Özgün Araştırma / Original Article

Tıpta Uzmanlık Sınavındaki Oftalmoloji ile İlgili Sorularda ChatGPT-4 Omni ve Gemini 1.5 Pro'nun Performansı

Performance of ChatGPT-4 Omni and Gemini 1.5 Pro on Ophthalmology-Related Questions in the Turkish Medical Specialty Exam

Mehmet Cem Sabaner, Zübeyir Yozgat

Kastamonu Üniversitesi Tıp Fakültesi; Kastamonu Eğitim ve Araştırma Hastanesi, Göz Hastalıkları Anabilim Dalı, Kastamonu, Türkiye

Öz

Amaç: Tıpta Uzmanlık Sınavları'nda yer alan oftalmoloji ile ilgili çoktan seçmeli sorulara (ÇSS) yanıt vermede iki yapay zeka (YZ) tabanlı büyük dil modeli (BDM) platformunun yanıt ve yorumlama yeteneklerini değerlendirmek.

Gereç ve Yöntem: 2006-2024 yılları arasında yapılan toplam 37 sınavın tüm ÇSS'ları incelendi. Oftalmoloji ile ilgili sorular belirlendi ve bölümlere ayrıldı. İlk olarak, sorular ChatGPT-4o ve Gemini 1.5 Pro YZ tabanlı BDM sohbet robotlarına hem Türkçe hem de İngilizce olarak belirli komutlarla soruldu. İkinci olarak, tüm ÇSS'lar herhangi bir etkileşim gerçekleştirilmeden yeniden soruldu. Son adımda ise yanlış yanıtlar için geri bildirim oluşturuldu ve tüm sorular yeniden soruldu.

Bulgular: Her iki YZ tabanlı BDM'de de 7312 ÇSS'dan 220 oftalmoloji ile ilgili soru değerlendirildi. Otuz üç bölümün (32 tam sınav ve 2022-2024 yılları arasındaki sınavların paylaşılan %10'luk kısmı) ortalama 6,47±2,91 (aralık: 2-13) ÇSS sayısı idi. Son adımdan sonra ChatGPT-4o, Gemini 1.5 Pro'ya kıyasla hem Türkçe (%97,3) hem de İngilizce'de (%97,7) daha yüksek doğruluk oranı elde etti (%94,1 ve %93,2) ve bu fark İngilizce'de istatistiksel olarak anlamlı iken (p=0,039) ancak Türkçe'de değildi (p=0,159). Bölümlerin YZ'ler arası karşılaştırmasında veya diller arası karşılaştırmada istatistiksel olarak anlamlı bir fark yoktu.

Sonuç: Her iki YZ platformu da oftalmolojiyle ilgili ÇSS'ları ele almada güçlü bir performans göstermiş olsa da, ChatGPT-4o şimdilik bir adım önde. Bu YZ tabanlı BDM'ler, yalnızca ÇSS'ların yanıtlarını doğru bir şekilde seçerek değil, aynı zamanda ayrıntılı açıklamalar sağlayarak oftalmolojik tıp eğitimini geliştirme potansiyeline sahiptir.

Anahtar Kelimeler: Yapay zeka, büyük dil modelleri, ChatGPT-4 omni, Gemini 1.5 Pro, tıp eğitimi, oftalmoloji, e-öğrenme

Abstract

Objectives: To evaluate the response and interpretative capabilities of two pioneering artificial intelligence (AI)-based large language model (LLM) platforms in addressing ophthalmology-related multiple-choice questions (MCQs) from Turkish Medical Specialty Exams.

Materials and Methods: MCQs from a total of 37 exams held between 2006-2024 were reviewed. Ophthalmology-related questions were identified and categorized into sections. The selected questions were asked to the ChatGPT-4o and Gemini 1.5 Pro AI-based LLM chatbots in both Turkish and English with specific prompts, then re-asked without any interaction. In the final step, feedback for incorrect responses were generated and all questions were posed a third time.

Results: A total of 220 ophthalmology-related questions out of 7312 MCQs were evaluated using both AI-based LLMs. A mean of 6.47±2.91 (range: 2-13) MCQs was taken from each of the 33 parts (32 full exams and the pooled 10% of exams shared between 2022 and 2024). After the final step, ChatGPT-4o achieved higher accuracy in both Turkish (97.3%) and English (97.7%) compared to Gemini 1.5 Pro (94.1% and 93.2%, respectively), with a statistically significant difference in English (p=0.039) but not in Turkish (p=0.159). There was no statistically significant difference in either the inter-AI comparison of sections or interlingual comparison.

Conclusion: While both AI platforms demonstrated robust performance in addressing ophthalmology-related MCQs, ChatGPT-4o was slightly superior. These models have the potential to enhance ophthalmological medical education, not only by accurately selecting the answers to MCQs but also by providing detailed explanations.

Keywords: Artificial intelligence, large language model, ChatGPT-4 omni, Gemini 1.5 Pro, medical education, ophthalmology, e-learning

Cite this article as: Sabaner MC, Yozgat Z. Performance of ChatGPT-4 Omni and Gemini 1.5 Pro on Ophthalmology-Related Questions in the Turkish Medical Specialty Exam.
Turk J Ophthalmol. 2025;55:177-185

Yazışma Adresi/Address for Correspondence: Mehmet Cem Sabaner, Kastamonu Üniversitesi Tıp Fakültesi; Kastamonu Eğitim ve Araştırma Hastanesi, Göz Hastalıkları Anabilim Dalı, Kastamonu, Türkiye
E-posta: drmcemsabaner@yahoo.com **ORCID-ID:** orcid.org/0000-0002-0958-9961
Geliş Tarihi/Received: 23.08.2024 **Kabul Tarihi/Accepted:** 10.06.2025

DOI: 10.4274/tjo.galenos.2025.27895

Giriş

"Çağımızın bu harikalarını, düşünen makineleri anlamak, şeytani bir zeka gerektirmez; daha ziyade, basit sağduyu yeterlidir."[1] Düşünen makinelerin geliştirilmesi ve bunların eğitim ve iş ortamına etkin bir şekilde entegre edilmesi insanlık için önemli bir gelişmedir. Yapay zeka (YZ), başta tıp bilimleri olmak üzere eğitim platformlarında ciddi bir potansiyele sahiptir.[2,3] YZ şu anda sadece eğitim yaklaşımlarını geliştirerek değil, aynı zamanda tanı ve tedavi önerilerini geliştirerek de

çağdaş tıbba katkıda bulunmaktadır.[2,3,4,5,6,7,8,9] YZ'nın neredeyse hiçbir zaman insanların yerini alamayacağı varsayılsa da, tıp ve tıp eğitimine olası katkıları oldukça ilgi çekicidir. ChatGPT-4o (omni) ve Gemini 1.5 Pro, çok çeşitli dillerde güvenilir ve içeriğe duyarlı çıktılar oluşturmak için tasarlanmış son teknoloji modellerdir. Temel olarak, çeşitli dil özelliklerini öğrenmek için büyük veri kümeleri ile eğitilmiş gelişmiş derin öğrenme çerçeveleri olan büyük dil modelleri (BDM) olarak kategorize edilirler.[3,10] YZ tabanlı BDM sohbet robotları artık özellikle dijital eğitim, kişiselleştirilmiş sağlık hizmetleri, otonom sistemler, müşteri desteği, veri bilimi ve yazılım mühendisliği gibi birçok alanda giderek daha fazla kullanılmaktadır.[10]

YZ odaklı e-öğrenme, eğitim paradigmalarını ve uygulamalarını küresel ölçekte etkilemeye devam etmektedir.[10] Ev ödevleri yapmak, araştırma yapmak, çoktan seçmeli soruları (ÇSS) cevaplamak ve hatta akademik tezler yazmak gibi görevler artık YZ tabanlı BDM'ler kullanılarak verimli bir şekilde yapılabilmektedir.[11] Bununla birlikte, bu YZ odaklı yaklaşımların güvenilirliği ve etkinliği incelenmeye devam etmektedir. Önceki araştırmalarda, YZ'nın ÇSS'ları çözebilme potansiyeli gösterilmiş ve doğru bilgiye ulaşmayı kolaylaştırmaki rolüne işaret edilmiştir.[6] Bununla birlikte, oftalmoloji ile ilgili ÇSS'ları yanıtlamada YZ sohbet robotlarının yeteneklerini özel olarak değerlendiren çalışmaların sayısı azdır.[12,13,14,15,16,17] Sonuç olarak, bu çalışmada, hem İngilizce hem de Türkçe olarak oftalmoloji ile ilgili ÇSS'lar için sohbet robotlarının nasıl etkili bir şekilde kullanılabileceğinin araştırması amaçlanmıştır. Bu amaçla çalışmada bu iki sohbet robotunun Türk Tıpta Uzmanlık Sınavı'nda (TUS) çıkan ilgili ÇSS'lara verdikleri yanıtlar değerlendirilmiştir.

Gereç ve Yöntem

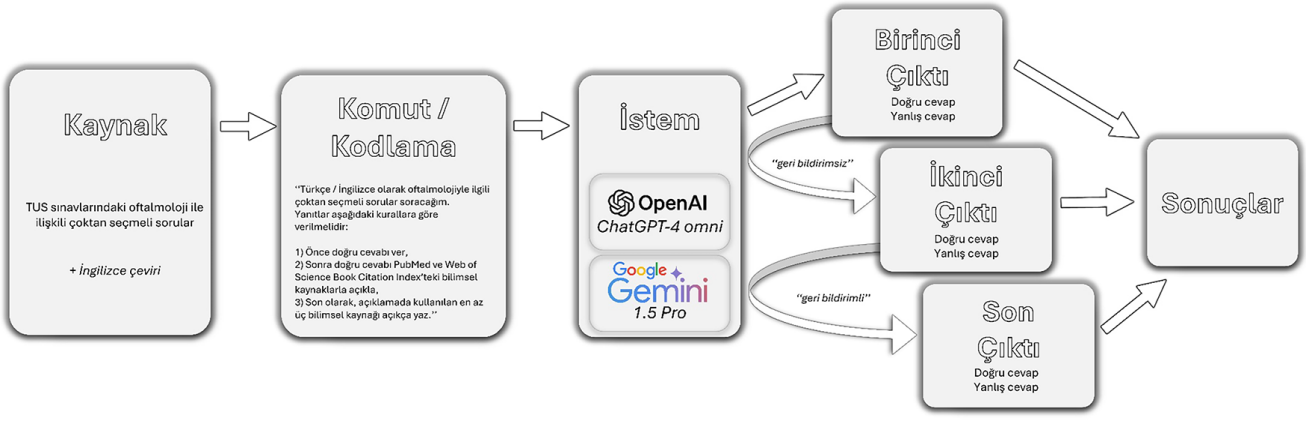
Çalışma Tasarımı ve Veri Toplama

Bu kesitsel çalışmada, geçmiş TUS sınavlarında çıkan oftalmoloji ile ilgili ÇSS'ları yanıtlamada iki YZ tabanlı BDM sohbet robotu modelinin performansı değerlendirilmiştir. TUS, Türkiye Öğrenci Seçme ve Yerleştirme Merkezi (ÖSYM) tarafından tıbbi uzmanlık eğitimine kabul için yılda iki kez yapılan ülke çapında standartlaştırılmış bir sınavdır. TUS, temel ve klinik tıp bilimleri testleri olmak üzere iki bölümden oluşmaktadır.

2006-2021 yılları arasında yapılan toplam 32 sınavda[18] yer alan tüm sorular ve 2022-2024 yılları arasında yapılan 5 sınavda[19,20,21] yer alan soruların %10'unun telif hakları ÖSYM tarafından Fikir ve Sanat Eserleri Kanunu kapsamında eser olarak alınmış ve çoğaltma, dağıtım ve yeniden yayınlama kısıtlamaları ile açık erişim olarak kullanıma sunulmuştur. Bu sorular iki kıdemli göz hekimi tarafından detaylı olarak gözden geçirildi. Oftalmoloji ile ilgili sorular fikir birliği ile belirlendi ve çalışmaya dahil edildi. Bu sorular ayrıca oftalmoloji alt başlıkları altında sınıflandırıldı.

Sohbet robotlarının oftalmoloji ile ilgili ÇSS'ları yanıtlama kapasitesi, en güncel premium sürümlerine Gemini Advanced ve ChatGPT Plus platformları aracılığıyla erişilen Gemini 1.5 Pro (Google, Mountain View, CA, ABD) ve ChatGPT-4o (OpenAI, San Francisco, CA, ABD) kullanılarak değerlendirildi. İstemler ve yanıt değerlendirmesini içeren genel etkileşim süreci Şekil 1'de özetlenmiştir. Her sohbet robotu oturumuna standart bir istem ile başlandı ve modelden üç adımlı bir formatı izleyerek ÇSS'ları Türkçe veya İngilizce olarak yanıtlaması istendi: (1) Doğru cevabı belirtin, (2) Web of Science (WoS) Citation Index ve PubMed'de indekslenen bilimsel kaynakları kullanarak cevabı gerekçelendirin ve (3) en az üç atıf listeleyin. Görsel veri içeren sorular için sohbet robotlarının resim yükleme özelliklerinden yararlanıldı. Değerlendirme için üç farklı girişimde bulunuldu.

İlk denemede, seçilen tüm oftalmoloji ile ilgili ÇSS'ler, Türkçe'den başlayarak sohbet robotlarına tek tek sunuldu. Yanıtların doğru mu yanlış mı olduğu konusunda herhangi bir geri bildirim verilmedi. Aynı sorular daha sonra profesyonel olarak İngilizceye çevrildi, ardından iki araştırmacı tarafından geri çeviri ve çapraz doğrulama yapıldı. Dilsel yapının getirdiği yanlılığı en aza indirmek için çeviri sırasında cevap seçenekleri yeniden sıralandı. İngilizce sorular, yanıtların doğru olup olmadığı konusunda herhangi bir geri bildirim verilmeden sohbet botlarına tek tek soruldu.



Şekil 1. Çalışmanın akış şeması
ÇSS: Çoktan seçmeli sorular, TUS: Tıpta Uzmanlık Sınavları

İkinci denemede, her iki dilde de daha önce kullanılan tüm sorular geri bildirim verilmeden tekrar girildi.

Son denemede, sohbet robotu yanıtlarından hatalı olanlar "başparmak aşağı" simgesiyle işaretlendi ve neden olarak "Gerçekte doğru değil" seçildi. Daha sonra tüm sorular yeniden değerlendirilmek üzere bir kez daha sunuldu.

Her deneme yeni bir sohbet robotu oturumunda gerçekleştirildi. Her denemede, yanıtların doğruluğu yalnızca resmi yanıt anahtarlarına göre değerlendirildi. Bir yanıtın, açıklamanın niteliğine bakılmaksızın doğru olarak işaretlenmesi için doğru seçeneğin seçilmesi gerekiyordu. Aksine, yanlış bir seçenek seçildiyse, doğru bir açıklama olsa bile, cevap yanlış kabul edildi. Her deneme için doğruluk oranı, doğru cevapların yüzdesi olarak hesaplandı.

Atıf yapılan kaynaklar da dahil olmak üzere sohbet robotları tarafından oluşturulan her açıklama, iki kıdemli göz doktoru tarafından bağımsız olarak değerlendirildi. Her yanıtın istenen oftalmolojik bilgiyle ilgili olup olmadığını değerlendirmek için 4'lü Likert ölçeği kullanıldı: 1 = İlgili değil, 2 = Biraz ilgili, 3 = Oldukça ilgili ve 4 = Son derece ilgili. Dört noktadan üçü açıklamanın bilimsel temelini değerlendirmek için belirlenirken, diğer nokta atıfta bulunulan referansların güvenilirliğine odaklanmıştır. Yayın tarihleri veya internet bağlantı adreslerinin hatalı olması gibi küçük yanlışlar cezalandırılmadı; ancak puanlama sırasında yazar adları, makale başlıkları veya dergi kaynaklarındaki tutarsızlıklar dikkate alındı. İki veya daha fazla referans eksik veya hatalıysa, puan kesintisi yapıldı. Madde düzeyinde kapsam geçerlik indeksi (M-KGİ), her maddeye 3 veya 4 puan veren hakemlerin oranı olarak hesaplandı.[22] Genel geçerliliği değerlendirmek için, her iki sohbet robotunun her girişim için tüm maddelere verdiği yanıtların M-KGİ puanlarının ortalaması alınarak ortalama KGİ değerleri hesaplandı. Polit ve Beck[22] tarafından belirlenen kriterler izlenerek ortalama KGİ değerinin ≥0,80 olması kapsam geçerliliğinin kabul edilebilir düzeyde olduğu şeklinde yorumlandı.

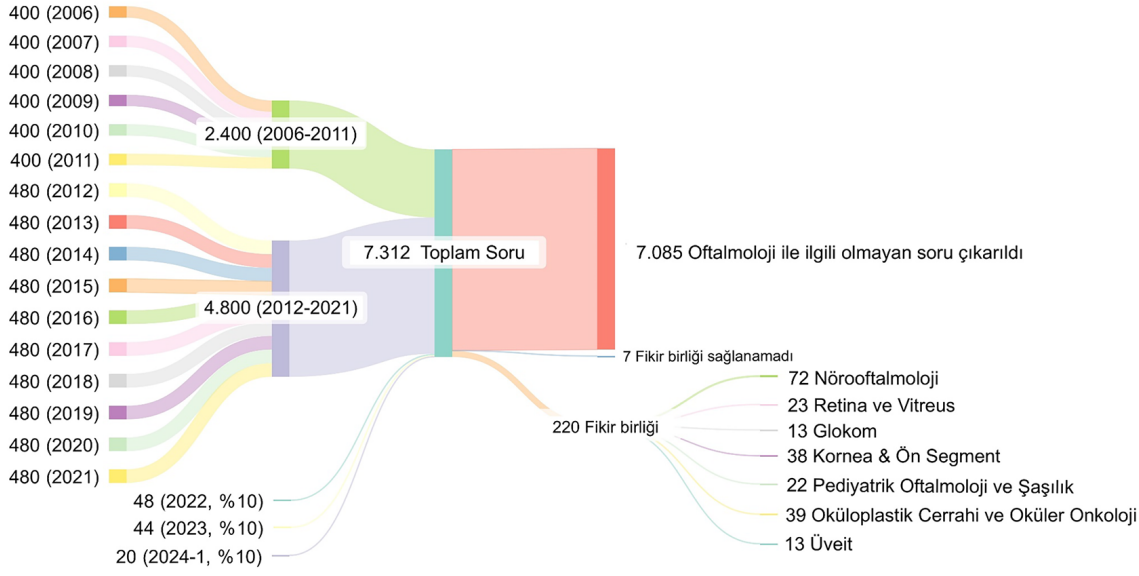
Bu YZ tabanlı BDM sohbet robotu değerlendirme çalışması için bilgilendirilmiş onam ve yerel etik kuruldan onay alınması gerekmedi.

İstatistiksel Analiz

Verileri GraphPad Prism (v10,2.3, San Diego, CA, ABD) ve IBM SPSS Statistics (v22,0, Armonk, NY, ABD) yazılımları kullanılarak analiz edildi. Soru akışını ve kategorizasyonu gösteren Sankey diyagramı, SankeyMATIC çevrimiçi aracı kullanılarak oluşturuldu. Tanımlayıcı istatistikler, ortalama ± standart deviasyon (SD) veya medyan ve çeyrekler açıklığı (25.-75. persentil) değerlerinden uygun olanı kullanılarak sunuldu. Kategorik değişkenler için öncelikle Pearson ki-kare testi kullanıldı. Ancak, frekans sayımlarının varsayımları karşılanmadığı durumlarda Fisher kesin olasılık testi veya Yates süreklilik düzeltmesi (örneğin; beklenen sayısı sırasıyla <5 veya 5-25 ise) kullanıldı. Dörtten fazla kategorik grubu içeren karşılaştırmalarda, analizler varsayılan yöntem olarak Pearson ki-kare testi ile yapıldı. İki sohbet robotu arasında açıklamaların kelime sayısındaki farklılıklar, verilerin normal olmayan dağılımı göz önüne alınarak, parametrik olmayan Mann-Whitney U testi ile analiz edildi. Likert ölçekli değerlendirmelerde denemeler arasındaki tutarlılığı ve hakemlerin uyumunu değerlendirmek için sınıf içi korelasyon katsayıları hesaplandı. İstatistiksel anlamlılık p değerinin 0,05'ten küçük olması olarak tanımlandı ve tüm analizler %95 güven aralığında (GA) yapıldı.

Bulgular

Geçmiş 37 TUS sınavında yer alan 7312 ÇSS'dan toplam 220 soru oftalmoloji ile ilgili bulundu ve daha ileri analiz için seçildi. Soru seçim süreci ve yan dal dağılımına ilişkin detaylı bilgiler Şekil 2'de görselleştirilmiştir. ÖSYM'nin telif hakkı kısıtlamaları nedeniyle soru ve cevapların tam metni yayınlanamadı. Bununla birlikte, çalışmaya dahil edilen sorularla



Şekil 2. Tıbbi Uzmanlık Sınavı soru seçimi ve alt uzmanlık dağılımını gösteren Sankey çalışma diyagramı

ilgili ayrıntılara Ek 1'de yer verilmiştir. TUS sorularına benzer sorular ve sohbet robotu cevap örnekleri Tamamlayıcı Bilgi'de sunulmuştur. Nöro-oftalmoloji en sık soru sorulan alt uzmanlık alanıyken (n=72), glokom ve üveit hakkında en az (n=13) soru sorulan alanlardı. Değerlendirilen sınavlarda (32 sınavın tamamı ile 2022-2024 yılları arasında yapılan sınavların paylaşılan %10'luk bölümü), ortalama oftalmoloji soru sayısı minimum 2 ve maksimum 13 olmak üzere 6,47 (SD=2,91) idi (Şekil 3).

Dile (Türkçe ve İngilizce) ve YZ modeline göre sınıflandırılmış her üç girişimde elde edilen doğruluk oranları ayrıntılı olarak Tablo 1'de sunulmuştur. Son denemede, ChatGPT-4o ile elde edilen doğruluk oranı, Gemini 1.5 Pro'ya (sırasıyla %94,1 ve %93,2) kıyasla hem Türkçe'de (%97,3) hem de İngilizce'de (%97,7) daha yüksekti. Bu fark İngilizce için istatistiksel anlamlılığa ulaşırken (p=0,039), Türkçe'de anlamlı değildi (p=0,159). Her iki model için de ardışık girişimlerde doğrulukta aşamalı bir artış gözlenmesine rağmen, bu değişiklikler istatistiksel olarak anlamlı değildi (p>0,05). Genel analizde (n=220), ChatGPT-4o, doğru yanıt sayısı açısından tüm girişimlerde üstün performans göstermiştir. Türkçede, ChatGPT-4o ile 209, 210 ve 214 doğru cevap elde edilirken, Gemini 1.5 Pro ile sırasıyla 202, 204 ve 207 doğru yanıt alındı (tüm karşılaştırmalar için p>0,05). İngilizce'de, aradaki fark son denemede istatistiksel anlamlılığa ulaştı (215'e kıyasla 205; p=0,039), ancak daha önceki denemelerde anlamlı fark yoktu (p=0,312).

Oftalmolojik alt uzmanlık alanları tek tek değerlendirildiğinde, iki YZ platformu arasında istatistiksel olarak anlamlı bir fark gözlenmedi. Ayrıca, aynı soru grubu Türkçe ve İngilizce olarak yanıtlandığında her iki model için de istatistiksel olarak anlamlı bir diller arası varyasyon izlenmedi. Sınav yıllarına göre detaylı performans dağılımı Tablo 2'de verilmiştir.

Değerlendirilen ÇSS sorulardan sadece ikisinde görsel vardı. Her iki sohbet robotu da bu soruları her iki dilde de doğru bir şekilde yanıtladı ve test edilen koşullar altında görsel yorumlama yeteneklerinin yeterliydi.

Ortalama KGİ değerleri ile değerlendirilen içerik geçerliliği, her iki sohbet robotu için tüm girişimlerde ve her iki dilde yüksek uyum gösterdi (Tablo 3). Bu iyi sonuçlara rağmen, her iki model de zaman zaman gerçekle örtüşmeyen halüsinasyonlar gördü veya fabrikasyon kaynaklar üretti. Yazar adları veya dergi başlıklarının uyumsuz olması gibi durumlar, M-KGİ puanlaması sırasında sistematik olarak hesaplamaya dahil edildi.

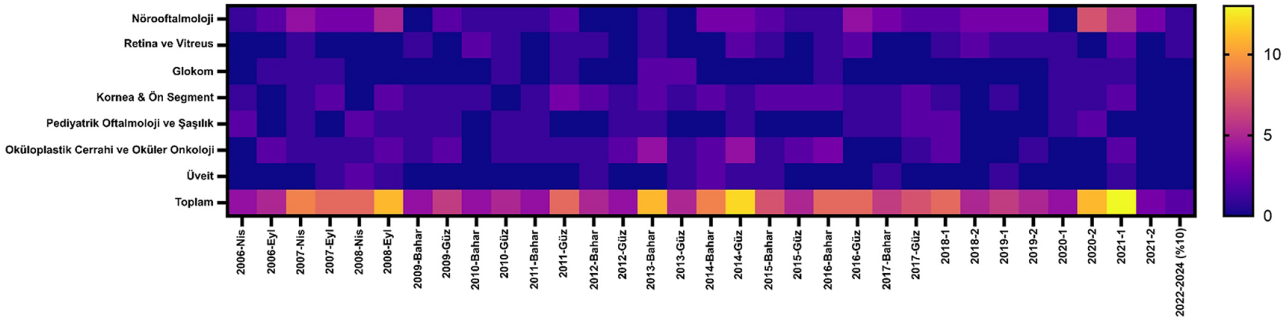
Açıklama uzunluğu açısından, her iki modelin Türkçe ve İngilizce cevapları arasında istatistiksel olarak anlamlı farklılık olduğu bulundu. İngilizce olarak oluşturulan açıklamalar, Türkçe karşılıklarından anlamlı düzeyde daha uzundu (ChatGPT-4o: medyan 178'e kıyasla 88 kelime; Gemini 1.5 Pro: medyan 124'e kıyasla 81,5 kelime; tüm karşılaştırmalar, p<0,001). Ayrıca, her iki dilde de ChatGPT-4o, Gemini 1.5 Pro'dan daha uzun yanıtlar üretti (hem Türkçe hem de İngilizce karşılaştırmalar için p<0,001).

Girişimler arasındaki yanıt tutarlılığını değerlendirmek için, her YZ modeli için Cohen kappa (κ) değerleri hesaplandı. Türkçede κ değerleri ChatGPT-4o için 0,974 (%95 GA, 0,967-0,980) ve Gemini 1.5 Pro için 0,967 (%95 GA, 0,957-0,975) idi. İngilizce'de, her iki model de κ değeri, mükemmel uyuma işaret eden, 1,000 bulundu. Bu sonuçlar, sohbet robotlarına geri bildirim verilmeyen birinci ve ikinci girişimler arasında mükemmele yakın tekrarlanabilirlik olduğunu göstermektedir.

Tartışma

Bu çalışma, son teknoloji BDM sohbet robotlarının oftalmoloji ile ilgili ÇSS'lere hem Türkçe hem de İngilizce olarak yüksek doğruluk seviyelerinde yanıt verebildiğini göstermektedir. Özellikle, ChatGPT-4o, İngilizce olarak yapılan son değerlendirme denemesinde Gemini 1.5 Pro'dan daha iyi performans gösterdi ve aradaki fark istatistiksel olarak anlamlıydı. Bu farklılığa rağmen, her iki YZ platformu da farklı dillerle yapılan farklı girişimlerde iyi bir performans sergiledi. Bu, oftalmoloji eğitimi ve değerlendirmesinde ek araçlar olarak kulanılma potansiyelleri olduğuna işaret etmektedir.

Oftalmolojinin diğer tıp disiplinlerinden farklı konumu ve nispeten izole olması nedeniyle, oftalmolojik sorular sağlık çalışanları için zor olabilir. Buna paralel olarak, sağlık çalışanlarının güncel oftalmolojik bilgilere ulaşmak için çevrimiçi kaynaklara yöneliminin artması, YZ tabanlı BDM'lerin tıp eğitimindeki öneminin artmakta olduğunu göstermektedir. Bu modeller, anında, yapılandırılmış ve referans destekli yanıtlar sağlayarak geleneksel eğitimi destekleyebilen araçlar olarak ortaya çıkmaktadır ve dijital öğrenmede yarattıkları dönüştürücü etki hızla kabul görmeye başlamıştır Böylece, YZ tabanlı



Şekil 3. Tıpta Uzmanlık Sınavı sorularının yıllara göre alt uzmanlık alanlarına dağılımının ısı haritası tablosu

Tablo 1. Sohbet robotlarının tüm girişimlerde bölüm sorularına verdikleri yanıtların diller arası ve modeller arası karşılaştırması

	n	Doğru cevap sayısı ve doğruluk oranı (%)												p*			
		ChatGPT-4o						Gemini 1.5 Pro									
		TR			EN			TR			EN			YZ'lar arası		Diller arası	
		Birinci	İkinci	Son	Birinci	İkinci	Son	Birinci	İkinci	Son	Birinci	İkinci	Son	TR	EN	ChatGPT-4o	Gemini 1.5 Pro
Nöro-oftalmoloji	**72**	67 (93,1)	68 (94,4)	70 (97,2)	66 (91,7)	66 (91,7)	70 (97,2)	65 (90,3)	65 (90,3)	67 (93,1)	64 (88,9)	64 (88,9)	64 (88,9)	0,763 0,745 0,441	0,779 0,779 0,097	>0,99 0,745 >0,99	>0,99 0,779 0,561
Retina ve vitreus	**23**	22 (95,7)	22 (95,7)	23 (100)	23 (100)	23 (100)	23 (100)	22 (95,7)	22 (95,7)	22 (95,7)	22 (95,7)	22 (95,7)	22 (95,7)	>0,99 >0,99 >0,99	>0,99 >0,99 >0,99	>0,99 >0,99 >0,99	>0,99 >0,99 >0,99
Glokom	**13**	12 (92,3)	12 (92,3)	13 (100)	13 (100)	13 (100)	13 (100)	12 (92,3)	12 (92,3)	13 (100)	12 (92,3)	12 (92,3)	13 (100)	>0,99 >0,99 >0,99	>0,99 >0,99 >0,99	>0,99 >0,99 >0,99	>0,99 >0,99 >0,99
Kornea ve ön segment	**38**	38 (100)	38 (100)	38 (100)	38 (100)	38 (100)	38 (100)	36 (94,7)	36 (94,7)	36 (94,7)	38 (100)	38 (100)	38 (100)	0,493 0,493 0,493	>0,99 >0,99 >0,99	>0,99 >0,99 >0,99	0,493 0,493 0,493
Pediatrik göz hastalıkları ve şaşılık	**22**	21 (95,5)	21 (95,5)	21 (95,5)	21 (95,5)	21 (95,5)	21 (95,5)	20 (90,9)	20 (90,9)	20 (90,9)	20 (90,9)	20 (90,9)	20 (90,9)	>0,99 >0,99 >0,99	>0,99 >0.99 >0,99	>0,99 >0,99 >0,99	>0,99 >0,99 >0,99
Oküloplasti ve oküler onkoloji	**39**	37 (94,9)	37 (94,9)	37 (94,9)	38 (97,4)	38 (97,4)	39 (100)	36 (92,3)	36 (92,3)	37 (94,9)	37 (94,9)	37 (94,9)	37 (94,9)	>0,99 >0,99 >0,99	>0,99 >0,99 0,494	>0,99 >0,99 0,494	>0,99 >0,99 >0,99
Üveit	**13**	12 (92,3)	12 (92,3)	12 (92,3)	11 (84,6)	11 (84,6)	11 (84,6)	11 (84,6)	12 (92,3)	12 (92,3)	11 (84,6)	11 (84,6)	11 (84,6)	>0,99 >0,99 >0,99	>0,99 >0,99 >0,99	>0,99 >0,99 >0,99	>0,99 >0,99 >0,99
Total	**220**	209 (95)	210 (95,5)	214 (97,3)	21 (95,5)	210 (95,5)	215 (97,7)	202 (91,8)	203 (92,3)	207 (94,1)	204 (92,7)	204 (92,7)	205 (93,2)	0,249 0,233 0,159	0,312 0,312 **0,039**	>0,99 >0,99 >0,99	0,858 >0,99 0,845

TR: Türkçe, EN: İngilizce, YZ: Yapay zeka, BDM: Büyük dil modelleri. YZ tabanlı BDM'ler alfabetik sıraya göre listelenmiştir. Birinci, ikinci ve son girişimler için diller arası ve YZ'ler arası karşılaştırmaların p değerleri sıralanmıştır. *Dört değişkenden en az birinin beklenen frekansı 5'in altında ise "Fisher kesin olasılık testi ", 5 ile 25 arasında ise "Yates süreklilik düzeltmeli ki-kare testi" kullanılmıştır. %95 güven aralığında $p<0,05$ istatistiksel olarak anlamlı kabul edilmiştir

Tablo 2. ChatGPT-4o ve Gemini 1.5 Pro'nun sınav yıllarına göre değerlendirilmesi: Dil ve model karşılaştırmaları

Yıl	n	Doğru cevap sayısı ve doğruluk oranı (%)								p*			
		TR				EN				Diller arası		YZ'lar arası	
		ChatGPT-4o		Gemini 1.5 Pro		ChatGPT-4o		Gemini 1.5 Pro		ChatGPT-4o	Gemini 1.5 Pro	TR	EN
		Birinci	İkinci	Birinci	İkinci	Birinci	İkinci	Birinci	İkinci	Son	Son				
2006	**9**	9 (100)	9 (100)	9 (100)	9 (100)	9 (100)	9 (100)	9 (100)	9 (100)	9 (100)	9 (100)				
2007	**17**	17 (100)	17 (100)	15 (88,2)	15 (88,2)	17 (100)	17 (100)	15 (88,2)	15 (88,2)	15 (88,2)	15 (88,2)				
2008	**19**	19 (100)	19 (100)	18 (94,7)	18 (94,7)	19 (100)	19 (100)	19 (100)	19 (100)	19 (100)	19 (100)				
2009	**10**	10 (100)	10 (100)	9 (90)	9 (90)	10 (100)	10 (100)	10 (100)	10 (100)	10 (100)	10 (100)				
2010	**9**	9 (100)	9 (100)	9 (100)	9 (100)	9 (100)	9 (100)	9 (100)	9 (100)	9 (100)	9 (100)				
2011	**12**	12 (100)	12 (100)	12 (100)	12 (100)	12 (100)	12 (100)	12 (100)	12 (100)	12 (100)	12 (100)				
2012	**9**	7 (77,8)	7 (77,8)	8 (88,9)	8 (88,9)	8 (88,9)	8 (88,9)	8 (88,9)	8 (88,9)	8 (88,9)	8 (88,9)				
2013	**16**	16 (100)	16 (100)	15 (93,8)	15 (93,8)	16 (100)	16 (100)	15 (93,8)	15 (93,8)	15 (93,8)	15 (93,8)				
2014	**21**	19 (90,5)	19 (90,5)	21 (100)	21 (100)	20 (95,2)	20 (95,2)	19 (90,5)	19 (90,5)	19 (90,5)	19 (90,5)	>0,99	>0,99	>0,99	>0,99
2015	**12**	11 (91,7)	11 (91,7)	12 (100)	12 (100)	10 (83,3)	10 (83,3)	10 (83,3)	10 (83,3)	10 (83,3)	10 (83,3)	>0,99	>0,99	>0,99	>0,99
2016	**16**	15 (93,8)	15 (93,8)	13 (81,3)	13 (81,3)	15 (93,8)	15 (93,8)	14 (87,5)	14 (87,5)	14 (87,5)	14 (87,5)	>0,99	>0,99	>0,99	>0,99
2017	**13**	13 (100)	13 (100)	12 (92,3)	12 (92,3)	12 (92,3)	12 (92,3)	11 (84,6)	11 (84,6)	11 (84,6)	11 (84,6)				
2018	**13**	13 (100)	13 (100)	11 (84,6)	11 (84,6)	13 (100)	13 (100)	12 (92,3)	12 (92,3)	12 (92,3)	12 (92,3)				
2019	**11**	10 (90,9)	11 (100)	8 (72,7)	9 (81,8)	10 (90,9)	10 (90,9)	10 (90,9)	10 (90,9)	10 (90,9)	10 (90,9)				
2020	**15**	13 (86,7)	13 (86,7)	14 (93,3)	15 (100)	13 (86,7)	13 (86,7)	15 (100)	15 (100)	15 (100)	15 (100)				
2021	**16**	14 (87,5)	14 (87,5)	14 (87,5)	14 (87,5)	15 (93,8)	15 (93,8)	14 (87,5)	14 (87,5)	14 (87,5)	15 (93,8)				
2022-2024 (10%)	**2**	2 (100)	2 (100)	2 (100)	2 (100)	2 (100)	2 (100)	2 (100)	2 (100)	2 (100)	2 (100)				

TR: Türkçe, EN: İngilizce, YZ: Yapay zeka, BDM: Büyük dil modelleri. *2016-2024 yılları arasındaki karşılaştırmalar ki-kare testi ile yapılmıştır. YZ tabanlı BDM'ler alfabetik sınava göre listelenmiştir. Diller arası ve YZ'lar arası karşılaştırmaların p değerleri birinci, ikinci ve son girişimler için sınayla listelenmiştir. %95 güven aralığında p<0,05 istatistiksel olarak anlamlı kabul edilmiştir

sohbet robotları, öğrencilerin farklı dil ve konularda karmaşık ÇSS'leri yorumlamalarına yardımcı olabilecek erişilebilir destek mekanizmaları olmaya başlamıştır.

Son birkaç yılda, tıp eğitiminde ChatGPT ve Gemini gibi BDM sohbet robotlarına ilgi giderek artmaktadır.[5,6,7,8,9] ChatGPT-3.5 ve Google Bard gibi ilk sürümlerin, TUS, Amerika Birleşik Devletleri Tıbbi Lisanslama Sınavı ve özel oftalmoloji soru bankaları gibi farklı kaynaklarda yer alan sorularda orta düzeyde başarı sağlayarak %50 ila %70 arasında değişen oranlarda doğru yanıt verdikleri bildirilmiştir.[15,23,24,25,26] Bu sonuçlar, umut verici olmakla birlikte, akıl yürütme derinliği, alana özgü hassasiyet ve çok dilli güvenilirlikte kısıtlılıklar olduğuna işaret etmektedir.

ChatGPT-4 ve Gemini 1.5 Pro gibi yeni modeller ortaya çıktıkça belirgin bir gelişme görülmüştür. Bazı çalışmalarda, özellikle yapılandırılmış, çoktan seçmeli sınav formatlarında, tıbbi yeterlilik sınavı veya oftalmoloji kurulu yeterlilik sınavı gibi testlerde çoğunlukla %70'i geçen ve bazı durumlarda %90'ı aşan skorlar elde edildiği bildirilmiştir.[13,14,16,17,27]

Yine de, mevcut araştırmaların çoğu açık uçlu sorulara veya genel tıbbi içeriğe odaklanmıştır. Çok az sayıda çalışmada, son derece uzmanlaşmış ve görsel bir alan olan oftalmoloji alanı araştırılmıştır. Bu çalışmaların da daha az bir bölümü bu modellerin farklı dillerde nasıl performans gösterdiğini değerlendirmiştir. Bu çalışma, ChatGPT-4o ve Gemini 1.5 Pro'yu, bilimsel gerekçe ve atıf gerektiren standartlaştırılmış istemler kullanarak, iki dilli (Türkçe ve İngilizce) oftalmoloji ile ilgili ÇSS'larda doğrudan karşılaştırarak bu boşluğu gidermek için tasarlanmıştır. Amacımız sadece modelin doğruluğunu değerlendirmek değil, aynı zamanda özel bir klinik

Tablo 3. Sohbet robotları tarafından oluşturulan açıklamaların geçerliliğinin değerlendirilmesi				
			Ortalama KGİ	SKK (%95 GA)
ChatGPT-4o	TR	Birinci	0,95	0,849 (0,756-0,901)
		İkinci	0,96	0,850 (0,774-0,897)
		Son	0,97	0,834 (0,753-0,885)
	EN	Birinci	0,96	0,951 (0,936-0,963)
		İkinci	0,96	0,942 (0,924-0,956)
		Son	0,98	0,885 (0,850-0,912)
Gemini 1.5 Pro	TR	Birinci	0,93	0,862 (0,771-0,911)
		İkinci	0,94	0,878 (0,820-0,915)
		Son	0,95	0,850 (0,757-0,902)
	EN	Birinci	0,92	0,927 (0,893-0,949)
		İkinci	0,93	0,925 (0,890-0,947)
		Son	0,93	0,918 (0,877-0,944)
TR: Türkçe, EN: İngilizce, KGİ: Kapsam geçerlilik indeksi, SKK: Sınıf içi korelasyon katsayısı, GA: Güven aralığı				

alanda YZ destekli öğrenmenin pedagojik ve dilsel boyutlarını araştırmaktı. İlginç bir şekilde, bulgularımızda doğruluk oranlarının göreceli olarak yüksek olması, çalışmamızın çeşitli metodolojik açıdan güçlü yönlerine bağlanabilir. İlk olarak, her iki YZ platformunun en yeni sürümlerini kullandık. Bu platformların her biri ChatGPT-3.5 veya Google Bard gibi önceki sürümlere kıyasla oldukça gelişmiştir. İkincisi, sadece doğru cevapları değil, aynı zamanda kanıta dayalı akıl yürütmeyi de talep eden yapılandırılmış istemler kullanılmış olması, model çıktılarının kalitesini artırmış olabilir. Üçüncüsü, çalışmanın iki dilli tasarım, eğitim sırasında yeterince temsil edilmeyen dillerde modelin davranışına ilişkin değerli bilgiler sunan diller arası kontrollü karşılaştırmayı mümkün kılmıştır. Teknolojinin kullanımı, istem komutlarının titizlikle hazırlanması ve iki dilin kullanılmış olması, bu çalışmayı önceki çalışmalardan farklı kılmaktadır ve BDM'lerin tıp eğitiminde araç olarak kullanılabileceğini düşündürmektedir.

Yıllar boyunca yapılan sınavları değerlendirdiğimizde, YZ'lar arası ve diller arası farklılığın olmadığını bulduk, ancak son denemede tüm İngilizce ÇSS'ler için YZ'lar arası karşılaştırmada anlamlı fark vardı. Bu sonuçlar çelişkili olarak görülmemelidir, çünkü muhtemelen yıllar içinde yapılan sınavların soru türlerinin heterojen dağılımı ve zorluk seviyeleri bu sonuçta etkili olmuştur.

Bu çalışmadaki ilgi çekici bulgulardan biri, kullanıcı geri bildiriminin sohbet robotu performansı üzerindeki etkisidir. Her iki model de test sırasında insanlara göre geleneksel anlamda gerçekten "öğrenmezken", hem ChatGPT-4o hem de Gemini 1.5 Pro, yanlış cevaplar için standartlaştırılmış bir negatif geri bildirim sinyali aldıktan sonra son denemelerinde ılımlı gelişme gösterdiler. Bu önemli bir soruyu gündeme getiriyor: BDM'ler, kalıcı bellek olmasa bile, yanıtlarını yapılandırılmış ipuçlarına göre ne ölçüde değiştiriyor? Bu gözlemler, bu modellerin olgusal tutarlılığa ve bağlamsal akıl yürütmeye nasıl öncelik vereceklerini kontrol eden temel bir eğitim mekanizması olan insan geri bildirimiyle pekiştirmeli öğrenmenin etkisine işaret edebilir.[28,29] Kullanıcı etkileşimi sırasında gerçek zamanlı öğrenme gerçekleşmese de, bir yanıta verilen "gerçekte yanlış" gibi geri bildirim sinyalleri, modelin odağını geçici olarak daha temkinli, kanıta dayalı akıl yürütme kalıplarına kaydırabilir.[28,29,30] Pratik açıdan, bu basit düzeltme bile, özellikle tıp gibi yüksek riskli alanlarda, sohbet robotunu daha doğru ve akademik temelleri olan bir cevaba yönlendirebileceğini göstermektedir. Antaki ve ark.[15] tarafından daha önce vurgulandığı gibi, BDM'lerin eğitimsel değeri sadece doğru cevaplar üretme yeteneklerinde değil, aynı zamanda akıl yürütme ve düşünmeyi kolaylaştırma potansiyellerinde yatmaktadır. Bu, tıp alanında eğitim veren ve sınav hazırlayan kişilere yeni olanaklar sunar. Dikkatli bir şekilde kullanılırsa, kontrollü geri bildirimler, yalnızca statik yanıtlar için değil, eleştirel düşünmeyi ve yinelemeli öğrenmeyi teşvik eden araçlar olarak da sohbet robotlarının pedagojik rolünü artırabilir.

Geçerlilik analizi, her iki sohbet robotunun da Türkçe ve İngilizce olarak çıktı içeriğinin tatmin edici düzeyde olduğunu göstermiştir. Uzman değerlendirmelerinden tutarlı şekilde yüksek puan almaları bunun göstergesidir. Özellikle, İngilizce olarak oluşturulan açıklamalar, her iki model için de Türkçe açıklamalardan daha ayrıntılıydı, bu da kullanıcıların İngilizce olarak etkileşime girdiklerinde daha zengin içeriğe erişebileceklerini düşündürmektedir. Özellikle ChatGPT-4o, her iki dilde de daha uzun ve daha kapsamlı yanıtlar verdi. Bu nedenle daha detaylı yanıtlar almak isteyen öğrencilerin ChatGPT-4o tercih etmesi uygun olabilir. Ayrıca, her iki modelin de yanıtlarında sıklıkla alternatif seçeneklerin neden yanlış olduğuna dair kısa yorumlar mevcuttu. Çeldiricilerin açıklanmasına yönelik bu uygulama, çoktan seçmeli değerlendirmelerin altında yatan akıl yürütme sürecinin daha iyi anlaşılmasını teşvik ederek sohbet robotu etkileşimlerinin eğitimsel değerini artırabilir.

İnsanların parmaklarının ucunda kolayca erişilebilen YZ tabanlı BDM sohbet robotu teknolojisi hızla gelişmeye devam etmekte ve oftalmoloji alanında da bu gelişme sürmektedir.[31,32,33]

Örneğin; ChatGPT'nin önceki sürümleri Eylül 2021'e kadar olan bilgilerle sınırlıydı.[34,35,36] Ancak, en son güncellemelerle birlikte ChatGPT, internette gezinme ve güncel içeriğe ulaşma yeteneği kazanarak doğruluk oranlarını kademeli olarak iyileştirebileceğini gösterdi. Bu gelişme araştırma açısından umut verici olsa da, aynı zamanda yayınların hızla değerini kaybetmesi gibi bir dezavantajı da beraberinde getirmektedir.[35] Ayrıca, aynı sohbet robotunun farklı sürümleri arasında tutarsızlık olmasına da yol açabilir. Bu da akademik ve profesyonel ortamda sohbet robotlarının tutarlı ve güvenilir şekilde kullanılabilmesine engel teşkil edebilir. Bu önemli gelişmelere rağmen, sohbet robotları hala halüsinasyon görebilir ve gerçekte olmayan kaynaklar üretmeye çalışabilir.[35] Bu nedenle, bu tür araçları kullanırken yanıtların dikkatle değerlendirilmesi, güvenilirliği sağlamak için esastır.

Çalışmanın Kısıtlılıkları

Bu çalışma, TUS sınavlarında çıkan oftalmoloji ile ilgili ÇSS'lerin değerlendirilmesinde ChatGPT-4o'nun Gemini 1.5 Pro'ya göre üstün olduğu yönleri açıklığa kavuşturmuş olsa da bazı kısıtlılıkları da mevcuttur. 1) Performans yalnızca Türkçe ve İngilizce dillerinde değerlendirilmiştir, 2) Açık uçlu soru performansı değerlendirilmemiştir ve 3) Birçok başka model mevcut olmasına rağmen yalnızca iki YZ tabanlı BDM kullanılmıştır. Bu çalışmanın bir diğer önemli kısıtlılığı, pratisyen hekimlerin temel oftalmoloji bilgilerini değerlendirmek için özel olarak tasarlanmış TUS sorularını kullanarak BDM'lerin etkinliğini değerlendirmeye odaklanılmasıdır. Sonuç olarak, burada sunulan bulgular, BDM'lerin oftalmoloji eğitimindeki potansiyel rolünü tam olarak aydınlatacak kadar kapsamlı değildir. BDM'lerin bu alandaki kullanılabilirliğini daha iyi anlamak için oftalmolojinin çeşitli alt alanlarına odaklanan daha ayrıntılı çalışmalara ihtiyaç vardır. Ayrıca çalışmamızda sadece resmi olarak yayınlanan cevap anahtarları ve soru iptalleri dikkate alınmıştır. Nadiren de olsa, bu tür sınavlarda sorulara daha sonra itiraz edildiği ve soru iptali için yasal işlemlerin başlatıldığı durumlar olmuştur. Ancak, bu değişiklikler genellikle resmi olarak yayınlanan cevap anahtarlarına yansıtılmaz ve bu nedenle analizde dikkate alınamamıştır. Bu tür tartışmalı soruların çalışmaya dahil edilmemesi çalışmanın bir kısıtlılığıdır çünkü verilerin daha kapsamlı değerlendirilmesine engel olmuş olabilir.

Son derece nadir olmasına rağmen, her iki YZ tabanlı modelde de mantıksal doğru açıklamanın yapıldığı ancak yanlış seçeneğin seçildiği veya yanlış açıklama yapıldığı ancak doğru seçeneğin seçildiği gözlemlenmiştir. YZ'nın da herkes gibi hata yapabileceği her zaman hatırlanmalıdır, bu nedenle sonuçları kontrol etmek her zaman akıllıca olacaktır. Ayrıca, sohbet robotları uydurma referanslar ve halüsinasyonlar üretme eğiliminde olduğundan, özellikle referansların doğruluğunu değerlendirmeyi amaçlayan özel bir geçerlilik analizinin olmaması bir kısıtlılık olarak kabul edilebilir. Son olarak, bu sınavlarda katılımcıların doğruluk oranları bilinmediğinden ve bu bilgi kamuya açık olmadığından, insan doğruluk oranları ile YZ'lerin doğruluk oranları arasında bir karşılaştırma yapılamamıştır.

Bu kısıtlılıklar ile bile, bildiğimiz kadarıyla, çalışmamız, ChatGPT-4o'nun TUS sınavlarında oftalmoloji ile ilgili ÇSS'leri değerlendirmede Gemini 1.5 Pro'ya göre ılımlı derecede daha iyi bir performans sergilediğini ortaya koyan ilk karşılaştırmalı YZ çalışmasıdır. Ek olarak, çok sayıda ÇSS'nin (n=220) değerlendirilmiş olması ve geri bildirim yapılan ve yapılmayan üç ardışık girişim ile gereçekleştirilmiş olması çalışmamızın güçlü yönleridir. Ayrıca, bilimsel açıklamaların PubMed ve WoS Sitasyon Endeksine dayalı olmasının bir koşul olması sonuçları etkilemiş olabilir. En güncel YZ versiyonlarının kullanılmış olması da çalışmayı güçlendiren bir başka faktördür. Son olarak, diğer çalışmaların çoğundan farklı olarak, bu çalışmada şekil içeren sorular da değerlendirilmiştir.

Sonuç

Her iki YZ tabanlı BDM, oftalmoloji ile ilgili ÇSS'ları yanıtlamada iyi performans göstermiştir. Sadece oftalmoloji ile ilgili ÇSS'lara cevapları doğru bir şekilde belirlemekle kalmayıp aynı zamanda açıklamalar hazırlayarak oftalmoloji eğitimini geliştirme potansiyellerini göstermişlerdir. Her iki YZ platformu da kullanışlı olsa da, ChatGPT-4o anlamlı şekilde önde gitmektedir. Özellikle tıp öğrencileri ve oftalmoloji asistanlarına YZ odaklı e-öğrenmenin katkıları hakkında yapılacak daha ileri araştırmalar, bu nispeten yeni ortaya çıkan teknolojik alan için çok değerli olacaktır.

Etik

Etik Kurul Onayı: Bu çalışmaya insan katılımcı dahil edilmediği veya hayvan deneyi yapılmadığı için etik onay ve bilgilendirilmiş onam alınması gerekmemiştir.

Hasta Onayı: Bu makale için hasta onamı gerekmemektedir.

Teşekkür

Bu makalenin dil açısından incelemesindeki katkıları için anadili İngilizce olan ve İngiltere'de ikamet eden Dr. Hasan Nadir Rana'ya teşekkür ederiz.

Beyan

Yazarlık Katkıları

Cerrahi ve Medikal Uygulama: M.C.S., Z.Y., Konsept: M.C.S., Z.Y., Dizayn: M.C.S., Z.Y., Veri Toplama veya İşleme: M.C.S., Z.Y., Analiz veya Yorumlama: M.C.S., Z.Y., Literatür Arama: M.C.S., Z.Y., Yazan: M.C.S., Z.Y.

Çıkar Çatışması: Tüm yazarlar, bu makalede tartışılan konu veya materyallerle ilgili herhangi bir finansal veya finansal olmayan çıkarı olan herhangi bir kurum veya kuruluşla bağlantıları olmadığını bildirirler. Çoktan Seçmeli Soruların kullanımı yalnızca bilimsel amaçlıdır.

Finansal Destek: Chatbot platformlarının premium abonelikleri ile ilgili tüm masraflar yazarlar tarafından şahsen karşılanmıştır. Bu çalışmanın yürütülmesi için herhangi bir dış kaynaktan finansal destek alınmamıştır.

Kaynaklar

1. Cahit A. Can machines think and how can they think? Atatürk Üniversitesi 1958-1959 Öğretim Yılı Halk Konf. 1959:91-103.
2. Keskinbora K, Güven F. Artificial intelligence and ophthalmology. Turk J Ophthalmol. 2020;50:37-43.
3. Shemer A, Cohen M, Altarescu A, Atar-Vardi M, Hecht I, Dubinsky-Pertzov B, Shoshany N, Zmujack S, Or L, Einan-Lifshitz A, Pras E. Diagnostic capabilities of ChatGPT in ophthalmology. Graefes Arch Clin Exp Ophthalmol. 2024;262:2345-2352.
4. Tsui JC, Wong MB, Kim BJ, Maguire AM, Scoles D, VanderBeek BL, Brucker AJ. Appropriateness of ophthalmic symptoms triage by a popular online artificial intelligence chatbot. Eye (Lond). 2023;37:3692-3693.
5. Güler MS, Baydemir EE. Evaluation of ChatGPT-4 responses to glaucoma patients' questions: can artificial intelligence become a trusted advisor between doctor and patient? Clin Exp Ophthalmol. 2024;52:1016-1019.
6. Chen JS, Reddy AJ, Al-Sharif E, Shoji MK, Kalaw FGP, Eslani M, Lang PZ, Arya M, Koretz ZA, Bolo KA, Arnett JJ, Roginiel AC, Do JL, Robbins SL, Camp AS, Scott NL, Rudell JC, Weinreb RN, Baxter SL, Granet DB. Analysis of ChatGPT responses to ophthalmic cases: can ChatGPT think like an ophthalmologist? Ophthalmol Sci. 2024;5:100600.
7. Carlà MM, Gambini G, Baldascino A, Giannuzzi F, Boselli F, Crincoli E, D'Onofrio NC, Rizzo S. Exploring AI-chatbots' capability to suggest surgical planning in ophthalmology: ChatGPT versus Google Gemini analysis of retinal detachment cases. Br J Ophthalmol. 2024;108:1457.
8. Ming S, Yao X, Guo X, Guo Q, Xie K, Chen D, Lei B. Performance of ChatGPT in ophthalmic registration and clinical diagnosis: cross-sectional study. J Med Internet Res. 2024;26:e60226.
9. David D, Zloto O, Katz G, Huna-Baron R, Vishnevskia-Dai V, Armarnik S, Zauberman NA, Barnir EM, Singer R, Hostovsky A, Klang E. The use of artificial intelligence based chat bots in ophthalmology triage. Eye (Lond). 2025;39:785-789.
10. Halaweh M. ChatGPT in education: strategies for responsible implementation. Contemp Educ Technol. 2023;15:e421.
11. Tlili A, Shehata B, Adarkwah MA, Bozkurt A, Hickey DT, Huang R, Agyemang B. What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. Smart Learn Environ. 2023;10:1-24.
12. 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:e57348.
13. Panthier C, Gatinel D. Success of ChatGPT, an AI language model, in taking the French language version of the European Board of Ophthalmology examination: a novel approach to medical knowledge assessment. J Fr Ophtalmol. 2023;46:706-711.
14. Wu JH, Nishida T, Liu TYA. Accuracy of large language models in answering ophthalmology board-style questions: a meta-analysis. Asia Pac J Ophthalmol (Phila). 2024;13:100106.
15. 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;3:100324.
16. Sakai D, Maeda T, Ozaki A, Kanda GN, Kurimoto Y, Takahashi M. Performance of ChatGPT in board examinations for specialists in the Japanese Ophthalmology Society. Cureus. 2023;15:e49903.
17. Moshirfar M, Altaf AW, Stoakes IM, Tuttle JJ, Hoopes PC. Artificial intelligence in ophthalmology: a comparative analysis of GPT-3.5, GPT-4, and human expertise in answering StatPearls questions. Cureus. 2023;15:e40822.
18. Previous MSEs in 2006-2021. https://www.osym.gov.tr/TR,15072/tus-cikmis-sorular.html.
19. 10% of MSE questions in 2022. https://www.osym.gov.tr/TR,22532/2022.html.
20. 10% of MSE questions in 2023. https://www.osym.gov.tr/TR,25279/2023.html.
21. 10% of MSE questions in 2024. https://www.osym.gov.tr/TR,29136/2024.html.
22. Polit DF, Beck CT. The content validity index: are you sure you know what's being reported? Critique and recommendations. Res Nurs Health. 2006;29:489-497.
23. Oztermeli AD, Oztermeli A. ChatGPT performance in the medical specialty exam: an observational study. Medicine (Baltimore). 2023;102:e34673.
24. Ilgaz HB, Çelik Z. The significance of artificial intelligence platforms in anatomy education: an experience with ChatGPT and Google Bard. Cureus. 2023;15:e45301.
25. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, Madriaga M, Aggabao R, Diaz-Candido G, Maningo J, Tseng V. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2:e0000198.
26. Mihalache A, Huang RS, Popovic MM, Muni RH. Performance of an upgraded artificial intelligence chatbot for ophthalmic knowledge assessment. JAMA Ophthalmol. 2023;141:798-800.
27. Sabaner MC, Hashas ASK, Mutibayraktaroglu KM, Yozgat Z, Klefter ON, Subhi Y. The performance of artificial intelligence-based large language models on ophthalmology-related questions in Swedish proficiency test for medicine: ChatGPT-4 Omni vs Gemini 1.5 Pro. AJO Int. 2024:100070.
28. Yang Z, Wang D, Zhou F, Song D, Zhang Y, Jiang J, Kong K, Liu X, Qiao Y, Chang RT, Han Y, Li F, Tham CC, Zhang X. Understanding natural language: potential application of large language models to ophthalmology. Asia Pac J Ophthalmol (Phila). 2024;13:100085.
29. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. 2023;29:1930-1940.
30. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright CL, Mishkin P, Agarwal S, Slama K, Ray A, Schulman J, Hilton J, Kelton F, Miller L, Simens M, Askell A, Welinder P, Christiano P. Training language models to follow instructions with human feedback. Adv Neural Inf Process Syst. 2022;35:27730-27744.
31. Aydın FO, Aksoy BK, Ceylan A, Akbaş YB, Ermiş S, Kepez Yıldız B, Yıldırım Y. Readability and appropriateness of responses generated by ChatGPT 3.5, ChatGPT 4.0, Gemini, and Microsoft Copilot for FAQs in refractive surgery. Turk J Ophthalmol. 2024;54:313-317.
32. Sabaner MC, Anguita R, Antaki F, Balas M, Boberg-Ans LC, Ferro Desideri L, Grauslund J, Hansen MS, Klefter ON, Potapenko I, Rasmussen MLR, Subhi Y. Opportunities and challenges of chatbots in ophthalmology: a narrative review. J Pers Med. 2024;14:1165.
33. Postacı SA, Dal A. The ability of large language models to generate patient information materials for retinopathy of prematurity: evaluation of readability, accuracy, and comprehensiveness. Turk J Ophthalmol. 2024;54:330-336.
34. Vaishya R, Misra A, Vaish A. ChatGPT: is this version good for healthcare and research? Diabetes Metab Syndr. 2023;17:102744.
35. Gurnani B, Kaur K. Leveraging ChatGPT for ophthalmic education: a critical appraisal. Eur J Ophthalmol. 2024;34:323-327.
36. Pushpanathan K, Zou M, Srinivasan S, Wong WM, Mangunkusumo EA, Thomas GN, Lai Y, Sun CH, Lam JSH, Tan MCJ, Lin HAH, Ma W, Koh VTC, Chen DZ, Tham YC. Can OpenAI's new o1 model outperform its predecessors in common eye care queries? Ophthalmol Sci. 2025;5:100745.