai	Activities of Daily Life
c	Alternate Step Test
C	Body Mass Index
CR	Cross Step moving on Four Stops
$fc(.)$	Dynamic Bayesian Networks
ΔH	Demura's Fall Risk Assessment Chart
λi	Electromyography
Ω	Faculdade de Engenharia da Universidade do Porto

KISALTMA LİSTESİ

CLS	Classification
DDİ	Doğal Dil İşleme
EHR	Elektronik Sağlık Kayıtları
EM	Exact Match
GPT	Generative Pre-trained Transformer
GPT	Grafik İşleme Ünitesi
i2b2	Informatics for Integrating Biology and the Bedside
LLM	Büyük Dil Modeli
SEP	Separator
TPU	Tensor İşleme Ünitesi
UNK	Unknown

ŞEKİL LİSTESİ

ÖZET

Sağlık Alanında LLM'ler

Muhammed Kayra BULUT

Bilgisayar Bilimleri Anabilim Dalı
Yüksek Lisans Tezi

Danışman: Prof. Dr. Banu DİRİ

Düşünün ki her doktorun ve hemşirenin her durumda, hastalarla konuşmaktan tutun da tıbbi araştırmalara derinlemesine girmeye kadar yardımcı olacak bir AI asistanı var. Bu, imkansız bir hayal değil; sağlık sektörüne özgü dil modelleri sayesinde gerçekleşmekte olan bir gerçek. Bu özel araçların oyunun kurallarını nasıl değiştirdiğine yakından bakalım.

Arka Plan: Büyük Dil Modelleri'nin (LLM) gelişimi, insanları sağlık sektörü dahil çeşitli sektörlerde de bu modellerin kullanılabilirliğini sorgulamaya itmiştir. Bu proje, sağlık alanında özelleştirilmiş dil araçları olan CLAIR-Short [1], CLAIR-Long [1], Clinical-Longformer [2], Clinical-BigBird [2] ve Almanac'ın [3], genel araçlar olan ChatGPT, Bing ve Bard vb. ile karşılaştırıldığında nasıl bir performans gösterdiğini incelemektedir. Amaç, klinik kararların kalite ve hızını artırmada, hastalarla konuşmada ve tıbbi araştırmaları ne kadar etkin ve hızlı şekilde yapabileceğimizde özelleştirilmiş dil araçlarının etkisini ortaya koymaktadır.

Yöntemler: Sağlık hizmetlerinde dil araçlarının nasıl performans gösterdiğine dair birkaç önemli çalışmayı detaylı bir şekilde inceledik. Bu, hasta mesajlarına cevap oluşturmak için ince ayar (Fine-Tune) yapılmış araçlar olan CLAIR-Short ve CLAIR-Long'un performansının yanı sıra, tedavi önerileri için dikkatlice seçilmiş tıbbi bilgilerden alınan bilgilerle dil araçlarını güçlendiren Almanac modelinin incelenmesini içerdi. Uzun tıbbi metinleri ele almak için Clinical-Longformer ve Clinical-BigBird araçlarının performansına da baktık. Değerlendirme kriterlerimiz, klinik ortamlarda doğru, tam ve kullanıcı dostu yanıtlar üretebilme yetenekleri, hassas tıbbi verileri güvenle işleme kapasiteleri ve güvenliklerine odaklandı.

Sonuçlar: Özelleştirilmiş araçlar, gerçeklik, tamamlık, kullanıcı tercihi ve zararlı kullanıcılara karşı güvenlik dahil birçok alanda, daha genel olanlara göre iyi performans gösterdi. Mesela, CLAIR-Short ve CLAIR-Long, hastalar ve doktorlar arasında net ve faydalı yanıtlar üreterek iletişimi büyük ölçüde hızlandırdı ve iyileştirdi. Almanac modelinin arama destekli olması, en güncel tıbbi kaynaklara erişim sağlayarak karar destek sistemini geliştirdi. Ayrıca, Clinical-Longformer ve Clinical-BigBird gibi özelleştirilmiş araçlar, tıbbi araştırma ve literatür inceleme süreçlerini akıcı hale getirme potansiyelini gösteren uzatılmış tıbbi belgelerle test edildiğine, yüksek başarımlar verdi.

Sonuç: Bu projeden elde edilen bulgular, özellikle araçlar belirli tıbbi kullanımlar için özelleştirildiğinde, sağlık hizmetlerinde dil araçlarının çok büyük potansiyele sahip olduğunu vurgulamaktadır. CLAIR-Short, CLAIR-Long, Clinical-Longformer, Clinical-BigBird ve Almanac gibi özelleştirilmiş dil araçları, sadece tıbbi bilgileri işlemede daha yüksek doğruluk ve güvenilirlik sağlamakla kalmaz, aynı zamanda yüksek güvenlik sağlar ve kullanıcılar tarafından daha çok tercih edilen yanıtlar sağlar. Bu iyileştirmeler, günlük tıbbi işlere dil araçlarının entegre edilmesi gerektiğini göstermekte olup, Yapay Zeka'nın sağlık hizmetileri ve tıbbi araştırmalara büyük fark oluşturacağı bir geleceğe işaret etmektedir.

Anahtar Kelimeler: Büyük Dil Modelleri, Sağlık Hizmetleri, CLAIR Modelleri, Almanac, Clinical-Longformer, Clinical-BigBird, Tıpta Yapay Zeka, Hasta İletişimi, Tıbbi Araştırma.

ABSTRACT

Thesis Title

Muhammed Kayra BULUT

Department of Computer Science
Master of Science Thesis

Supervisor: Prof. Dr. Banu DIRI

Imagine a world where every doctor and nurse has an AI buddy, ready to offer a helping hand with everything from talking to patients to diving deep into medical research. That's not a distant dream; it's becoming a reality, thanks to the magic of language models tailored for healthcare. Let's dive into how these specialized tools are changing the game.

Background: The advancement of Large Language Models (language tools) has shown significant promise in various sectors, including healthcare. This project focuses on checking out how well of specialized language tools, namely CLAIR-Short [1], CLAIR-Long [1], Clinical-Longformer [2], Clinical-BigBird [2], and Almanac [3], in comparison to general tools such as ChatGPT, Bing, and Bard, within the healthcare domain. The aim is to show what's possible of tailored language tools in making clinical decisions better, talking to patients, and how efficiently we can do medical research.

Methods: We took a close look at several key studies that dug into how language tools in healthcare. This included a detailed examination of the performance of fine-tuned tools like CLAIR-Short and CLAIR-Long for coming up with answers to patient messages, and the Almanac model, which boosts language tools with retrieval capabilities from carefully selected medical info for advice on guidelines and treatments. We also looked into the performance of Clinical-Longformer and Clinical-BigBird tools for handling long medical texts. Our assessment criteria focused on the tools' ability to generate accurate, complete, and user-preferred responses in clinical settings, along with their safety and reliability in handling

sensitive medical data.

Results: The specialized tools demonstrated better performance in many areas, including factuality, completeness, user preference, and adversarial safety, compared to the more general ones. For instance, CLAIR-Short and CLAIR-Long showed amazing ability in generating clear and helpful answers to patient messages, greatly improving how patients and doctors talk to each other. The Almanac model's search-boosted framework showcased enhanced decision-making support by getting to the latest medical guidelines. Additionally, specialized tools like Clinical-Longformer and Clinical-BigBird effectively managed extended medical documents, illustrating the potential for these language tools in streamlining medical research and literature review processes.

Conclusion: The findings from this project underscore the game-changing potential of language tools in healthcare, particularly when tools are sharpened for specific medical uses. Specialized language tools like CLAIR-Short, CLAIR-Long, Clinical-Longformer, Clinical-BigBird, and Almanac not only offer better accuracy and reliability in processing medical information but also ensure better safety and happier users. These improvements suggest integrating language tools into day-to-day medical work, pointing to a future where AI makes a big difference to healthcare delivery and medical research.

Keywords: Large Language Models, Healthcare, CLAIR Models, Almanac, Clinical-Longformer, Clinical-BigBird, AI in Medicine, Patient Communication, Medical Research.

1
GİRİŞ

1.1 LLM Nedir?

LLM'ler, veya Büyük Dil Modelleri, insan diliyle etkileşim kurabilen ve DDİ teknolojilerini kullanarak devasa veri setlerinden öğrenen yapay zeka sistemleridir. Bu modeller, metinleri anlama, çıkarımlar yapma ve dil bazlı görevleri yerine getirme gibi birçok yeteneğe sahiptir. Aynı zamanda LLM'ler, benzeri görülmemiş özellikleri nedeniyle hem akademide hem de endüstride giderek daha popüler hale geliyor[4]. Son yıllarda, GPT gibi modellerin geliştirilmesiyle, LLM'lerin yetenekleri önemli ölçüde artmıştır.

1.1.1 Son Yıllarda LLM'leri Eğitmek İçin Gerekli Donanım ve Veri Setleri

LLM'lerin eğitimi için gerekli temel kaynaklar arasında büyük ekran kartları ve geniş veri setleri bulunmaktadır. Bu modellerin eğitimi, yüksek işlem gücü gerektiren bir süreçtir ve genellikle çoklu GPU veya hatta TPU gibi özel donanımlar kullanılarak gerçekleştirilir. Eğitim süreci, büyük ve çeşitli veri kümelerinden yararlanarak modellerin geniş bir dil bilgisine ve çeşitli konularda derinlemesine bilgiye sahip olmasını sağlar.

Büyük veri setleri, modellerin dilin karmaşıklığını ve nuanslarını öğrenmesi için çok önemlidir. Bu veri setleri, internetten toplanan milyarlarca kelimeyi içerebilir ve farklı dillerdeki metinler, konuşma dilindeki varyasyonlar ve çeşitli konulardaki bilgiler gibi geniş bir yelpazeyi kapsar. LLM'ler, bu büyük veri setlerini analiz ederek, dilin doğal kullanımını anlama ve taklit etme yeteneklerini geliştirir.

1.1.2 LLM'ler İnsan Dilini Nasıl Anlar?

LLM'ler, insan dilini anlama ve işleme yeteneklerini, metinlerdeki kelime ve ifadeleri matematiksel vektörlere dönüştürerek gerçekleştirir. Bu süreç, genellikle tokenizer yapısı ve özel token'lar (special tokens) kullanılarak yönetilir.

Tokenizerlar, metni daha küçük parçalara ayırırken, özel token'lar ise modele dilin belirli özelliklerini ve yapılarını öğretmek için kullanılır. Örneğin cümlenin bir soru cümlesi olduğunu modelin daha iyi anlaması için "<|soru|>", ya da cümlenin sonunun geldiğini belirtmek için <|endoftext|> gibi token'lar kullanılabilir.

1.1.2.1 Tokenizer Yapısı

Tokenizerlar, metni, modele girebilecek şekilde kelime, kelime parçaları veya semboller gibi daha küçük birimlere (token'lara) ayırır. Bu işlem, modelin dilin yapısal özelliklerini ve kelime arası ilişkileri daha iyi anlamasını sağlar. Örneğin, "Enfotext" kelimesi bir tokenizer tarafından "Enfo", "##text" gibi parçalara ayrılabilir. Bu ayrım, modelin hem "Enfo" hem de "text" anlamlarını öğrenmesine ve daha geniş bir dil bilgisini kavramasına yardımcı olur.

1.1.2.2 Özel Token'lar (Special Tokens)

Özel token'lar, modele metin hakkında ek bilgiler sağlamak için kullanılır. Bunlar, cümlenin başlangıcını ve sonunu belirten [CLS] ve [SEP], metindeki bilinmeyen kelimeleri temsil eden [UNK] gibi token'lardır. Örneğin, [CLS] token'ı, sınıflandırma görevlerinde modelin çıktıyı üreteceği yer olarak işaretlenirken; [SEP] token'ı, iki farklı cümlenin veya metin parçasının birbirinden ayrılmasını sağlar.

Token'lar ve özel token'lar, modelin hem yerel (kelime ve kelime grupları) hem de genel (cümle ve paragraf düzeyinde) dil yapısını anlamasına olanak tanır. Bu, LLM'lerin metni anlamlandırmasında, bağlamı korumasında ve dilin nüanslarını modellemesinde temel bir rol oynar.

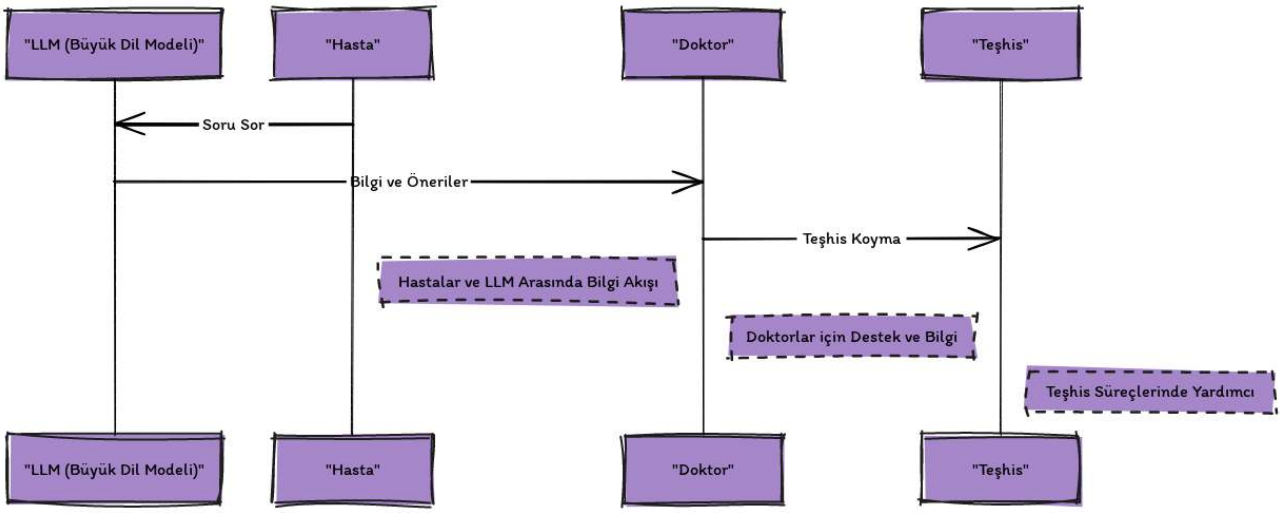
1.2 LLM'lerin Sağlık Sektöründe Kullanımı

Sağlık sektöründe, LLM'lerin kullanımının giderek artması beklenmektedir ve bu modeller gelecekte, doktorlar, sağlık çalışanları ve hastalar tarafından çeşitli şekillerde değerlendirilebilir. Bu teknolojiler ileride, sağlık alanında bir devrim oluşturma potansiyeline sahip olup, teşhis, tedavi ve hasta takibi süreçlerini iyileştirme fırsatları sunar[5]. Bu potansiyeline karşın LLM'ler şu anda neredeyse hiç kullanılmamaktadır. Bunun sebepleri olarak ahlaki kaygılar ve henüz potansiyelin gerçeklenememesi gösterilebilir.

1.2.1 Doktorlar ve Sağlık Çalışanları İçin Kullanım

Doktorlar ve sağlık çalışanları, LLM'leri, teşhis koyma süreçlerini desteklemek ve tedavi planlaması yapmak için kullanabilirler. Örneğin, LLM'ler, hasta semptomlarını analiz ederek potansiyel hastalıklar hakkında ön bilgiler sunabilir ve doktorların daha hızlı ve doğru teşhisler koymasına yardımcı olabilir. Ayrıca, LLM'ler, medikal literatürdeki en güncel bilgileri analiz ederek, belirli bir hastalık için en etkili tedavi yöntemleri hakkında önerilerde bulunabilir.

Örneğin *Şekil 1.1*'deki akışa baktığımızda doktor hastanın semptomlarını LLM'e danışarak daha etkili ve hızlı bir biçimde inceleyip teşhis koyabilir.



Şekil 1.1 Teşhis Koyma Akışı

1.2.2 Hastalar İçin Kullanım

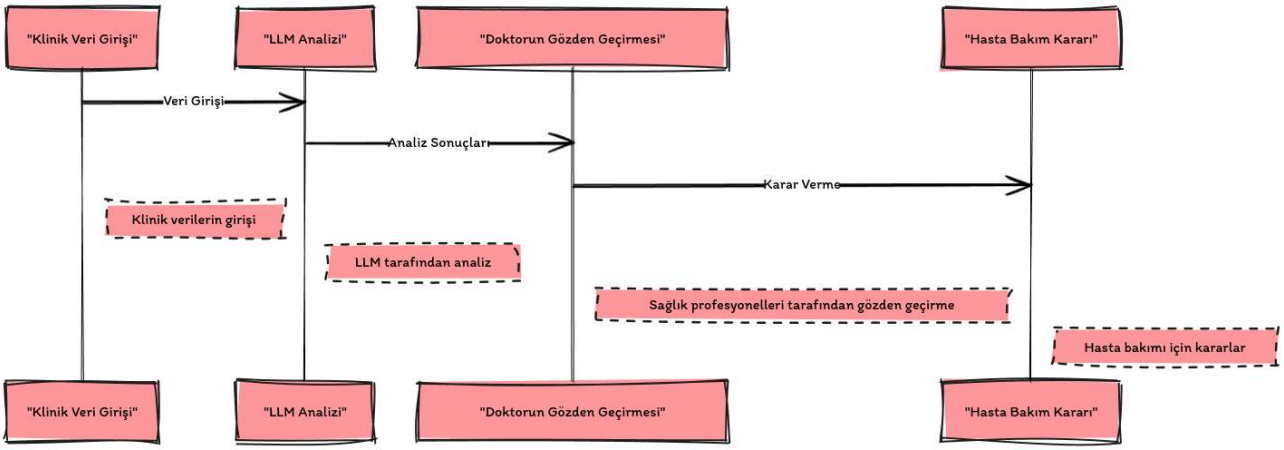
Hastalar, sağlık durumları hakkında bilgi almak ve basit sağlık sorunları için önerilerde bulunmak amacıyla LLM'leri kullanabilirler. Hastalar, semptomlarını bir LLM modeline girerek, olası hastalıklar ve alınabilecek önlemler hakkında bilgi alabilirler. Bu, özellikle doktora ulaşımın zor olduğu durumlarda veya hızlı bilgiye ihtiyaç duyulduğunda faydalı olabilir.

Örneğin *Şekil 1.1*'deki akışa baktığımızda hasta, LLM'e soru sorup hastalığı hakkında daha nokta atışı tavsiye ve bilgilere sahip olabilir. Bu sayede, eğer hastanın haberi olmadığı ve hastalığın erken teşhis edilmesi gerektiği durumlarda LLM'ler hayat kurtarıcı bir rol oynar.

1.2.3 Tahlil ve Görüntüleme Sonuçlarının Değerlendirilmesi

Tahlil sonuçları veya MR görüntüleri gibi medikal veriler, LLM'ler tarafından ön değerlendirme için kullanılabilir. Bu modeller, görüntü işleme ve metin analizi yetenekleri sayesinde, verileri analiz ederek önemli bulguları saptayabilir ve doktorların dikkatini çekmesi gereken alanlara işaret edebilir. Bu, teşhis sürecini hızlandırabilir ve doktorların daha etkili kararlar almasına yardımcı olabilir.

Örneğin *Şekil 1.2*'de gösterilen süreçte, klinik veri girişi ile başlayan analiz ve değerlendirme zinciri, LLM'lerin katkısıyla doktorun gözden geçirmesine doğruluk, hız ve güncellik açısından katkı sağlar. Bu, özellikle hastalığın erken evrede teşhis edilmesi gibi önemli durumlarda, doktorlara daha kesin ve hızlı karar verme imkânı sunarak, bir doktorun aynı zamanda, daha doğru ve fazla teşhis koyabilmesinin önünü açar.



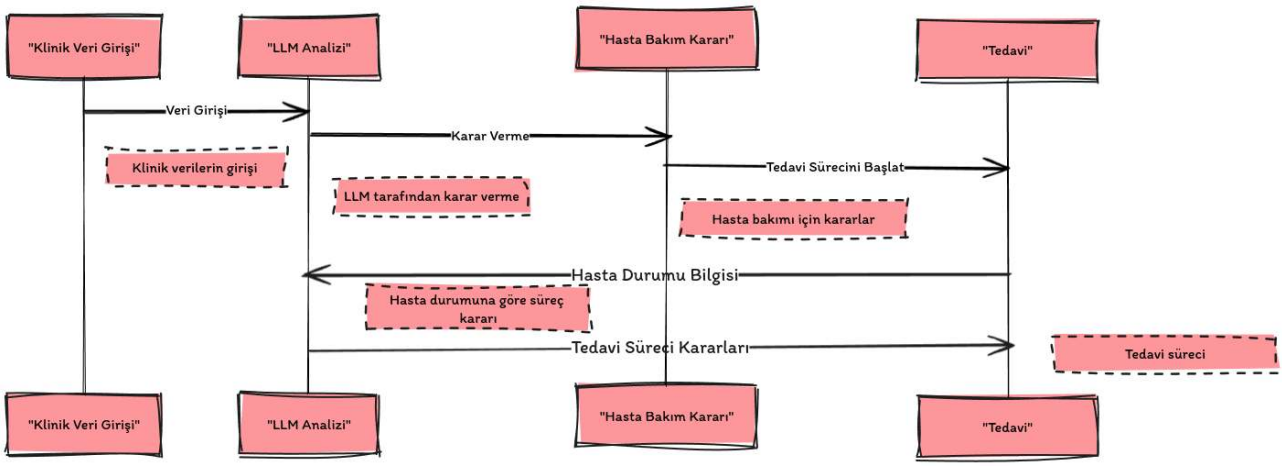
Şekil 1.2 Klinik Veri Akışı

1.2.4 Tedavi Takibi ve Yönetimi

LLM'ler, tedavi süreçlerinin takibi ve yönetiminde de kullanılabilir. Hastalar, tedavi planlarını ve ilaç kullanımını LLM'ler aracılığıyla takip edebilir, yan etkiler veya tedaviye yönelik soruları hakkında bilgi alabilirler. Ayrıca, LLM'ler, hastaların sağlık durumları ve tedaviye uyumları hakkında geri bildirim sağlayarak, daha kişiselleştirilmiş ve etkili bir tedavi sürecinin yönetilmesine olanak tanır.

Örneğin *Şekil 4.1*'de gösterilen süreçte, klinik veri girişi ile başlayan analiz ve değerlendirme zinciri, LLM'lerin karar vermesiyle hasta bakım kararına ve sonrasında da tedavi sürecine kadar uzanır. Sonrasında tedavi uygulamaları sağlık personelleri tarafından gerçekleştirilirken, dönütler LLM'e verilir ve LLM tedavi sürecini de bu dönütlere göre yönetir. Hasta durum bilgisine göre tedaviyi ağırlaştırma, hafifletme ve bitirme kararlarını LLM verebilir. Bu sayede doktorlara

göre erişilebilir, hızlı, ucuz ve konforlu bir tedavi süreci gerçekleşmiş olur. Tabi ki günümüz teknolojisiyle bunların tamamının doktor kontrolünde olması gerekir.



Şekil 1.3 Tedavi Takip Akışı

2
Özelleştirilmiş ve Genel Amaçlı LLM'lerin Karşılaştırılması

2.1 Genel Amaçlı LLM'ler

Genel amaçlı LLM'ler, geniş bir dil bilgisine sahip olup, çeşitli konular ve görevler üzerinde çalışabilir. Bu modeller, büyük veri setlerinden öğrenir ve çok yönlü kullanım imkanı sunar. Örneğin, GPT-4, Bard, Bing gibi modeller, basit medikal soruları yanıtlamaktan, hasta danışmanlığına ve hatta medikal literatür taramasına kadar bir dizi görevi yerine getirebilir.

Ancak, genel amaçlı LLM'lerin sağlık sektörüne özgü terminoloji ve konseptleri anlamada sınırlılıkları vardır. Bu modeller, spesifik medikal bilgilere veya karmaşık klinik senaryolara cevap vermekte zorlanabilir, çünkü eğitim verileri genellikle medikal olmayan metinlerden oluşur. Aynı zamanda bu LLM'leri eğitmek ve çalıştırmak oldukça uzun sürer ve çok maliyetlidir.

2.1.1 Genel Amaçlı Modellerin Eğitimi ve Çalıştırılması

Genel amaçlı LLM'lerin eğitimi ve çalıştırılması, hem maliyetli hem de devasa derecede zaman alıcı bir süreçtir ve bu süreç, geniş kapsamlı veri setleri ve güçlü bilgisayar donanımı gerektirmektedir. Bu modeller, çeşitlilik ve geniş kapsamda bilgi sunabilmek için milyarlarca kelimeyi içeren devasa veri setleri üzerinde eğitilir. Eğitim süreci, dilin çeşitli yönlerini ve nüanslarını öğrenmek için geniş bir veri yelpazesini kapsar. Bu da modeli eğitmek için yüksek performanslı bilgisayar sistemleri ve büyük ölçekli veri işleme kapasitesine ve aynı zamanda yüksek paraya ihtiyaç olduğu anlamına gelir.

2.1.1.1 Donanım Gereksinimleri

Eğitim ve çalıştırma süreçleri, genel amaçlı LLM'lerin geliştirilmesi ve sürekli iyileştirilmesi için çok önemlidir. Bu süreçlerin başarıyla yürütülmesi, güçlü

GPU'lar veya TPU'lar gibi yüksek donanım kaynakları gerektirir. Genel amaçlı modellerin eğitimi, özellikle büyük ve karmaşık modeller için, önemli bir maliyet ve zaman yatırımı anlamına gelir.

2.1.1.2 Donanım Kaynakları ve Maliyet

Güçlü GPU'lar ve TPU'lar, genel amaçlı LLM'lerin eğitim sürecinde paralel işlemleri hızlı bir şekilde gerçekleştirebilir ve böylece eğitim süresini önemli ölçüde kısaltabilir. Ancak, bu yüksek performanslı donanım birimlerinin maliyeti, özellikle çok sayıda birimin eş zamanlı olarak kullanılması gerektiğinde, projenin bütçesine büyük bir yük getirebilir. Bu donanımların enerji tüketimi de göz önünde bulundurulduğunda, operasyonel maliyetler daha da artar.

2.1.1.3 Sürekli Güncelleme ve Geliştirme İhtiyacı

Genel amaçlı LLM'ler, dilin ve bilginin sürekli evrimi nedeniyle düzenli olarak güncellenmeli ve yeniden eğitilmelidir. Bu sürekli güncelleme ve optimizasyon süreci, modeli günümüzde tutmak için gereklidir. Ancak, her güncelleme ve optimizasyon döngüsü, yeni eğitim verilerini toplama, işleme ve modeli yeniden eğitme maliyetleri anlamına gelir. Bu durum, özellikle sürekli geliştirme ve yenilik arayışında olan kurumlar için önemli bir maliyet kalemi oluşturur.

2.1.1.4 Artan Karmaşıklık ve Kaynak İhtiyacı

Modellerin boyutu ve karmaşıklığı arttıkça, eğitim ve çalıştırma için gereken donanım kaynaklarının miktarı da artar. Büyük modeller, daha fazla hafıza ve işlem gücü gerektirir, bu da donanım yatırımlarını daha da önemli hale getirir. Ayrıca, büyük modellerin eğitimi ve çalıştırılması sırasında karşılaşılan teknik zorluklar, uzman personel gereksinimini ve dolayısıyla iş gücü maliyetlerini artırabilir.

2.1.1.5 Ölçeklendirme Zorlukları

Genel amaçlı LLM'lerin eğitimi ve sürekli geliştirilmesi süreçlerinin ölçeklendirilmesi, teknik ve mali zorluklar içerir. Modelin boyutu arttıkça, eğitim için gereken veri miktarı ve işlem gücü de artar. Bu, büyük ölçekli eğitim operasyonlarının yönetilmesini karmaşıklaştırır ve etkili bir ölçeklendirme stratejisi gerektirir. Ölçeklendirme çabaları, veri depolama ve işleme altyapısının yanı sıra, donanım ve yazılım optimizasyonlarını da kapsamalıdır.

2.1.1.6 Güncel Bilgilere Uyumu

Genel amaçlı LLM'ler, sağlık sektörü gibi hızla gelişen alanlarda güncel bilgileri içerecek şekilde sürekli güncellenmelidir. Ancak, bu sürekli güncelleme işlemi, hem zaman hem de maliyet açısından zorlayıcı olabilir. Modelin eğitim verilerini güncel tutmak ve modeli en son bilgilerle yeniden eğitmek, devasa boyutlarda maliyete sahiptir.

2.2 Özelleştirilmiş LLM'ler

Özelleştirilmiş LLM'ler, belirli bir alana veya göreve özel olarak tasarlanmış ve eğitilmiş modellerdir. Bu özelleştirme, özellikle sağlık sektörü gibi karmaşık ve spesifik terminolojiyi içeren alanlarda, modellerin hız ve doğruluğunu önemli ölçüde artırır. Bunun yanında modelin boyutunu ve modelin çalıştırılması için gereken donanım kaynağını da önemli ölçüde azaltır. Sağlık sektöründe kullanılan özelleştirilmiş modeller, klinik notlar, medikal araştırma makaleleri, hasta raporları ve diğer alana özgü veri setleriyle eğitilir. Bu süreç, modellere medikal terminoloji ve klinik senaryolar hakkında çok ayrıntılı bir cevap verebilme yeteneği kazandırır ve bu da onların teşhis koyma, tedavi önerileri ve medikal literatür analizi gibi spesifik görevlerde daha doğru, hızlı ve güvenilir sonuçlar sunmasını sağlar.

2.2.1 Hız ve Verimlilik

Özelleştirilmiş LLM'ler, eğitildikleri spesifik görevlerde genel amaçlı modellere göre daha hızlı ve verimli çalışır. Bu modeller, belirli bir alandaki veri setlerine odaklanarak eğitildiklerinden, gereksiz bilgileri süzme ve alakasız içeriği işleme konusunda daha az zaman harcarlar. Örneğin, Clinical-Longformer[2] ve Clinical-BigBird[2] gibi modeller, uzun klinik metinleri hızlı bir şekilde işleyebilir ve önemli bilgileri etkili bir şekilde çıkarabilir, bu da sağlık çalışanlarının karar verme süreçlerini hızlandırır.

2.2.2 Fiyat/Performans

Özelleştirilmiş LLM'lerin geliştirilmesi ve eğitilmesi, genel amaçlı modellere kıyasla daha verimli olabilir. Bu modeller, belirli bir görev veya alan için optimize edildiğinden, geniş kapsamlı ve binlerce konuyu içeren veri setlerinden elde edilen bilgiyi işlemek zorunda kalmazlar. Daha küçük ve özel verilerden elde edilen bilgi için gereken kaynaklar daha azdır. Ayrıca, özelleştirilmiş modellerin sürekli olarak büyük ölçekli veri setleri ile eğitilmesi gerekmez, daha küçük veri setleriyle daha ucuz ve hızlı bir şekilde eğitilebilirler. Bu da uzun vadede maliyet tasarrufu sağlar.

2.2.3 Doğruluk ve Kesinlik

Özelleştirilmiş LLM'ler, eğitimlerinde kullanılan özel veri setleri sayesinde, teşhis koyma, tedavi planlama ve medikal literatür inceleme gibi görevlerde yüksek doğruluk oranları ve kesinlik sunar. Bu modeller, karmaşık medikal terimleri ve klinik durumları anlama konusunda özel olarak geliştirildiklerinden, genel amaçlı modellere göre daha güvenilir sonuçlar üretebilirler. Özelleştirilmiş modeller, klinik karar destek sistemlerinin geliştirilmesinde ve hasta bakımının iyileştirilmesinde efektif ve önemli bir rol oynar.

2.2.4 Uygulama Esnekliği

Özelleştirilmiş LLM'ler, sağlık sektöründeki çeşitli ihtiyaçlara göre özelleştirilebilir ve uyarlanabilir. Bu, farklı medikal disiplinlerde veya spesifik sağlık hizmeti görevlerinde modellerin etkili bir şekilde kullanılmasını sağlar. Örneğin, bir model, kanser teşhisi koyma üzerine özelleştirilebilirken, başka bir model, kalp hastalıklarının yönetimi için optimize edilebilir. Bu esneklik, sağlık hizmetlerinin geniş bir yelpazede iyileştirilmesine olanak tanır.

2.3 Almanac ile Genel Modellerin Kıyaslanması

Bu bölümde, özel amaçlı bir dil modeli olan Almanac [3]'ın, genel amaçlı dil modelleri Bard, Bing ve GPT-4 ile performans karşılaştırması yapılacaktır. Sağlık sektörüne özel olarak geliştirilmiş Almanac modeli, bu alanda daha etkin ve doğru bilgi sağlama potansiyeline sahiptir. Genel amaçlı modellerse daha geniş bir uygulama alanına sahip olmakla birlikte, bu modellerin sağlık sektörüne özgü gereklilikleri karşılamada bazı sıkıntılar yaşayabileceği düşünülmektedir.

Karşılaştırmada, her bir modelin sağlık sektöründe karşılaşılabilecek gerçek dünya sorunlarına nasıl cevap verdiği, gerçeklik (facutality), tamamlılık (completeness) ve kullanıcı tercihi (preference) gibi metrikler üzerinden değerlendirilecektir. Ayrıca modellerin doğru bilgiye ne kadar atıf yaptığı ve aynı zamanda yanıltıcı komutlara karşı duyarlılığı da incelenecektir.

Bu kıyaslama, Almanac'ın uzmanlık alanı olan sağlık sektöründe, genel amaçlı modellere göre sunduğu avantajları ve sağlık sektöründe, genel amaçlı modellerin sahip olduğu sıkıntılar detaylı bir şekilde ele alınacaktır. Karşılaştırmanın amacı, her modelin güçlü ve zayıf yönlerini ortaya koymak ve özellikle sağlık sektörü gibi kritik alanlarda kullanım için yol gösterici bilgiler sağlamaktır.

2.3.1 Gerçeklik - Tamamlılık - Kullanıcı Tercihi

Özellikle sağlık sektörüne özgü LLM'lerin performansını değerlendirmek için çeşitli metrikler ön plana çıkar. Bu metrikler arasında gerçeklik (facutality), tamamlılık (completeness) ve kullanıcı tercihi (preference) bulunmaktadır. Gerçeklik, bir modelin ürettiği bilginin doğruluğunu; tamamlılık, verilen bilginin konuyu ne derece kapsadığını; kullanıcı tercihi ise kullanıcıların modelin yanıtlarını ne kadar yararlı bulduğunu gösterir.

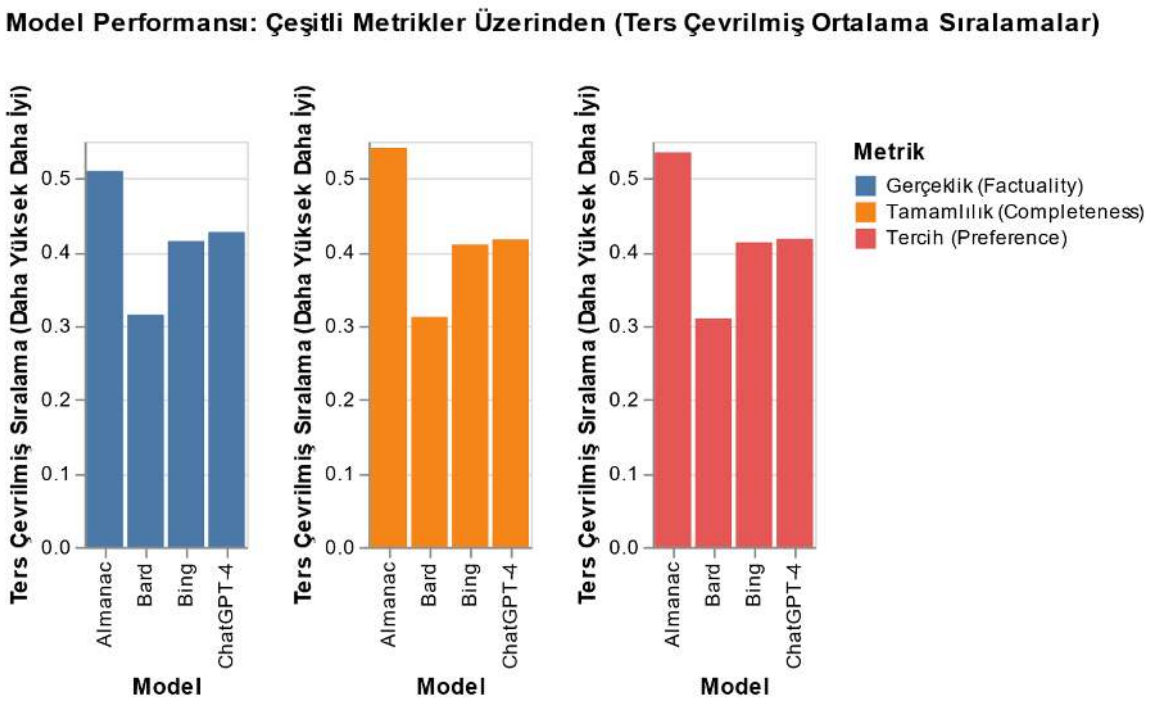
Şekil 2.1 model performansını üç önemli metrik üzerinden kıyaslamaktadır. Grafiğe baktığımızda gerçeklik değerleri açısından Almanac birinci, ChatGPT-4 ikinci, Bing üçüncü, Bard ise dördüncü olmuştur. Bing ve ChatGPT-4 arasındaki fark önemsenecek derecede fazla değildir. Aynı zamanda Bard açık ara en başarısız model olmuştur. Almanac da bir o kadar çok yüksek başarım göstermiştir. Tamamlılık değerleri açısından Almanac birinci, ChatGPT-4 ikinci, Bing üçüncü, Bard ise dördüncü olmuştur. Yine Bing ve ChatGPT-4 arasındaki fark önemsenmeyecek derecede azdır. Aynı zamanda Bard açık ara en başarısız model olmuştur. Almanac bu sefer farkı biraz daha açmıştır. Tercih değerleri açısından Almanac birinci, ChatGPT-4 ikinci, Bing üçüncü, Bard ise dördüncü olmuştur. Yine Bing ve ChatGPT-4 arasındaki fark önemsenmeyecek derecede azdır. Aynı zamanda Bard açık ara en başarısız model olmuştur. Almanac da bir o kadar çok yüksek başarım göstermiştir. Bu veriler, Almanac'ın gerçeklik ve tamamlılıkta diğer genel amaçlı modellerden daha yüksek bir performans sergilediğini ve açık ara bu alandaki en iyi model olduğunu göstermektedir.

2.3.2 Doğru Atıf - Kötü Prompt

LLM'lerin değerlendirilmesinde önemli bir diğer metrik, modelin doğru atıfları ne derece başarılı bir şekilde gerçekleştirdiği ve kötü niyetli ya da yanıltıcı kullanım girişimlerine karşı duyarlılığıdır.

Kötü niyetli promptlara karşı duyarlılık, bir dil modelinin kötü niyetli veya yanıltıcı kullanım girişimlerine karşı verdiği tepkiyi ve bu tür girişimleri nasıl yönettiğini ifade eder. Kötü niyetli promptlar, bilinçli olarak yanıltıcı, zararlı, manipülatif ya da toksik yanıtlar üretmeye teşvik eden girdilerdir. Bir modelin kötü niyetli promptlara karşı duyarlılığı yüksekse, bu tür kötü amaçlı girdilere karşı direnç göstermesi ve güvenli, etik, doğru yanıtlar üretmesi beklenir.

Doğru atıf yüzdesi ise, bir modelin referans verirken kaynakları ne kadar doğru bir şekilde belirttiğini gösterir. Bu, özellikle bilgi tabanlı yanıtlarda, modelin sunduğu bilgilerin güvenilir kaynaklara dayandığını ve doğru bir şekilde atıfta



Şekil 2.1 Almanac Genel Metrik Kıyaslaması [3]

bulunulduğunu gösteren bir metriktir. Yüksek bir doğru atıf yüzdesi, modelin bilgiyi doğru kaynaklardan aldığını ve bu bilgiyi doğru bir şekilde kullanıcıya sunduğunu gösterir ki bu, özellikle akademik araştırma veya sağlık bilgisi gibi doğruluk gerektiren alanlarda çok önemlidir.

Şekil 2.2'de sunulan grafik, özellikle doğru atıflar ve kötü niyetli promptlara duyarlılık olmak üzere iki özel metrik üzerinden Almanac'ın Bard, ChatGPT-4 ve Bing modelleriyle kıyaslamasını göstermektedir. Bu kıyaslama, modelin bilgiyi ne kadar doğru şekilde referanslandığını ve kötü niyetli promptlara karşı ne derece dayanıklı olduğunu ölçer.

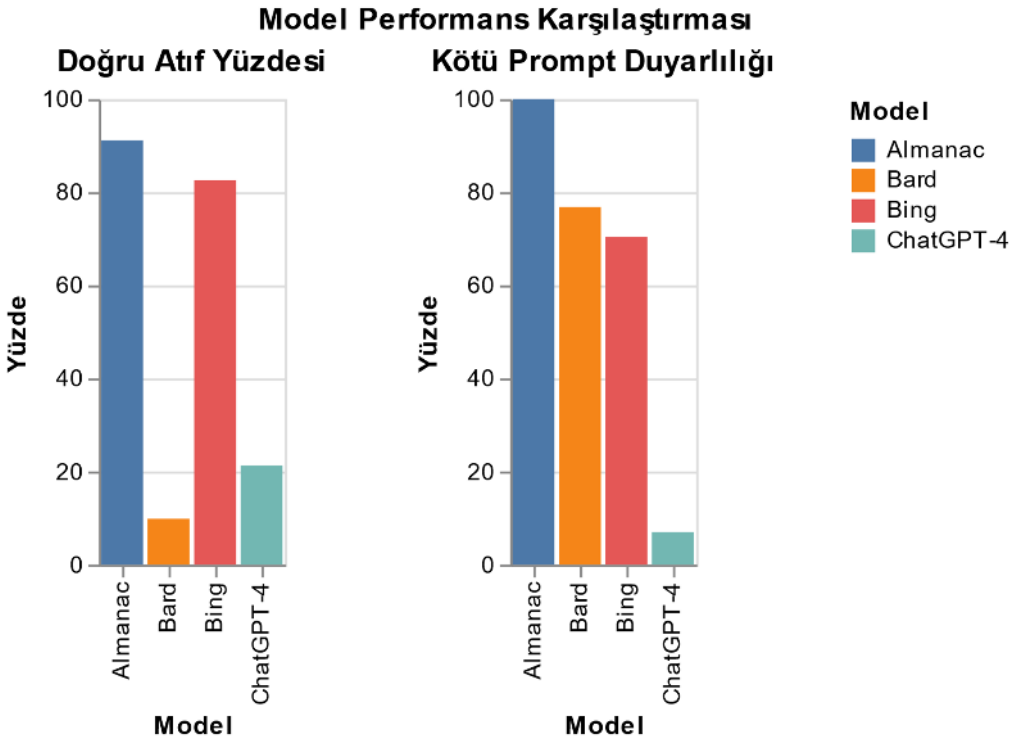
Doğru atıf yüzdesine bakıldığında, Almanac %91.11 ile en yüksek başarıyı gösterirken, bu oran Bard için %9.84, ChatGPT-4 için %21.27 ve Bing için %82.54 olarak belirlenmiştir. Bu sonuçlar, Almanac'ın referans vermede diğer genel amaçlı modellere kıyasla daha yüksek bir doğruluk sergilediğini ortaya koymaktadır. Aynı zamanda Bard aşırı düşük bir başarım gösterirken, ChatGPT-4 önceki metriklere göre bu sefer Bing'e göre devasa düşük başarım göstermiştir. Bing'in başarımı ne kadar yüksek olsa da Almanac'a göre kayda değer seviyede geride kalmıştır.

Adversarial prompt duyarlılığına gelince, Almanac %100 ile tüm kötü niyetli girişimlere karşı doğru yanıt vermiş, bu alanda kusursuz bir performans sergilemiştir. ChatGPT-4'ün %7 ile oldukça düşük bir duyarlılık gösterdiği, Bard ve Bing'in ise sırasıyla %76.80 ve %70.40 ile daha yüksek duyarlılık

sergilediği gözlenmektedir. Aynı zamanda bu sefer en başarısız model Bard değil ChatGPT-4'dür. Bard şaşırtıcı bir şekilde önceki metriklere göre bu sefer Bing'e göre yüksek bir başarım göstermiştir. Ama yine de Bard'ın başarımı ne kadar yüksek olsa da Almanac'a göre kayda değer seviyede geride kalmıştır.

Bu sonuçlar, Almanac'ın sadece doğru bilgi sağlamada değil, aynı zamanda potansiyel olarak yanıltıcı veya zararlı girdilere karşı dayanıklılıkta da öne çıktığını göstermektedir. Sağlık hizmetleri gibi kritik ve doğruluk gerektiren alanlarda, Almanac gibi özelleştirilmiş modellerin kullanılması, verilen bilginin doğruluğunu ve güvenilirliğini artırabilir. Bu, özelleştirilmiş modellerin sağlık hizmetlerinde kullanımının önemini ve değerini bir kez daha göstermektedir.

Aynı zamanda, özel eğitilmiş modellerin genel amaçlı modellere göre daha kararlı bir başarım grafiği çiziği de ortadadır. Çünkü örneğin ChatGPT-4 gerçeklik, tamamlılık ve tercih gibi metriklerde en başarılı ikinci model olsa da (başarım oranı da hayli yüksek), doğru atıf yüzdesinde devasa başarım farkıyla üçüncü, kötü promt duyarlılığındaysa yine devasa bir farkla dördüncü olmmuştur. Yine benzer şekilde Bard kötü promt duyarlılığı metriği hariç tüm metriklerde sonuncu sıradayken, kötü prompt duyarlılığında ikinci olmuştur. Bu da genel amaçlı modellerin metriklerarası, başarımlarının kararsız olduğunu açıkça göstermektedir.



Şekil 2.2 Almanac Özel Metrik Kıyaslaması [3]

3
Sağlık Alanında LLM'lerin Değerlendirilmesi ve Uygulamaları

3.1 LLM'lerin Performansının Değerlendirilmesi

LLM'lerin performansının değerlendirilmesi, yapay zeka ve dil işleme teknolojilerinin sağlık alanında uygulanabilirliğini ve güvenilirliğini ölçmek için kritik bir süreçtir. Almanac gibi modeller, genel amaçlı LLM'lere kıyasla tıbbi veritabanlarına ve güncel medikal bilgilere erişim sağlayarak, sağlık profesyonellerinin karşılaştığı karmaşık sorulara yanıt verme potansiyeline sahiptir. Bu performans değerlendirmesi için ClinicalQA veri seti kullanılmıştır.

3.1.1 ClinicalQA veri seti

ClinicalQA veri seti, LLM'lerin tıbbi bilgiyi anlama ve uygulama yeteneğini değerlendirmek için özel olarak tasarlanmış bir veri setidir. Bu veri seti, klinik ortamda sık karşılaşılan senaryoları içeren ve doktorlar tarafından sıklıkla sorulan soruları içerir. Örneğin, semptomların teşhise dönüştürülmesi, tedavi seçeneklerinin değerlendirilmesi ve ilaç etkileşimlerinin belirlenmesi gibi.

ClinicalQA, genel sağlık bilgisinin yanı sıra özel tıbbi uzmanlık gerektiren konuları kapsar ve modellerin bu konulardaki yanıtlarını gerçek dünya standartlarına göre ölçer. Bu veri seti, modellerin klinik bilgiyi nasıl işlediğini, karar verme süreçlerini nasıl destekleyebileceğini ve sonuçta hasta bakımını nasıl iyileştirebileceğini anlamak için kullanılır.

ClinicalQA, LLM'lerin doğruluğunu, anlayış derinliğini ve klinik bilgi üzerine kurulu çıkarım yapma becerilerini test eden kaliteli bir veri setidir. Bu, modellerin medikal alandaki kullanımının potansiyelini ve sınırlamalarını değerlendirmek için önemli bir veridir. Modelin bu veri setinde sergilediği performans, sağlık hizmetleri uygulamalarında potansiyel kullanımının bir göstergesi anlamına gelir.

3.1.2 Almanac modelinin ClinicalQA görevlerindeki başarısı

Almanac, sağlık alanında özel olarak geliştirilmiş bir dil modeli olarak, ClinicalQA veri setindeki verilerde önemli bir başarım göstermiştir. Bu veri seti, 314 açık uçlu klinik sorudan oluşmakta ve çeşitli tıbbi uzmanlık alanlarını içermektedir. Veri, hekimlerin iş hayatlarında karşılaştıkları gerçek dünya verilerinden alınmış olup, tedavi yönergelerinden klinik hesaplamalara kadar geniş bir alanı kapsamaktadır. Almanac modelinin bu verilerde sergilediği performans, gerçeklik, tamamlılık ve kullanıcı tercihi gibi önemli metrikler açısından diğer genel amaçlı modellere göre önemli bir üstünlük sağlamıştır.

Bu başarının altında yatan temel faktörler, Almanac'ın sağlık hizmetleri alanına özgü derinlemesine eğitimi ve bu alandaki spesifik ihtiyaçlara göre optimize edilmiş olmasıdır. Model, klinik notlar, medikal araştırma makaleleri, hasta raporları ve diğer sağlık hizmetleriyle ilgili veri kaynaklarından elde edilen zengin ve çeşitli verilerle beslenmiştir. Bu, Almanac'ın tıbbi terminoloji, prosedürler ve tedavi süreçleri gibi konularda ayrıntılı bilgiye sahip hale getirmiştir.

Almanac'ın ClinicalQA veri setindeki etkileyici performansı, LLM'lerin sağlık alanında karşılaşılan zorlukları nasıl çözebileceğine dair önemli bir kaynak teşkil etmektedir. Model, tedavi yönergelerini doğru bir şekilde yorumlayabilme, karmaşık klinik hesaplamaları yapabilme ve hekimlerin karşılaştığı gerçek dünya sorunlarına pratik çözümler sunabilme kapasitesine sahiptir. Bu özelliği, Almanac'ı sağlık hizmetleri alanında klinik karar destek aracı olarak kullanma potansiyelini gözler önüne sermektedir.

Ayrıca, Almanac'ın kullanıcılar tarafından tercih edilmesindeki bir diğer önemli faktör, modelin kullanıcıların sorduğu sorulara doğru, anlaşılır ve kullanışlı yanıtlar sunabilmesidir. Bu, özellikle klinik uygulamalarda ve hasta bakımında bilgiye dayalı kararlar alınmasında önemli bir faktördür. Almanac'ın sağlık sektörüne özel olarak geliştirilmesi, modelin bu alandaki özel istekleri karşılayacak şekilde tasarlanmış olması, sağlık çalışanları ve hasta bakımı açısından gayet önemlidir.

3.1.3 Kötü niyetli kullanım güvenliği ve kullanıcı tercihleri

Almanac modeli, kötü niyetli veya yanıltıcı kullanım girişimlerine karşı dayanıklılık konusunda, diğer modellere kıyasla kayda değer derecede öndedir. Bu tür adversarial güvenlik, LLM'lerin sağlık sektörü gibi kritik bilgi gerektiren alanlarda güvenli bir şekilde kullanılabilmesi için büyük önem taşır. Kötü niyetli girişimler, modelin yanlış yönlendirilmesine, yanıltıcı bilgilerin üretilmesine ve hatta özel verilerin sızdırılmasına yol açabilir, bu da ciddi etik ve güvenlik

sorunlarına neden olabilir.

Almanac'ın adversarial senaryolarda %100 doğru cevap verme başarısı, modelin bu tür risklere karşı maksimum derecede dayanıklı olduğunu ve sağlık hizmetleri alanında güvenilir bir araç olarak kullanılabileceğini göstermektedir. Bu başarı, modelin geliştirilme sürecindeki dikkatli tasarım, kapsamlı testler ve etik standartlara uygunluk gibi faktörlere bağlanabilir. Almanac'ın yüksek düzeyde adversarial güvenliği, sağlık personelleri ve hastalar arasında güven oluştururken, yanıltıcı bilgi ve veri sızıntısı risklerini minimize etme konusunda gayet önemlidir.

Kullanıcı tercihleri açısından, Almanac'ın ürettiği çıktıların kullanıcıların ihtiyaçlarına ve tercihlerine daha uygun bulunması, modelin sağlık sektörüne özgü karmaşık tıbbi kavramları anlayabilme ve etkili bir şekilde iletebilme kapasitesini vurgular. Kullanıcıların modelin çıktılarını tercih etmeleri, Almanac'ın kullanım kolaylığı, anlaşılırlığı ve bilginin doğruluğu gibi alanlarda diğer modellere göre avantaj sağladığını gösterir. Bu, özellikle hasta bakımı ve klinik karar verme süreçleri gibi kritik alanlarda, doğru ve zamanında bilgilere erişimin önemini ortaya koyar.

3.2 Uzun Klinik Metinler için LLM'ler

Sağlık sektöründe kullanılan elektronik sağlık kayıtları, klinik notlar ve araştırma makaleleri gibi uzun metinler, genellikle karmaşık ve uzun tıbbi bilgiler içerir. Bu metinlerin etkili bir şekilde işlenmesi ve analiz edilmesi, hasta bakımının iyileştirilmesi, hızlı ve doğru teşhislerin konulması ve tedavi yöntemlerinin geliştirilmesi için gayet önemlidir. Uzun metinleri işlemek için tasarlanmış LLM'ler, bu alanda önemli avantajlar sunar.

3.2.1 Clinical-Longformer ve Clinical-BigBird modellerinin avantajları

Clinical-Longformer ve Clinical-BigBird, uzun klinik metinlerin işlenmesi için özel olarak tasarlanmış LLM'lerdir. Bu modeller, standart dönüşüm tabanlı modellerin sınırlamalarını aşarak, uzun metinleri daha etkili bir şekilde işleme yeteneğine sahiptirler.

Clinical-Longformer, uzun dökümanları işleyebilmek için tasarlanmış bir modeldir. Dikkat mekanizmasını optimize ederek, çok uzun metinlerde bile önemli bilgileri tespit edebilir ve anlamlandırabilir. Bu özellik, klinik notlar ve hasta raporları gibi uzun metinlerin analiz edilmesinde, önemli tıbbi bilgilerin hızla çıkarılmasını sağlar.

Clinical-BigBird ise, Clinical-Longformer'ın daha kapsamlı halidir diyebiliriz. Daha da geniş bir dikkat mekanizması ile Clinical-Longformer'ın kapsamını genişletir. Bu model, metin içerisindeki geniş bağlamı daha iyi anlayarak, daha karmaşık klinik senaryolarda bile detaylı bilgi çıkarımı yapabilir. Ayrıca, Clinical-BigBird, metinler arasındaki ilişkileri daha iyi anlama ve çeşitli klinik belgeler arasında bağlantılar kurma kapasitesine sahiptir.

Her iki model de, uzun metinlerdeki önemli bilgileri anlama ve çıkarım yapma konusunda gelişmiş kabiliyetlere haizdir. Sağlık hizmetlerinde karar destek sistemlerinin geliştirilmesine önemli katkılar sunabilir. Bu modeller sayesinde, sağlık çalışanları, kapsamlı klinik verileri daha hızlı ve etkili bir şekilde analiz edebilir, bu da tedavi süreçlerini hızlandırır ve hasta sonuçlarını iyileştirir.

3.2.2 Uzun metin işlemede LLM'lerin önemi

Uzun metin işlemede LLM'lerin kullanımı, sağlık hizmetlerinde bilgiye dayalı karar verme süreçlerini önemli ölçüde iyileştirebilir. Bu modeller, büyük miktardaki klinik veriyi ayrıntılı bir şekilde işleyerek, teşhis, tedavi ve hasta yönetimi gibi önemli alanlarda ayrıntılı bilgi sağlarlar. Uzun metin işleme kabiliyetine sahip LLM'ler, klinik araştırmaların hızlandırılmasına, tedavi protokollerinin geliştirilmesine ve hasta bakım kalitesinin artırılmasına katkıda bulunur.

Ayrıca, bu modeller, sağlık çalışanlarının önemli ölçüde zamandan tasarruf etmelerine olanak tanır. Uzun klinik metinlerin manuel olarak incelenmesi zaman alıcı ve zor olabilirken, LLM'ler bu süreci otomatize ederek, sağlık çalışanlarının ya da akademisyenlerin vakitlerini araştırmayla heba etmemelerine imkan verir. Bu, özellikle büyük veri setlerinin hızla arttığı sağlık sektöründe, verimliliği ve etkinliği artırmada önemli bir avantaj sağlar.

3.3 Clinical-Longformer ve Clinical-BigBird Modellerinin Başarımları

Performans ölçümünde i2b2 ve emrQA veri setlerinin varyasyonları kullanılmıştır.

3.3.1 i2b2

i2b2 veri seti, klinik araştırmalar ve sağlık hizmetleri bilgi işlemi için tasarlanmış zengin bir kaynaktır. i2b2 projesi, biyolojik ve tıbbi bilgilerin entegrasyonunu kolaylaştırmayı amaçlar ve bu süreçte, hastane kayıtları, klinik notlar ve diğer sağlık verilerinden oluşan geniş bir veri havuzunu kullanır. Performans ölçümünde

kullanılan i2b2 veri setlerinin varyasyonları, genellikle hastalık teşhisi, hasta yönetimi ve tedavi sonuçları gibi spesifik klinik görevlere odaklanır. Bu veri setleri, LLM'lerin tıbbi metin işleme ve anlayışı, hastalık teşhisi koyma, tedavi önerileri sunma ve klinik karar destek sistemleri geliştirme gibi alanlardaki performansını değerlendirmek için kullanılır. i2b2 veri setleri, gerçek dünya klinik senaryolarını yansıttığı için, modelin pratik klinik uygulamalarda nasıl performans gösterdiğine dair değerli bilgiler sunar. Buradaki testte veri setinin 2006 - 2010 - 2012 - 2014 varyasyonları kullanılmıştır.

3.3.2 emrQA

emrQA veri seti, Elektronik Sağlık Kayıtları (EHR) üzerine kurulu, doğal dil işleme (DDİ) ve bilgi çıkarma için tasarlanmış bir soru-cevap platformudur. Bu platform, sağlık hizmetlerindeki elektronik kayıtları kullanarak DDİ modellerinin bu kayıtlardaki bilgileri nasıl anladığını ve çıkardığını değerlendirmek üzere geliştirilmiştir. emrQA'nın performans ölçümleri, modellerin hasta bilgilerini anlama, klinik soruları yanıtlama ve sağlık hizmetlerinde karşılaşılan özgün sorunlara çözümler sunma kapasitesini test eder. Bu testlerde kullanılan "medication", "relation" ve "heart disease" gibi varyasyonlar, emrQA veri setinin kapsamlılığını ve LLM'lerin sağlık bilgisi işleme ve çıkarım yapma yeteneklerini derinlemesine değerlendirmenin önemini vurgular.

- **Medication Varyasyonu:** Medication (ilaç) varyasyonu, elektronik sağlık kayıtlarındaki ilaç bilgileriyle ilgili sorulara odaklanır. Bu varyasyon, LLM'lerin ilaç adları, dozajları, kullanım zamanları ve yan etkileri gibi ilaçla ilgili karmaşık detayları nasıl anladığını ve bu bilgileri nasıl kullanarak ilgili soruları yanıtladığını değerlendirir. Medication varyasyonu, özellikle ilaç yönetimi ve hasta güvenliği açısından kritik öneme sahiptir.

- **Relation Varyasyonu:** Relation (ilişki) varyasyonu, hastalıklar, semptomlar, test sonuçları ve tedaviler arasındaki ilişkileri keşfetmeye yöneliktir. Bu varyasyon, LLM'lerin elektronik sağlık kayıtlarındaki çeşitli sağlık durumları ve müdahaleleri arasındaki bağlantıları nasıl kurduğunu ve bu bağlantıları anlamlandırdığını test eder. Relation varyasyonu, LLM'lerin klinik karar destek sistemlerindeki potansiyel kullanımlarını ve hasta bakımında bilgiye dayalı karar vermeyi destekleme yeteneğini ortaya koyar.

- **Heart Disease Varyasyonu:** Heart Disease (kalp hastalığı) varyasyonu, kalp hastalıklarıyla ilgili sorulara ve senaryolara odaklanır. Bu varyasyon, LLM'lerin kalp hastalıklarının teşhisi, tedavisi ve yönetimi hakkında bilgiyi

nasıl işlediğini ve ilgili klinik soruları nasıl yanıtladığını değerlendirir. Heart Disease varyasyonu, LLM'lerin spesifik tıbbi alanlardaki derinlemesine bilgiyi nasıl kullanabildiğini ve bu bilgileri klinik uygulamalarda nasıl etkili bir şekilde iletebildiğini gösterir.

Şekil 2.1 model performansını üç önemli metrik üzerinden kıyaslamaktadır. Grafiğe baktığımızda gerçeklik değerleri açısından Almanac birinci, ChatGPT-4 ikinci, Bing üçüncü, Bard ise dördüncü olmuştur. Bing ve ChatGPT-4 arasındaki fark önemsenecek derecede fazla değildir. Aynı zamanda Bard açık ara en başarısız model olmuştur. Almanac da bir o kadar çok yüksek başarım göstermiştir. Tamamlılık değerleri açısından Almanac birinci, ChatGPT-4 ikinci, Bing üçüncü, Bard ise dördüncü olmuştur. Yine Bing ve ChatGPT-4 arasındaki fark önemsenmeyecek derecede azdır. Aynı zamanda Bard açık ara en başarısız model olmuştur. Almanac bu sefer farkı biraz daha açmıştır. Tercih değerleri açısından Almanac birinci, ChatGPT-4 ikinci, Bing üçüncü, Bard ise dördüncü olmuştur. Yine Bing ve ChatGPT-4 arasındaki fark önemsenmeyecek derecede azdır. Aynı zamanda Bard açık ara en başarısız model olmuştur. Almanac da bir o kadar çok yüksek başarım göstermiştir. Bu veriler, Almanac'ın gerçeklik ve tamamlılıkta diğer genel amaçlı modellerden daha yüksek bir performans sergilediğini ve açık ara bu alandaki en iyi model olduğunu göstermektedir.

3.4 Modellerin i2b2 ile Test Edilmesi

Şekil 3.1'de modellerin performansınıb F-1 Skor metriği üzerinden i2b2 veri setinin varyasyonlarıyla kıyaslanmış halini görüyoruz. Her grafiği ayrı ayrı inceleyelim.

3.4.1 i2b2 2006 Veri Setinde Model Performansı

Şekil 3.1'de i2b2 2006 veri setinde, BERT modeli %93.9 F1 skoru ile diğer modellere göre düşük bir başarım sergilemiştir. BioBERT (%94.8 F1) modeli ClinicalBERT (%95.1 F1) modeli tarafından hafif bir marjla geride bırakılmıştır. RoBERTa modeli %95.6 F1 skoru ile daha iyi bir performans gösterirken, Clinical-Longformer %97.4 F1 skoru ile bu grupta en iyi performansı sergilemiştir. Clinical-BigBird ise %96.7 F1 skoru ile başarı sıralamasında ikinci olmuştur.

3.4.2 i2b2 2010 Veri Setinde Model Performansı

Şekil 3.1'de i2b2 2010 veri setinde, BERT modelinin performansı %83.5 F1 skoru ile önceki veri setine kıyasla hayli daha düşüktür, ki bu özel eğitim gerektiren sağlık

alanında oldukça yetersiz kalabilir. BioBERT modeli %86.5 F1 skoru ile daha iyi bir sonuç elde etmiş, ClinicalBERT hemen ardından %86.1 F1 skoru ile yer almıştır. RoBERTa modeli %85.1 F1 ile bu veri setinde beklenenden biraz daha düşük bir performans gösterirken, Clinical-Longformer (%88.7 F1) ve Clinical-BigBird (%87.2 F1) modelleri, bu görev için özel olarak eğitilmiş olduklarını ve alana özgü veri setlerinde önemli bir performans avantajı yakalamışlardır. Görüldüğü üzre tüm modellerin başarımı bu veri setinde, diğer veri setlerine göre daha yüksektir.

3.4.3 i2b2 2012 Veri Setinde Model Performansı

Şekil 3.1'de i2b2 2012 veri setinde, BERT modeli yine %75.9 F1 skoru ile en düşük performansı sergilemiş, bu kez RoBERTa %76.7 F1 skoru ile ona yakın bir performans göstermiştir. Bu sıralamada BioBERT ve ClinicalBERT sırasıyla %78.9 F1 ve %77.3 F1 skorları ile biraz daha yüksek başarı elde etmişlerdir. Yine bu sette de Clinical-Longformer %80.0 F1 skoruyla en iyi performansı sergileyen model olurken, Clinical-BigBird %78.7 F1 skoruyla onu izlemiştir. Bu veriler, sağlık alanında uzmanlaşmış modellerin, genel modellere kıyasla daha karmaşık ve spesifik veri setlerinde daha iyi performans gösterme eğiliminde olduklarını göstermektedir. Görüldüğü üzre tüm modellerin başarımı bu veri setinde, diğer veri setlerine göre düşük kalmıştır.

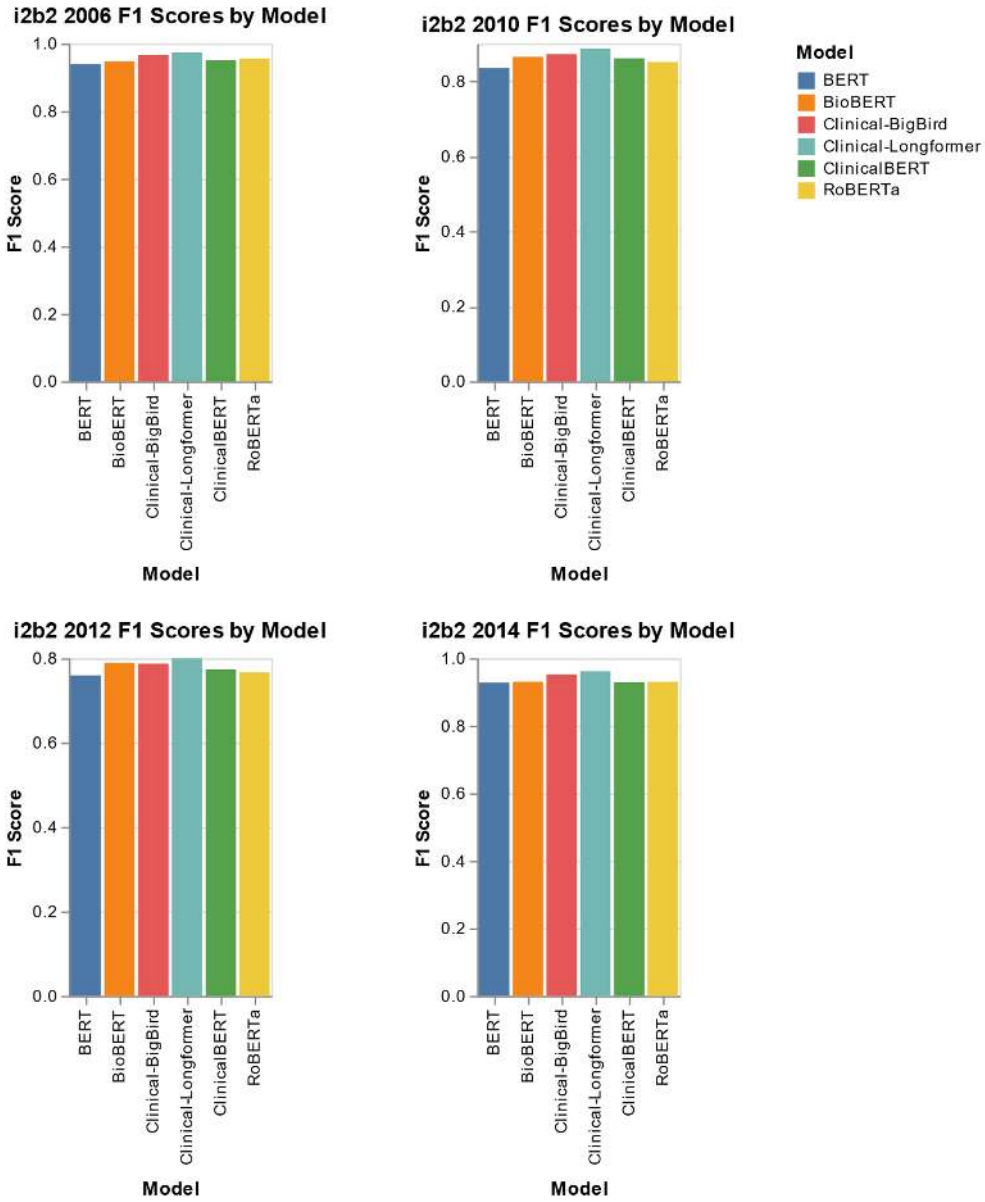
3.4.4 i2b2 2014 Veri Setinde Model Performansı

Şekil 3.1'de i2b2 2014 veri setinde, BERT modeli %92.8 F1 skoru ile diğer modellere göre en az başarılı olanıdır. ClinicalBERT ve BioBERT modelleri sırasıyla %92.9 F1 ve %93.0 F1 skorlarıyla benzer bir performans sergilerken, RoBERTa modeli aynı skoru elde etmiştir. Bu sette de Clinical-Longformer %96.1 F1 skoruyla açık ara öndeyken, Clinical-BigBird %95.2 F1 skoruyla onun hemen ardından gelmiştir. Bu veriler, i2b2 2014 gibi güncel ve spesifik veri setlerinde alana özel eğitilmiş modellerin özellikle yüksek başarım gösterdiğini ve bu modellerin gerçek dünya sağlık verileriyle çalışırken tercih edilmesi gerektiğini kanıtlar niteliktedir.

Tüm veri setlerindeki başarımlara genel bir bakış attığımızda, Clinical-Longformer'ın her birinde en üst sırayı aldığını görüyoruz; bu, onun tüm varyasyonlarda en yüksek F1 puanını elde ettiğini ve böylece kesin bir şekilde en başarılı model olduğunu kanıtlıyor. Diğer yandan, Clinical-BigBird modeli üç veri setinde ikinci sırayı kaparken, bir veri setinde oldukça yakın bir farkla üçüncü sıraya yerleşmiş. Bu sonuçlar, LLM'lerin özelleştirilmesinin, özellikle tıbbi bilgi işleme ve klinik karar destek görevlerinde belirgin bir performans artışı sağladığını

gösteriyor.

Özelleştirilmiş modeller, genel amaçlı modellere kıyasla, spesifik alan bilgisine ve terminolojiye daha hakim olduğundan, daha doğru ve alakalı çıktılar üretebilir. Bu durum, sağlık alanındaki karmaşık soruları cevaplarken veya uzun klinik metinlerden bilgi çıkarmada özellikle önemli role sahiptir. Clinical-Longformer ve Clinical-BigBird'in üstün performansları, LLM'lerin özelleştirilmesinin sadece teorik bir iyileştirme olmadığını, gerçek dünya uygulamalarında somut faydalar sağladığını ortaya koymaktadır.



Şekil 3.1 i2b2 F1-Skor Kıyaslaması [2]

3.5 Modellerin emrQA ile Test Edilmesi

Şekil 3.2'de belirtilen emrQA veri seti üzerinden yapılan test sonuçlarına göre, çeşitli modellerin medikal bilgi işleme yetenekleri incelenmiştir. EM ve F1 skorları, modellerin emrQA-veri setinin üç farklı bölümünde; medication, relation ve heart disease üzerinde nasıl performans gösterdiğini detaylandırır.

3.5.1 emrQA-Medication Performans Değerlendirmesi

Medication kategorisinde Clinical-Longformer, EM metriğinde %30.2 ve F1 skorunda %71.6 ile yüksek performans sergileyerek, özelleştirilmiş modeller arasında birinci olmuştur. Bu sonuçlar, modelin ilaç bilgisiyle ilgili verileri işleme ve doğru bilgileri sağlama konusundaki kalitesini göstermektedir. Clinical-BigBird, Clinical-Longformer'a oldukça yakın bir performansla EM'de %30.0 ve F1'de %71.5 ile takip etmiş, bu alanda özelleştirilmiş modellerin önemini tekrar vurgulamıştır.

Diğer önceden eğitilmiş modellerin karşılaştırmasında ise, BERT modeli EM'de %24.0 ve F1'de %67.5 ile bu alanda en düşük performansı göstermiştir. BioBERT, BERT'e göre biraz daha yüksek bir performansla EM'de %24.7 ve F1'de %70.0 skorlarına ulaşmıştır. ClinicalBERT, EM'de %29.7 ve F1'de %69.8 ile kendi kategorisindeki modeller arasında yüksek bir performans göstermiştir. RoBERTa ise, bu grubun içinde EM'de %28.0 ve F1'de %70.6 ile sağlam bir performans göstermiştir.

Bu veriler, genel olarak, medication konusunda Clinical-Longformer ve Clinical-BigBird modellerinin özelleştirilmiş yapılarının, ilaç bilgisinin doğruluğunu sağlamada önemli bir rol oynadığını göstermektedir. Özellikle ilaç yönetimi ve reçetelendirme süreçlerinin önemli olduğu sağlık sektöründe, bu tür özelleştirilmiş modellerin daha doğru ve güvenilir çıktılar sağlaması beklenir. Bunun yanı sıra, medication konusundaki yüksek EM ve F1 skorları, modellerin ilaçla ilgili veri işleme kapasitelerinin, klinik karar destek sistemleri ve hasta yönetimi uygulamalarında etkin bir şekilde kullanılabileceğini göstermektedir.

3.5.2 emrQA-Relation Performans Değerlendirmesi

Relation kategorisinde, modeller arasında Clinical-Longformer'ın EM skoru %91.1 ve F1 skoru %94.8 ile açık ara önde olduğunu görüyoruz. Bu skorlar, modelin hastalıklar, semptomlar ve tedaviler arasındaki ilişkileri anlama ve doğru yanıtlar verme konusunda gayet başarılı olduğunu gösteriyor. Clinical-BigBird ise %89.8 EM ve %94.4 F1 skoruyla Clinical-Longformer'ı yakından takip ediyor ve bu iki

modelin özelleştirilmiş yapısının, bu tür karmaşık klinik soruları anlamada ne kadar etkili olduğunu ortaya koyuyor.

Öte yandan, diğer modellerin skorları daha düşük seyrediyor. BioBERT, %83.6 EM ve %92.6 F1 ile sağlam bir performans sergilese de, ClinicalBERT bu alanda %84.9 EM ve %92.9 F1 ile biraz daha iyi bir sonuç elde ediyor. RoBERTa ise %82.5 EM ve %91.7 F1 skoruyla bu kategoride daha düşük bir performans gösteriyor. Bunlar, alana özel eğitilmiş modellerin, özellikle ilişki tespiti ve çıkarımı gerektiren klinik görevlerde, genel amaçlı modellere üstünlük sağladığını gösteren verilerdir.

Bu sonuçlar, Clinical-Longformer ve Clinical-BigBird'in özellikle emrQA-Relation kategorisinde gösterdiği başarı ile, sağlık alanında önemli karar verme ve teşhis koyma süreçlerine daha iyi ve doğru katkı sağlayabileceğini ve bu alanda sağlık çalışanlarına daha kaliteli bir yardımcı olacağını göstermektedir. Bu performanslar, LLM'lerin sağlık bilgilerini anlama ve ilişkilendirmeleri kurma konusunda ne kadar ileri gidebileceğini ve klinik uygulamalarda kullanılma potansiyellerini gözler önüne seriyor.

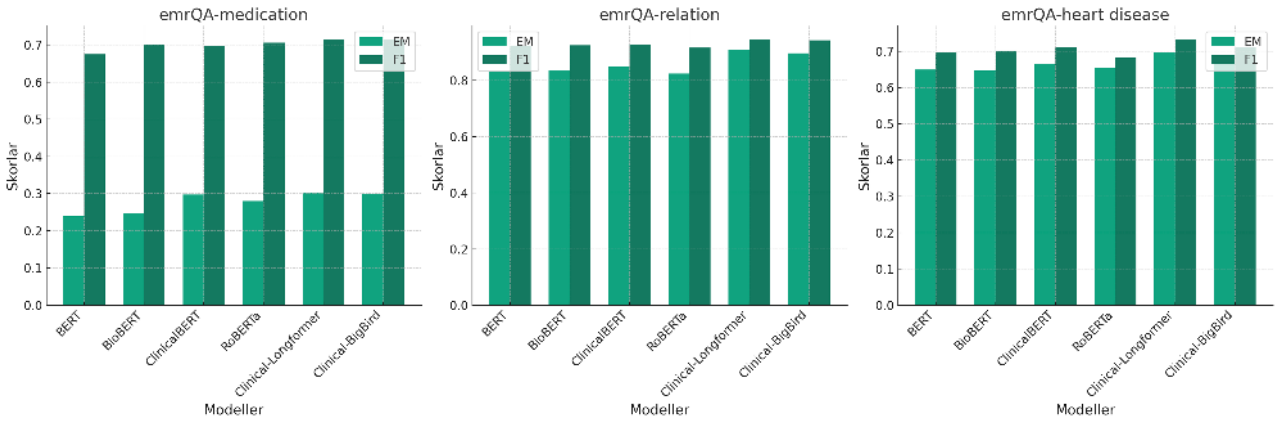
3.5.3 emrQA-Heart Disease Performans Değerlendirmesi

Kalp hastalıkları üzerine odaklanan emrQA-Heart Disease bölümünde, modellerin performansları incelendiğinde, Clinical-Longformer'ın EM metriğinde %69.8 ve F1 skorunda %73.4 ile en yüksek başarıyı elde ettiğini görüyoruz. Bu, modelin kalp hastalıkları gibi özel ve hayati bir tıbbi alanı anlama ve bilgi işleme kapasitesinin güçlü olduğunu göstermektedir. Clinical-BigBird ise EM'de %66.4 ve F1'de %71.1 ile onu takip ediyor ve özellikle F1 metriğinde, Clinical-Longformer'a oldukça yakın bir performans sergiliyor.

Diğer önceden eğitilmiş modellerin performanslarına baktığımızda, BERT EM'de %65.0 ve F1'de %69.8 ile bu alanda zorlanırken, BioBERT EM'de %64.7 ve F1'de %70.2 ile ona yakın bir performans gösteriyor. ClinicalBERT modeli biraz daha iyi bir sonuçla EM'de %66.6 ve F1'de %71.1 skorlarını almış. RoBERTa'nın performansı ise EM'de %65.5 ve F1'de %68.2 ile diğer özelleştirilmiş modellere göre daha düşük kalmıştır. ClinicalBERT'in bu veri setindeki performansının Clinical-BigBird'e göre daha iyi olduğu açıkça görülüyor.

Bu veriler, kalp hastalıkları gibi karmaşık tıbbi durumların değerlendirilmesinde özelleştirilmiş modellerin genel amaçlı modellere göre daha etkili olduğunu göstermektedir. Özelleştirilmiş modellerin bu alandaki yüksek F1 ve EM skorları, kritik tıbbi verilerin işlenmesindeki başarılarını gözler önüne seriyor. Klinik karar destek sistemlerinde bu tür modellerin kullanılmasının etkili bir yöntem

olacağını bize gösteriyor. Bu sonuçlar, sağlık sektöründeki LLM uygulamalarında özelleştirme çalışmalarının ne kadar hayati olduğunu ve bu tür teknolojilerin klinik ortamlarda karşılaşılan gerçek sorunları çözmede ne kadar yüksek bir başarıya sahip olduğunu gösteriyor.



Şekil 3.2 emrQA F1-Skor ve EM Skor Kıyaslaması [2]

4
SONUÇ

4.1 LLM'lerin Sağlık Sektörüne Etkisi

Sağlık sektöründe (LLM'lerin) kullanılması, veri analizi ve hastalık teşhisi alanında potansiyel bir devrim niteliği taşımaktadır. Bu modeller, büyük hacimli medikal veri setlerinden anlamlı bilgiler çıkararak ve klinik karar destek sistemlerine entegre olarak, hasta teşhis ve tedavisinde yeni bir dönem başlatabilir. LLM'lerin özelleştirilmesi, bu modellerin özel sağlık durumlarına ve hastalıklara özgü nüansları daha iyi anlamalarını ve böylece daha doğru teşhisler koymalarını sağlamaktadır.

4.1.1 Veri Analizi ve Hastalık Teşhisi Alanındaki Potansiyel Devrim

Özelleştirilmiş LLM'ler, klinik veri analizini ve hastalık teşhisini temelden değiştirecek potansiyele sahiptir. clinicalQA, i2b2 ve emrQA gibi veri setlerinde elde edilen yüksek F1, EM ve diğer metrik skorları, bu modellerin karmaşık medikal bilgileri işleyebilme ve doğru çıkarımlarda bulunabilme yeteneklerini göstermiştir. Özellikle Almanac, Clinical-Longformer ve Clinical-BigBird gibi modellerin, ilaç bilgisi, hastalık ilişkileri ve kalp hastalıkları, kötü promt duyarlılığı, doğru atıf yüzdesi, gerçeklik, tamamlılık, tercih gibi spesifik konularda yüksek performans göstermesi, gerçek dünya sağlık verilerinden elde edilen bilgilerin doğruluğu ve kullanışlılığını gözler önüne seriyor. Bu, LLM'lerin hem rutin tıbbi işlemlerde hem de acil durumlarda etkili birer destek personeli gibi kullanılabileceğini gösteriyor.

4.1.2 Bireyselleştirilmiş Hasta Bakımı ve Tıbbi Araştırmalarda Hızlandırma

LLM'ler, kişiselleştirilmiş hasta bakımını destekleme ve tıbbi araştırmaları hızlandırma konusunda devrim niteliğinde bir gelişmedir. Hasta geçmişi, genetik bilgi ve yaşam tarzı gibi bireysel faktörlere dayalı kişiye özgü tedavi önerilerinde bulunabilirler. Ayrıca, büyük veri setlerini analiz ederek, hastalıkların yayılımı, tedavi sonuçları ve yan etkileri hakkında hızlı, ayrıntılı ve doğru

bilgi sağlayabilirler. Bu, özellikle epidemiyolojik çalışmalara ve yeni ilaçların geliştirilmesine büyük katkı sağlar. Sağlık sektöründ, tedavi süreçlerini kişiselleştirmek ve hasta deneyimini iyileştirmek için LLM'ler biçilmiş kaftandır.

4.1.3 Veri Analizi ve Hastalık Teşhisi Alanındaki Potansiyel Devrim

Özelleştirilmiş ve gelişmiş LLM'lerin sağlık sektöründeki uygulamaları, veri analizi ve hastalık teşhisi konularında potansiyel bir devrim niteliği taşımaktadır. i2b2, emrQA gibi veri setlerinden alınan sonuçlar, bu modellerin tıbbi veri yorumlamada ve karar verme süreçlerinde önemli rol oynayabileceğini göstermektedir. Özelleştirilmiş LLM'lerin doğru teşhis koymada ve hastalık belirtileri ile ilişkili bilgileri tanıma yetenekleri, hasta sonuçlarını iyileştirmede ve sağlık hizmeti verimliliğini artırmada kullanılabilir. Bu modeller, karmaşık tıbbi bilgileri sadeleştirebilir, klinik önerilerde bulunabilir ve geniş veri setlerinden önemli bilgileri hızlıca çıkarabilir. Bu gelişme, tıbbi hizmetlerin kalitesini, hızını, fiyatını ve erişilebilirliğini artıracaktır.

4.2 Gelecek Vizyonu ve Potansiyel Sorunlar

4.2.1 Teknolojik Sınırlamalar ve Etik Standartlar

LLM'lerin sağlık sektöründeki uygulamalarının genişlemesiyle birlikte, karşılaşılacak temel sorunlardan biri teknolojik sınırlamalardır. Bu sınırlamalar, donanım kaynaklarından model eğitimi için gereken veri miktarına, işleme kapasitesinden modellerin öğrenme yeteneklerinin sınırlarına kadar geniş bir alanı kapsar. Bilhassa, özelleştirilmiş modellerin yüksek doğruluğa ulaşması için gereken spesifik ve yüksek kaliteli veri setlerinin oluşturulması, zaman alıcı ve maliyetli bir süreçtir.

Etik standartlar, özellikle sağlık sektörüne LLM'lerin entegrasyonunda çok fazla öneme haizdir. Modellerin hasta mahremiyeti ve veri güvenliği konularında nasıl davranması gerektiği, yanıltıcı veya zararlı içeriklere karşı nasıl koruma sağlayacağı gibi etik sorunlar, bu teknolojilerin güvenli ve adil bir şekilde kullanılmasını sağlamak için ele alınmalıdır.

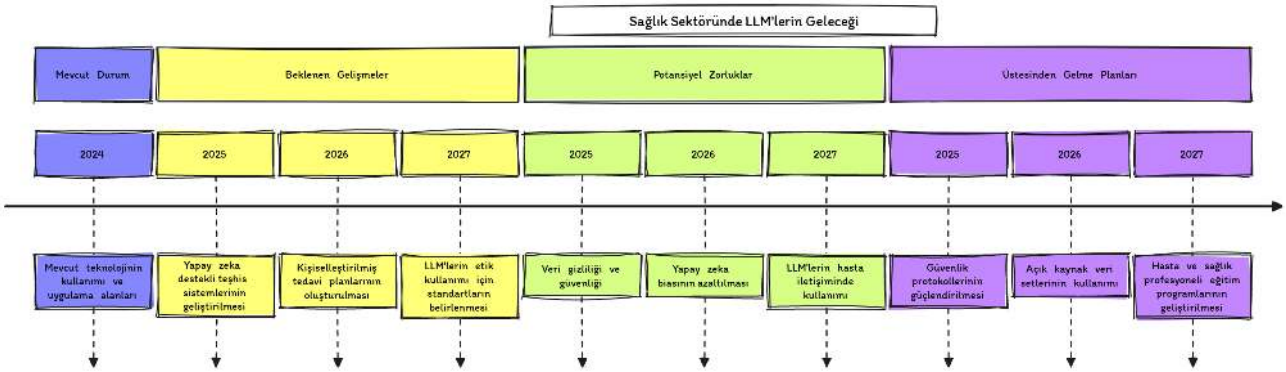
4.2.2 Yapay Zekânın Yanıltıcı Bilgi Üretme Riski

Sağlık alanında LLM'lerin yaygınlaşmasıyla beraber, yanıltıcı bilgi üretme riski de artmaktadır. Modellerin eğitim sürecinde yanlış veya taraflı veri setlerinin kullanılması, modellerin yanlış bilgiler üretmesine ve hatalı kararlar almalarına

yol açacaktır. Yapay zekânın yanlış bilgi üretme riski, özellikle hastalık teşhisi ve tedavi önerileri gibi hayati kararları etkileyebileceğinden, bu modellerin sürekli olarak doğruluk ve güvenilirlik açısından test edilmesi gerekmektedir.

Yapay zekânın sağlık alanında karşı karşıya olduğu bu ve benzeri potansiyel sorunlar, teknolojinin gelecekteki yönünü şekillendirecek ve hem faydalarını maksimize etmeye hem de risklerini minimize etmeye yönelik dikkatli bir yol haritası çizmeyi gerektirecektir. Yapay zeka sistemlerinin tasarımı ve uygulamaları, insan sağlığını ve refahını temel alan etik prensiplerle uyumlu olmalı ve güvenilir bir sağlık hizmeti sunmayı amaçlamalıdır.

Şekil 4.1'de yıllara göre potansiyel gelişmeler, zorluklar ve çözümler şematize edilmiştir. Bu öngörüye göre yaklaşık 3 yıl içinde LLM'ler (özellikle alana özel eğitilen modeller) sağlık dünyasında kullanılmaya başlanacaktır. Bu süreçte etik kaygılar da kısmen giderilmiş olacak ve LLM'ler sağlık çalışanlarıyla entegre bir biçimde kullanılmaya başlanacaktır. Kim bilir, belki de 10 yıl sonra bu kadar sağlık çalışanına da ihtiyaç kalmaz.



Şekil 4.1 Potansiyel Gelecek Akışı

KAYNAKÇA

[1] S. Liu *et al.*, "Leveraging large language models for generating responses to patient messages," *medRxiv*, pp. 2023–07, 2023.

[2] Y. Li, R. M. Wehbe, F. S. Ahmad, H. Wang, Y. Luo, "A comparative study of pretrained language models for long clinical text," *Journal of the American Medical Informatics Association*, vol. 30, no. 2, pp. 340–347, 2023.

[3] C. Zakka *et al.*, "Almanac—retrieval-augmented language models for clinical medicine," *NEJM AI*, vol. 1, no. 2, AIoa2300068, 2024.

[4] Y. Chang *et al.*, "A survey on evaluation of large language models," *ACM Transactions on Intelligent Systems and Technology*, 2023.

[5] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, D. S. W. Ting, "Large language models in medicine," *Nature medicine*, vol. 29, no. 8, pp. 1930–1940, 2023.