ChatGPT, ChatGPT Plus, Gemini ve Microsoft Copilot Büyük Dil Modellerinin Türkiye'deki Diş Hekimliği Uzmanlık Eğitimi Giriş Sınavı'nda Sorulan Ağız, Diş ve Çene Radyolojisi Sorularını Cevaplama Performanslarının Değerlendirilmesi

Assessment of the Performance of Large Language Models ChatGPT, ChatGPT Plus, Gemini, and Microsoft Copilot in Responding to Oral and Maxillofacial Radiology Questions from the Turkish Dentistry Specialty Entrance Exam

Dr. Öğr. Üyesi Ramazan Berkay PEKER
Trakya Üniversitesi, Diş Hekimliği Fakültesi, Ağız, Diş ve Çene Radyolojisi Anabilim Dalı, Edirne
ORCID ID: 0000-0003-3162-6236

Geliş tarihi: 08.10.2024

Kabul tarihi: 03.02.2025

doi: 10.5505/yeditepe.2025.93265

Yazışma adresi:
Dr. Öğr. Üyesi Ramazan Berkay PEKER
Trakya Üniversitesi, Diş Hekimliği Fakültesi,
Ağız, Diş ve Çene Radyolojisi Anabilim Dalı,
22030 Edirne, TÜRKİYE
Adres: Edirne, TÜRKİYE
Tel: +90 284 2364552
Faks: +90 284 2364550
E-posta: rberkaypeker@trakya.edu.tr

ÖZET

Amaç: Bu çalışmanın amacı biri ücretli (ChatGPT-4o), üçü ücretsiz (ChatGPT-4o mini, Gemini, Microsoft Copilot) dört büyük dil modelinin (BDM) 2012-2021 yılları arasında yapılan Diş Hekimliği Uzmanlık Eğitimi Giriş Sınavı'nda (DUS) sorulmuş olan Ağız, Diş ve Çene Radyolojisi (ADÇR) sorularını cevaplama performansını değerlendirip karşılaştırmaktır.

Gereç ve Yöntem: 2012-2021 yılları arasında DUS'ta sorulmuş olan 123 soru, "oral diagnoz" ve "radyoloji" kategorilerine ayrılarak dört BDM'ye soruldu. BDM'lerin verdiği cevabın doğruluğuna göre elde edilen veriler Pearson Ki Kare Testi, Monte Carlo düzeltmeli Fisher Exact Testi ve Bonferroni düzeltmeli z Testi kullanılarak analiz edildi ($p<0.05$).

Bulgular: Tüm sorulara verilen doğru cevap oranı ChatGPT-4o mini'de %74, ChatGPT-4o'da %91,1, Gemini'de %69,9 ve Microsoft Copilot'ta %86,2 olarak elde edilmiştir. Sadece radyoloji kategorisinde verilen cevaplarda BDM'ler arasında istatistiksel olarak anlamlı bir ilişki bulunmuştur ($p=0,054$; $p<0,001$).

Sonuç: Dört BDM arasında ücretli olan ChatGPT-4o ve ücretsiz olan Microsoft Copilot, istatistiksel olarak birbirine benzer ve sorulan soruların %80'inden fazlasına doğru cevap verme performansı sergilemiştir. Önceki çalışmalar analiz edildiğinde BDM'lerin hızlı bir gelişim gösterdiği gözlenmektedir. BDM'ler ilerleyen zamanlarda diş hekimliği eğitiminde etkin bir şekilde rol oynayabilecektir.

Anahtar Kelimeler: Yapay zeka; büyük dil modelleri; ChatGPT; Microsoft Copilot; Gemini; çoktan seçmeli soru.

ABSTRACT

Aim: The aim of this study was to evaluate and compare the performance of four large language models (LLM), one paid (ChatGPT-4o) and three free (ChatGPT-4o mini, Gemini, Microsoft Copilot), in answering Oral, Dental and Maxillofacial Radiology questions asked in the Dental Specialty Training Entrance Examination between 2012 and 2021.

Materials and Method: The 123 questions asked in Dental Specialty Training Entrance Examination between 2012 and 2021 were divided into "oral diagnosis" and "radiology" categories and asked to four LLMs. The data obtained according to the accuracy of the answers given by the LLMs were analyzed using Pearson Chi-Square Test, Fisher Exact Test with Monte Carlo correction and z Test with Bonferroni correction ($p<0.05$).

Results: The correct answer rate for all questions was 74% in ChatGPT-4o mini, 91.1% in ChatGPT-4o, 69.9% in Gemini and 86.2% in Microsoft Copilot. A statistically significant correla-

tion was found among the LLMs only in the answers given to the radiology category (p<0.001).

Conclusion: Among the four LLMs, ChatGPT-4o, a paid LLM, and Microsoft Copilot, a free LLM, performed statistically similar to each other and answered more than 80% of the questions correctly. When previous studies are analyzed, it is observed that LLMs are developing rapidly. LLMs will be able to play an effective role in dental education in the future.

Keywords: Artificial intelligence; large language models; ChatGPT; Microsoft Copilot; Gemini; multiple choice questioning.

GİRİŞ

Son yıllarda yapay zeka (YZ) tabanlı sohbet robotları, bilgiye erişim ve işleme süreçlerinde önemli değişimlere yol açmıştır. Büyük dil modelleri (BDM), geniş veri setleri üzerinde eğitilerek kullanıcıların sorularına yüksek doğrulukla yanıt verme kapasitesine sahiptir. Yapılan çalışmalar, bu sistemlerin farklı kaynaklardan gelen bilgileri sentezleyerek anlamlı çıktılar oluşturabildiğini ve özellikle tıp eğitiminde sınav başarısını değerlendirme süreçlerinde araştırmalara konu olduğunu göstermektedir.[1] Ancak, BDM'lerin bilgi işleme yetkinlikleri yüksek olmasına rağmen, insan uzmanlığının sağladığı deneyimsel karar verme becerisinden yoksun oldukları bilinmektedir. Bu nedenle, tıp ve diş hekimliği gibi alanlarda BDM'lerin akademik performanslarını ölçmek ve sınav başarısına etkilerini değerlendirmek önem arz etmektedir.

YZ destekli dil modelleri, son yıllarda eğitim, tıp ve sağlık bilimleri alanlarında giderek daha fazla ilgi görmektedir. Bu modeller, geniş veri setleriyle eğitilerek farklı disiplinlerde bilgi sentezleme ve problem çözme yetenekleriyle öne çıkmaktadır. Chat Generative Pre-Trained Transformer (ChatGPT) başta olmak üzere, Microsoft Copilot ve Google'ın Gemini gibi modelleri de tıp eğitiminde ve sınav değerlendirmelerinde potansiyel kullanım alanlarıyla dikkat çekmektedir. Yapılan araştırmalar, ChatGPT ve diğer büyük dil modellerinin tıbbi sınavlardaki performanslarını değerlendirmeye yönelik çalışmaların arttığını ve bu sistemlerin eğitim süreçlerine entegrasyonu üzerine tartışmaların sürdüğünü göstermektedir.[1,2,3]

BDM'ler, büyük hacimli eğitim verileriyle etkileşim sağlayarak eğitilir ve bilgi işleme yeteneklerini geliştirir. Bu eğitim, BDM'lerin doğru tahminlerde bulunabilmesi için veri setlerindeki düzenleri anlamalarına ve veriler arasındaki bağlantıları öğrenmelerine yardımcı olur. BDM'lerin bilgi işleme kapasitesi, yalnızca eğitildikleri veri miktarıyla değil, aynı zamanda bu verilerin özgüllüğü ve güncelliğiyle de doğrudan ilişkilidir. Daha küçük ve daha spesifik veri alt kümeleriyle daha fazla etkileşim sağlandığında, bu modellerin belirli bir alandaki doğruluk oranlarının arttığı gösterilmiştir. Ancak, veri setlerinin kapsamı ve kalitesine bağlı olarak, BDM'lerin tıbbi karar verme süreçlerindeki güvenilirliği değişkenlik gösterebilir.[1]

BDM'ler birçok alanda başarılı uygulamalar sergilemiş olmasına rağmen, sağlık alanındaki etkinlikleri henüz tam olarak netleşmemiştir. Bu durum, BDM'lerin tıbbi karar verme süreçlerinde insan uzmanlığına özgü deneyimsel muhakeme becerisine sahip olmamasından kaynaklanmaktadır. Özellikle karmaşık teşhislerin doğru konulabilmesi için klinik tecrübe ve bağlamsal yorumlama gereklidir. Bu eksikliğe rağmen, BDM'lerin sağlık hizmetlerinde yardımcı bir araç olarak potansiyelini değerlendirmeye yönelik ilgi giderek artmaktadır.

Üretken BDM'ler yalnızca teşhis süreçlerinde değil, aynı zamanda tıp ve diş hekimliği eğitiminde de giderek daha fazla araştırma konusu olmaktadır. Yapılan çalışmalar, bu modellerin öğrencilere bireyselleştirilmiş geri bildirim sağlama, tıbbi bilgiye erişimi kolaylaştırma ve sınav başarısını arttırma gibi çeşitli alanlarda potansiyel taşıdığını göstermektedir.[3] Özellikle, tıp eğitimi alanında BDM'lerin kanıta dayalı öğrenme süreçlerine katkı sağlayabileceği ve sınav performansını değerlendirmek için kullanılabileceği vurgulanmaktadır.[2,4] Ancak, eğitimdeki etkin kullanımları halen tartışmalı olup, etik, güvenilirlik ve doğruluk gibi unsurlar dikkate alınması gereken önemli faktörlerdir.

Diş Hekimliği Uzmanlık Eğitimi Giriş Sınavı (DUS), diş hekimliği lisans eğitimi sonrası uzmanlık eğitimi almak isteyen adaylar için 2012 yılından itibaren Ölçme, Seçme ve Yerleştirme Merkezi (ÖSYM) tarafından uygulanan merkezi bir sınavdır. DUS, 40 temel bilim ve 80 klinik bilim sorusundan oluşan toplam 120 soruluk çoktan seçmeli tek cevaplı olan kapsamlı bir değerlendirme sınavıdır. Son yıllarda BDM'lerin belirli bir standarda sahip mesleki bilgi ölçüm sınavlarındaki performansı ve eğitim süreçlerine entegrasyonu, giderek daha fazla akademik araştırmanın odağı haline gelmiştir.

Bu çalışmanın amacı; üçü ücretsiz (ChatGPT, Gemini, Microsoft Copilot) biri ücretli (ChatGPT Plus) toplam dört popüler ve baskın BDM'nin DUS'ta sorulan Ağız, Diş ve Çene Radyolojisi (ADÇR) sorularına verdikleri cevapların performanslarını ölçüp birbirleriyle kıyaslamaktır. Bu sayede bu dört farklı BDM'nin ADÇR alanına özgü bilgilerinin doğruluğu tespit edilip karşılaştırılmış olacaktır.

GEREÇ ve YÖNTEM

ÖSYM'nin 2012 yılından 2021 yılına kadar sorduğu tüm DUS soruları, https://www.osym.gov.tr/TR,15070/dus-cikmis-sorular.html adresinden indirildi. Her sınavın ADÇR'ye ait soruları seçilerek kaydedildi. Çalışmada kullanılacak BDM'lerin bir kısmında destek olmadığı için içeriğinde görüntü, resim, şekil, tablo olan sorular çalışma dışı bırakıldı. Böylece toplam 130 sorunun 7 tanesi çalışma dışı

bırakıldı.

Çalışmaya dahil edilen sorular, https://dosyamerkez.saglik.gov.tr/Eklenti/29523/0/agizdisveceneradyolojisimufredatv23pdf.pdf adresinde bulunan Ağız, Diş ve Çene Radyolojisi Uzmanlık Eğitimi Çekirdek Müfredatı baz alınarak iki kategoriye ayrıldı. Müfredatta bulunan Oral ve Perioral Bölgenin Değerlendirilmesi, Oral ve Perioral Bölge Yumuşak Doku Lezyonları (Enfeksiyöz), Oral ve Perioral Bölge Sert Doku Lezyonları, Oral ve Perioral Bölge Lezyonları, Sistemik Hastalıklarda Oral Bulgular başlıkları ile ilgili sorular "Oral Diagnoz", Görüntü Değerlendirmesi başlığı ile ilgili sorular "Radyoloji" adı altında kategorize edildi. Böylece 123 sorunun 30'u "Oral Diagnoz", 93'ü "Radyoloji" kategorisinde değerlendirildi.

Sorular sorulmadan önce her BDM'ye "Aşağıda soracağım çoktan seçmeli soruların doğru cevabını söyler misin?" komutu kullanılarak standardizasyon sağlandı. Bu komuttan sonra sorular, ChatGPT'nin ücretsiz versiyonu olan ChatGPT-4o mini'ye (OpenAI Incorporated, Mission District, San Francisco, United States), ChatGPT'nin ücretli versiyonu olup ChatGPT Plus olarak adlandırılan ChatGPT-4o'ya (OpenAI Incorporated, Mission District, San Francisco, United States), Google Gemini'ye (Alphabet Inc., CA, US) ve Microsoft Copilot'a (Microsoft Corporation, WA, US) simultane olarak soruldu. Soruların sorulma aşaması 15.09.2024-22.09.2024 tarihleri arasında yapıldı. Elde edilen veriler IBM SPSS V23 ile analiz edildi. Kategorik veriler arasındaki ilişkinin incelenmesinde Pearson Ki Kare Testi ve Monte Carlo düzeltmeli Fisher Exact Testi kullanıldı. Çoklu karşılaştırmalar Bonferroni düzeltmeli z Testi ile incelendi. Sonuçlar frekans (yüzde) şeklinde sunuldu. Önem düzeyi p<0,05 olarak alındı.

BULGULAR

Oral diagnoz kategorisine verilen cevaplar ve BDM'ler arasında istatistiksel olarak anlamlı bir ilişki bulunmamıştır (p=0,054). Oral diagnoz kategorisindeki sorulara verilen doğru cevap oranı ChatGPT-4o mini'de %93,3 iken ChatGPT-4o'da %100, Gemini'de %86,7 ve Microsoft Copilot'ta %100 olarak elde edilmiştir.

Radyoloji kategorisine verilen cevaplar ve BDM'ler arasında istatistiksel olarak anlamlı bir ilişki bulunmuştur (p<0,001). Radyoloji kategorisindeki sorulara verilen doğru cevap oranı ChatGPT-4o mini'de %67,7 iken ChatGPT-4o'da %88,2, Gemini'de %64,5 ve Microsoft Copilot'ta %81,7 olarak elde edilmiştir. ChatGPT-4o ve Microsoft Copilot ikilisi kendi arasında, ChatGPT-4o mini ve Gemini ikilisi de kendi arasında birbirine benzer performans göstermiştir.

Tüm sorular toplu olarak incelendiğinde bu sorulara verilen cevaplar ve BDM'ler arasında istatistiksel olarak anlamlı bir ilişki bulunmuştur (p<0,001). Tüm sorulara verilen doğru cevap oranı ChatGPT-4o mini'de %74 iken ChatGPT-4o'da %91,1, Gemini'de %69,9 ve Microsoft Copilot'ta %86,2 olarak elde edilmiştir. ChatGPT-4o ve Microsoft Copilot ikilisi kendi arasında, ChatGPT-4o mini ve Gemini ikilisi de kendi arasında birbirine benzer performans göstermiştir.

BDM'ler ve kategoriler arasındaki ilişki Tablo 1'de gösterilmiştir.

Tablo 1. Büyük dil modelleri ve kategoriler arasındaki ilişkinin incelenmesi.

	Büyük Dil Modeli				Toplam	Test İstatistiği	p
	ChatGPT-4o mini	ChatGPT-4o	Gemini	Microsoft Copilot			
Oral Diagnoz							
Doğru	28 (93,3)	30 (100)	26 (86,7)	30 (100)	114 (95)	6,260	0,054*
Yanlış	2 (6,7)	0 (0)	4 (13,3)	0 (0)	6 (5)		
Radyoloji							
Doğru	63 (67,7)^{ab}	82 (88,2)^{c}	60 (64,5)^{b}	76 (81,7)^{ac}	281 (75,5)	19,130	**<0,001** **
Yanlış	30 (32,3)	11 (11,8)	33 (35,5)	17 (18,3)	91 (24,5)		
Genel Toplam							
Doğru	91 (74)^{ab}	112 (91,1)^{c}	86 (69,9)^{b}	106 (86,2)^{ac}	395 (80,3)	23,152	**<0,001** **
Yanlış	32 (26)	11 (8,9)	37 (30,1)	17 (13,8)	97 (19,7)		

*Fisher Exact Testi; **Pearson Ki KareTesti; Frekans (Yüzde); a-c Aynı harfe sahip gruplar arasında fark yoktur.

BDM'lerin sınav yılları ile ilişkisi incelendiğinde (Tablo 2), sadece 2013 yılında sorulan sorulara verilen cevaplar ve BDM'ler arasında istatistiksel olarak anlamlı bir ilişki bulunmuştur (p=0,003). 2013 yılında sorulan sorulara verilen doğru cevap oranı ChatGPT-4o mini'de %78,9 iken ChatGPT-4o'da %100, Gemini'de %68,4 ve Microsoft Copilot'ta %100 olarak elde edilmiştir. ChatGPT-4o mini, diğer tüm BDM'lerle benzer performans sergilerken; Gemini, ChatGPT-4o ve Microsoft Copilot'tan farklı performans sergilemiştir.

Tablo 2. Büyük dil modelleri ve sınav tarihleri arasındaki ilişkinin incelenmesi.

	Büyük Dil Modeli				Toplam	Test İstatistiği	p
Yıllar	ChatGPT 4o mini	ChatGPT 4o	Gemini	Microsoft Copilot			
2012							
Doğru	15 (78,9)	18 (94,7)	14 (73,7)	17 (89,5)	64 (84,2)	3,804	0,332*
Yanlış	4 (21,1)	1 (5,3)	5 (26,3)	2 (10,5)	12 (15,8)		
2013							
Doğru	15 (78,9)^{abc}	19 (100)^{c}	13 (68,4)^{b}	19 (100)^{ac}	66 (86,8)	11,774	**0,003***
Yanlış	4 (21,1)	0 (0)	6 (31,6)	0 (0)	10 (13,2)		
2014							
Doğru	16 (80)	19 (95)	15 (75)	18 (90)	68 (85)	3,769	0,337*
Yanlış	4 (20)	1 (5)	5 (25)	2 (10)	12 (15)		
2015							
Doğru	6 (60)	8 (80)	7 (70)	7 (70)	28 (70)	1,077	0,962*
Yanlış	4 (40)	2 (20)	3 (30)	3 (30)	12 (30)		
2016							
Doğru	6 (60)	8 (80)	7 (70)	8 (80)	29 (72,5)	1,429	0,860*
Yanlış	4 (40)	2 (20)	3 (30)	2 (20)	11 (27,5)		
2017							
Doğru	7 (87,5)	8 (100)	6 (75)	6 (75)	27 (84,4)	2,651	0,714*
Yanlış	1 (12,5)	0 (0)	2 (25)	2 (25)	5 (15,6)		
2018							
Doğru	7 (70)	8 (80)	7 (70)	8 (80)	30 (75)	0,726	1,000*
Yanlış	3 (30)	2 (20)	3 (30)	2 (20)	10 (25)		
2019							
Doğru	6 (75)	8 (100)	5 (62,5)	8 (100)	27 (84,4)	5,424	0,126*
Yanlış	2 (25)	0 (0)	3 (37,5)	0 (0)	5 (15,6)		
2020							
Doğru	7 (70)	7 (70)	6 (60)	8 (80)	28 (70)	1,077	0,962*
Yanlış	3 (30)	3 (30)	4 (40)	2 (20)	12 (30)		
2021							
Doğru	6 (66,7)	9 (100)	6 (66,7)	7 (77,8)	28 (77,8)	4,108	0,305*
Yanlış	3 (33,3)	0 (0)	3 (33,3)	2 (22,2)	8 (22,2)		
Genel toplam							
Doğru	91 (74)^{ab}	112 (91,1)^{c}	86 (69,9)^{b}	106 (86,2)^{ac}	395 (80,3)	23,152	**<0,001***
Yanlış	32 (26)	11 (8,9)	37 (30,1)	17 (13,8)	97 (19,7)		

*Fisher Exact Testi; **Pearson Ki Kare Testi; Frekans (Yüzde); a-cAynı harfe sahip gruplar arasında fark yoktur.

TARTIŞMA

Çalışmanın ana bulguları yorumlandığında, YZ destekli BDM'lerin DUS ADÇR sorularının doğru cevabını tahmin etmesinde genel olarak yüksek bir doğruluk oranı mevcut olduğu görülmektedir. Microsoft Copilot ücretsiz bir BDM olmasına rağmen ücretli olan ChatGPT-4o ile istatistiksel olarak benzer performans göstermiştir. BDM'ler arasında

en düşük performansı Gemini göstermiştir.

Literatürde farklı sınavlarda BDM'lerin performanslarının incelendiği çalışmalar mevcuttur.[5,6,8] Japonya'da yapılan bir çalışmada, GPT-4'ün Japon Tıbbi Lisanslama Sınavı'nda GPT-3.5'e kıyasla belirgin bir üstünlük sağladığı rapor edilmiştir. Çalışmada, GPT-4'ün tüm soru kategorilerinde başarılı olduğu ve özellikle zorlu sorularda daha yüksek doğruluk oranları elde ettiği belirtilmiştir.[5]Bu durum, GPT-4'ün klinik akıl yürütme ve tıbbi bilgi açısından güvenilir bir araç olma potansiyelini ortaya koymaktadır.

Birleşik Devletler Tıbbi Lisanslama Sınavı USMLE'de, ChatGPT-3'ün sınav basamaklarına göre %60 doğruluk eşiğinde veya bu değere yakın bir performans gösterdiği bildirilmiştir.[6] Bu çalışmada ChatGPT'nin özellikle yüksek içsel tutarlılık sergilediği ve cevapları açıklarken klinik içgörüler sağladığı vurgulanmıştır. ChatGPT'nin öğrencilere yeni bir bakış açısı sağlayarak eğitim süreçlerini iyileştirme potansiyeline sahip olduğu öne sürülmüştür. Çalışmanın sonucunda yapılan çıkarımda sınavdaki doğruluk oranının içerik türü ve modelin eğitimiyle doğrudan bağlantılı olabileceği de belirtilmiştir.[6]

Benzer şekilde, USMLE sorularında BDM performanslarını inceleyen bir diğer çalışmada, ücretli olan GPT-4 modeli %90 doğruluk oranı sergilerken, ücretsiz olan ChatGPT-3.5 Turbo modelinin %62,5 doğruluk oranına ulaştığı bildirilmiştir.[4] GPT-4'ün yanıtlama tutarlılığının ChatGPT-3.5 Turbo'ya kıyasla daha yüksek olduğu da vurgulanmıştır.[4] Lee ve ark. yaptıkları çalışmada GPT-4 modelinin USMLE sorularında %90'dan fazla doğru yanıt oranı sergilediğini bildirmiştir.[7] Çalışmada GPT-4'ün başarısına rağmen ortaya çıkabilecek "halüsinasyon" olarak adlandırılan hatalı bilgi sunma riskinin de olduğu vurgulanmaktadır. Özellikle tıbbi karar verme senaryolarında bu riskin farkında olunması ve çıktıların profesyoneller tarafından doğrulanmasının gerekliliği vurgulanmıştır. Ek olarak, GPT-4'ten "komut mühendisliği" teknikleriyle daha verimli cevaplar alınabileceği belirtilmektedir.[7] Bizim çalışmamızda da bu prensiplerin uygulanması, BDM'lerin DUS ADÇR sorularında sundukları yanıtların kalitesini doğrudan etkileyebileceğini ortaya koymaktadır.

ChatGPT'nin radyoloji kurul sınavlarına benzer çoktan seçmeli sorular üzerindeki performansı da benzer eğilimler sergilemektedir. Bhayana ve ark. tarafından yapılan bir çalışmada, ChatGPT'nin radyoloji kurul sınavına yönelik 150 çoktan seçmeli soruda %69 doğruluk oranına ulaştığı bildirilmiştir.[8] Çalışmada ChatGPT'nin bilgi hatırlama ve temel kavramları anlama gibi düşük düzeyde düşünme gerektiren sorularda %84 doğruluk oranı elde ettiği, ancak kavramları uygulama ve analiz etme gibi daha üst düzey düşünme gerektiren sorularda bu oranın %60'a düştüğü belirtilmiştir. Ayrıca model, radyolojiye özgü fizik sorularında %40 doğruluk oranıyla sınırlı bir performans sergilemiş, bu oran genel klinik sorulardaki %73 doğruluk oranına kıyasla anlamlı bir farklılık ortaya koymuştur.[8] Bu durum, modellerin belirli alanlardaki bilgi derinliğinin farklılık gösterebileceğini ortaya koymaktadır. Gilson ve ark. çalışmasında, ChatGPT'nin USMLE 1.basamak ve 2.basamak sınavlarında %60 doğruluk eşiğini geçerek üçüncü sınıf tıp öğrencilerinin seviyesinde performans gösterdiği belirtilmiştir.[9] Farklı veri setlerinde yapılan incelemelerde, özellikle daha zor sorularda performansın düştüğü rapor edilmiştir. Bununla birlikte, doğru cevaplarda kullanılan mantıksal açıklama ve bağlamsal bilgilerin yüksek oranlarda olduğu gözlemlenmiştir.[9] Bu sonuçlar, ChatGPT'nin eğitim süreçlerinde küçük grup öğrenimi veya bireysel çalışma için potansiyel bir araç olarak değerlendirilebileceğini göstermektedir.[9]

Chen ve ark. 2024 yılında gerçekleştirdiği çalışmada, GPT-4o modelinin Amerika, İngiltere, Hong Kong ve Çin'de uygulanan tıbbi lisanslama sınavlarında, diğer modellere kıyasla üstün bir performans sergilediği rapor edilmiştir.[10] Üçü İngilizce, biri Çince uygulanan bu sınavlarda dil ve kültürel çeşitliliğin model performansı üzerindeki etkileri net bir şekilde ortaya konmuştur. İngilizce olan sınavlarda GPT-4o modeli diğer modellere kıyasla üstün performans sergilerken, Çince olan sınavda tüm BDM'lerin sergilediği performansın daha düşük olduğu gözlemlenmiştir.[10] Bu durum, modellerin dil yetkinliklerinin eğitim veri setlerinin kapsamına bağlı olarak değiştiğini göstermektedir. Bu bulgular, BDM'lerin daha çeşitli ve kapsayıcı veri setleriyle eğitilmesinin gerekliliğini açıkça vurgulamaktadır.

Yükseköğretimde yapılan kapsamlı bir inceleme, ChatGPT'nin çoktan seçmeli sınavlarda sergilediği performansın genel olarak başarılı olduğunu göstermiştir.[11] ChatGPT-4, genellikle ortalama bir öğrencinin performansına yakın sonuçlar vermiş, ancak ChatGPT-3.5 bu başarıyı yakalayamamıştır.[11] Bu fark, GPT-4'ün daha gelişmiş bir model olduğunu ve sınav sorularında üst düzey düşünme becerilerinin önemini ortaya koymaktadır.

Onkoloji alanında yapılan bir çalışmada ise ChatGPT-3.5'in Amerikan Klinik Onkoloji Derneği Öz Değerlendirmesi çoktan seçmeli test sınavı sorularında %56,1'lik bir doğruluk oranına ulaştığı, ancak bu oranın onkologlar için geçerli yeterlilik düzeylerinin altında olduğu bildirilmiştir.[12] Aynı çalışmada BDM'lerin onkolojinin uzmanlık alanlarına göre farklı başarı oranları sergilediği belirtilmiş, modelin sınırlı eğitim verisi ve gerçek zamanlı klinik bilgiye erişim eksikliği gibi temel problemleri olduğu vurgulanmıştır. Ayrıca BDM'lerde bulunan eğitim verisinin güncelliği ve erişilebilirliği gibi faktörlerin BDM'lerin gösterdiği performansı doğrudan etkilediği belirtilmiştir.[12]

Onkoloji alanında yapılan bir diğer çalışmada, ChatGPT-4'ün Amerikan Klinik Onkoloji Derneği ve Avrupa Klinik Onkoloji Derneği sınav sorularında %85 doğruluk oranı elde ettiği bildirilmiştir. Yanlış cevapların %81,8'inin klinik uygulamalarda orta-yüksek seviyede zarara yol

açabileceği vurgulanmıştır. Çalışmada, özellikle dinamik ve hızla değişen literatür bilgisinin gerektiği durumlarda BDM'lerin bilgiye erişimdeki zorluklarına dikkat çekilmiştir.[13] Bu bulgu, sağlık alanında kullanılan BDM'lerin sürekli güncellenmesi ve güvenli kullanım protokollerinin geliştirilmesi gerekliliğini ortaya koymaktadır. Ayrıca, doğru cevap oranının yüksek olmasına rağmen yanlış cevapların olası zarar seviyesinin düşük tutulması için insan denetiminin gerekliliği belirtilmiştir.[13]

Benzer şekilde, oftalmoloji alanında yapılan bir çalışmada, ChatGPT'nin farklı sürümlerinin performansı değerlendirilmiştir. Bu çalışmada, özellikle ChatGPT'nin genel tıbbi bilgilerde yüksek bir doğruluk oranına sahip olduğu ancak oftalmoloji gibi daha spesifik alt alanlarda sınırlamalar yaşadığı belirtilmiştir. Bu durum, temel eğitim veri setlerinin kapsamı ve uzmanlık alanlarına özgü veri yetersizliği ile ilişkilendirilmiştir.[14] Çalışmada, ChatGPT'nin genel tıbbi sorulara verdiği yanıtların doğruluk oranı %59,4'e kadar çıkarken, oftalmolojiye özgü daha karmaşık sorularda bu oran daha düşüktür. Bu bulgu, BDM'lerin sağlık gibi çok disiplinli alanlarda daha fazla uzmanlaşması gerektiğini göstermektedir. Bu nedenle, daha kapsamlı ve alanına özgü verilerle model eğitimi, BDM'lerin doğruluğunu ve kullanılabilirliğini artırmada önemli bir adım olacaktır.

Gemini ve ChatGPT-4'ün oftalmoloji alanındaki performansını inceleyen başka bir çalışmada, her iki modelin %80'in üzerinde doğruluk oranı sergilediği gösterilmiştir.[15] Gemini'nin "Bilmiyorum" seçeneğini daha sık kullanarak hata oranını azalttığı, ChatGPT-4'ün ise daha iddialı bir yaklaşım sergileyerek daha fazla hata yaptığı belirtilmiştir.[15] Bu farklılıklar, iki modelin eğitim süreçleri ve tasarım ilkeleri arasındaki farklara işaret etmektedir.

Diğer yandan, nöroşirurji alanında yapılan bir çalışmada, ChatGPT-3.5 %73,4 doğruluk oranı sergilerken, GPT-4 %83,4 oranıyla sınav barajını aşmış ve insan kullanıcılardan daha iyi bir performans göstermiştir.[16] Bu çalışmada, GPT-4'ün daha karmaşık girişleri işleme ve çok adımlı problem çözme konularında belirgin bir üstünlük sağladığı belirtilmiştir. Ancak, görüntü içeren sorularda her iki modelin de performansının düşük olduğu ifade edilmiştir.[16] Bu durum, BDM'lerin görüntüleme verilerinde sınırlı entegrasyona sahip olduğunu göstermektedir.

Schubert ve ark. ChatGPT-4'ün nöroloji kurul sınavlarına benzer sorular üzerinde %85 doğruluk oranı yakaladığını ve insan kullanıcılarının ortalama performansını (%73,8) geride bıraktığını bildirmiştir.[17] Çalışma, ChatGPT-4'ün özellikle davranışsal, bilişsel ve psikolojik kategorilerde insanlardan daha üstün performans sergilediğini göstermiştir. Ancak modelin, üst düzey düşünme gerektiren soruların yanı sıra görüntü veya kapsamlı tıbbi açıklamalar içeren sorularda daha düşük performans sergilediği belirtilmiştir.[17] Bu bulgu, BDM'lerin dinamik veri entegrasyonu ve tıbbi karar desteği konularındaki sınırlamalarına dikkat çekmektedir. Çalışmada ayrıca, ChatGPT-3.5 ve ChatGPT-4'ün, verdikleri bazı yanlış yanıtlar konusunda yüksek bir güven sergilediği belirtilmiş ve bu durumun, modellerden alınan çıktılar üzerinde profesyonel bir doğrulama yapılmasının önemini artırdığı ifade edilmiştir.[17]

Literatürde, tıbbi alanlarda BDM'lerin performansını değerlendiren çalışmalar, modellerin bilgi düzeyi ve alan uzmanlığına bağlı olarak değişen başarı oranlarını ortaya koymaktadır.[1,3,18,19] Botross ve ark. Oftalmoloji Kurulu Sınavı Alıştırma Soruları üzerine yaptığı bir çalışmada, Gemini olarak bilinen Bard'ın %62,4 doğruluk oranına ulaştığı, ancak bütünsel anlayış ve deneyime dayalı bilgiden yoksun olduğu belirtilmiştir.[1] Oftalmoloji alanındaki bir diğer çalışmada ise ChatGPT-4'ün %70'lik doğruluk oranına karşılık, ChatGPT-3.5'in yalnızca %55 doğruluk oranı sergilediği tespit edilmiştir.[8] İtalyan Üniversite ve Araştırma Bakanlığı ve İtalyan Eğitim Bakanlığı'ndan oluşan bir hükümet kuruluşu olan CINECA kuruluşunun yaptığı sınavın sorularının incelendiği bir çalışmada ChatGPT-4 ve Microsoft Copilot'ın, Gemini'den daha yüksek performans sergilediği bildirilmiştir.[3]

DUS sorularını değerlendiren bir çalışmada, Protetik Diş Tedavisi kategorisinde ChatGPT-3.5 %35,7'lik bir doğruluk oranı gösterirken, Bard (Gemini) %38,9'luk doğruluk oranına ulaşmıştır. ADÇR sorularında ise her iki modelin de %52,8'lik bir başarı elde ettiği belirtilmiştir.[19] Bu bulgular, modellerin performansının alan ve soru türüne göre değişkenlik gösterebileceğini ortaya koymaktadır.

Literatür incelendiğinde, çalışmamızda da olduğu gibi YZ destekli BDM'lerin aynı sorular üzerinde farklı doğruluk oranları sergilediği gözlemlenmiştir.[4,9] Bu durum ChatGPT ve Microsoft Copilot'ın GPT mimarisine dayanması, Gemini'nin ise LaMDA (Language Model for Dialogue Application) ve PaLM 2 (Pathways Language Model) teknolojilerini kullanmasıyla açıklanabilir.[20] Çalışmamızda, ChatGPT ve Microsoft Copilot'ın Gemini'ye kıyasla daha yüksek performans gösterdiği belirlenmiştir.

Farklı çalışmalar arasındaki değişkenliklerin başlıca nedeni, soruların aynı branştan olmamasıdır. Bununla birlikte, Turunç Oğuzman ve ark.[19] çalışmasında kullanılan sorular ile bu çalışmadaki sorular aynı olmasına rağmen, başarı oranlarının farklılık göstermesi dikkat çekicidir. Bu farklılığın, bu çalışmada kullanılan BDM'lerin sürümlerinin daha güncel olmasından kaynaklandığı düşünülmüştür. Ayrıca, literatürde GPT-4'ün sağlık alanında genel bilgi seviyesini artırmak için açık kaynak verilerle eğitildiği ve spesifik tıbbi alanlarda daha fazla gelişim gösterebileceği belirtilmektedir.[7] Bu bulgular, zamanla BDM'lerin bilgi seviyelerinin daha da artacağını ve gelişimlerinin süreceğini göstermektedir.

Çalışmamızın güçlü yönlerine gelecek olursak, bu araştırma, dört farklı BDM'nin DUS ADÇR sorularında verdiği cevapları aynı anda inceleyen ilk çalışma olarak öne çık-

maktadır. Ayrıca, ücretli ve ücretsiz modellerin performanslarının karşılaştırılması, çalışmanın literatüre olan dikkat değer bir diğer önemli katkısıdır. Bu yaklaşım, BDM'lerin mevcut sınav performanslarının analiz edilmesine ek olarak, bu modellerin eğitim süreçlerine nasıl entegre edilebileceğine dair değerli bilgiler sunmaktadır.

SONUÇ

BDM'ler yakın zamanda YZ alanındaki gelişmelere paralel olarak önemli ilerlemeler kaydetmiştir. Bu çalışma, BDM'lerin, ADÇR alanına özel yardımcı araçlar olarak kullanım potansiyelini ortaya koyarken bazı modellerin karmaşık durumlarda henüz tutarlı, güvenilir ve doğru cevaplar sağlayamadığını da göstermektedir. Bununla birlikte bu çalışmada, ücretli ve ücretsiz sürüm BDM'lerin DUS sınavındaki ADÇR sorularını cevaplama performanslarını karşılaştırılmasının mevcut literatüre değerli bir katkı sağladığı düşünülmektedir. Gelecek çalışmalarda, daha güncel BDM sürümlerinin kapsamlı bir şekilde incelenmesi ve bu doğrultuda modellerin geliştirilmesine yönelik katkı sağlanması gerektiği değerlendirilmelidir.

KAYNAKLAR

1. Botross M, Mohammadi SO, Montgomery K, Crawford C. Performance of Google's Artificial Intelligence Chatbot "Bard" (Now "Gemini") on Ophthalmology Board Exam Practice Questions. Cureus 2024; 16(3): e57348.

2. Terwilliger E, Bcharah G, Bcharah H, Bcharah E, Richardson C, et al. Advancing Medical Education: Performance of Generative Artificial Intelligence Models on Otolaryngology Board Preparation Questions With Image Analysis Insights. Cureus 2024; 16(7): e64204.

3. Rossettini G, Rodeghiero L, Corradi F, Cook C, Pillastrini P, et al. Comparative accuracy of ChatGPT-4, Microsoft Copilot and Google Gemini in the Italian entrance test for healthcare sciences degrees: a cross-sectional study. BMC Med Educ 2024; 24(1): 694.

4. Brin D, Sorin V, Vaid A, Soroush A, Glicksberg BS, et al. Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. Scientific Reports 2023; 13(1): 16492.

5. Takagi S, Watari T, Erabi A, Sakaguchi K. Performance of GPT-3.5 and GPT-4 on the Japanese Medical Licensing Examination: Comparison Study. JMIR Med Educ 2023; 9: e48002.

6. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health 2023; 2(2): e0000198.

7. Lee P, Bubeck S, Petro J. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. N Engl J Med 2023; 388(13): 1233-9.

8. Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a Radiology Board-style Examination: Insights into Current Strengths and Limitations. Radiology 2023; 307(5): e230582.

9. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, et al. How Does ChatGPT Perform on the United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge Assessment. JMIR Med Educ 2023; 9: e45312.

10. Chen Y, Huang X, Yang F, Lin H, Lin H, et al. Performance of ChatGPT and Bard on the medical licensing examinations varies across different cultures: a comparison study. BMC Med Educ 2024; 24(1): 1372.

11. Newton P, Xiromeriti M. ChatGPT performance on multiple choice question examinations in higher education. A pragmatic scoping review. Assessment & Evaluation in Higher Education 2024; 49(6): 781-98.

12. Odabashian R, Bastin D, Jones G, Manzoor M, Tangestaniapour S, et al. Assessment of ChatGPT-3.5's Knowledge in Oncology: Comparative Study with ASCO-SEP Benchmarks. Jmir ai 2024; 3: e50442.

13. Longwell JB, Hirsch I, Binder F, Gonzalez Conchas GA, Mau D, et al. Performance of Large Language Models on Medical Oncology Examination Questions. JAMA Netw Open 2024; 7(6): e2417641.

14. Antaki F, Touma S, Milad D, El-Khoury J, Duval R. Evaluating the Performance of ChatGPT in Ophthalmology: An Analysis of Its Successes and Shortcomings. Ophthalmol Sci 2023; 3(4): 100324.

15. Carlà MM, Giannuzzi F, Boselli F, Rizzo S. Testing the power of Google DeepMind: Gemini versus ChatGPT 4 facing a European ophthalmology examination. AJO International 2024; 1(3): 100063.

16. Ali R, Tang OY, Connolly ID, Zadnik Sullivan PL, Shin JH, et al. Performance of ChatGPT and GPT-4 on Neurosurgery Written Board Examinations. Neurosurgery 2023; 93(6): 1353-65.

17. Schubert MC, Wick W, Venkataramani V. Performance of Large Language Models on a Neurology Board-Style Examination. JAMA Netw Open 2023; 6(12): e2346721.

18. Haddad F, Saade JS. Performance of ChatGPT on Ophthalmology-Related Questions Across Various Examination Levels: Observational Study. JMIR Med Educ 2024; 10: e50842.

19. Turunç Oğuzman R, Yurdabakan ZZ. Performance of Chat Generative Pretrained Transformer and Bard on the Questions Asked in the Dental Specialty Entrance Examination in Turkey Regarding Bloom's Revised Taxonomy. Current Research in Dental Sciences 2024; 34(1): 25-34.

20. Giannakopoulos K, Kavadella A, Aaqel Salim A, Stamatopoulos V, Kaklamanos EG. Evaluation of the Performance of Generative AI Large Language Models ChatGPT, Google Bard, and Microsoft Bing Chat in Supporting Evidence-Based Dentistry: Comparative Mixed Methods Study. J Med Internet Res 2023; 25: e51580.