KLİNİK OPTİK ÇOKTAN SEÇMELİ SORULARDA YAPAY ZEKÂ SOHBET BOTLARI: CHATGPT-5, COPİLOT VE GEMİNİ'NİN TÜRKÇE-İNGİLİZCE KARŞILAŞTIRMASI

AI CHATBOTS ON CLINICAL OPTICS MULTIPLE-CHOICE QUESTIONS: A TURKISH–ENGLISH COMPARISON OF CHATGPT-5, COPILOT, AND GEMINI

Yusuf Samet Atlıhan[1*], Tevfik Serhat Bahar[2], Olgar Öcal[3]

[1*]Göz Hastalıkları Kliniği, Silopi Devlet Hastanesi, Şırnak, Türkiye
[2]Göz Hastalıkları Kliniği, Alanya Alaaddin Keykubat Üniversitesi Alanya Eğitim ve Araştırma Hastanesi, Alanya, Türkiye
[3]Göz Hastalıkları Kliniği, Akdeniz Üniversitesi Hastanesi, Antalya, Türkiye

ORCID ID: YSA: 0000-0001-8761-6104 **TSB:** 0000-0003-1145-1782 **OÖ:** 0000-0002-1842-0906

DOI: 10.52309/jaihs.1855936

MAKALE BİLGİLERİ

Anahtar Kelimeler:
Büyük dil modelleri;
ChatGPT-5;
Copilot;
Gemini;
Klinik optik;
Oftalmoloji eğitimi;
Yapay zekâ sohbet botları

ÖZET

Amaç: Bu çalışmanın amacı, klinik optik alanındaki çoktan seçmeli sorularda büyük dil modellerinin (BDM) doğruluk düzeylerini değerlendirmek ve soruların dili (Türkçe–İngilizce) ile model performansı arasındaki olası ilişkiyi incelemektir.
Gereç ve Yöntem: American Academy of Ophthalmology (AAO) 2024–2025 Basic and Clinical Science Course (BCSC) Clinical Optics and Vision Rehabilitation kitabındaki dört seçenekli 88 soru seçildi; İngilizce özgün sorular sertifikalı çevirmen tarafından Türkçeye çevrildi. ChatGPT-5, Copilot ve Gemini'ye standart yönerge ile her soru ayrı oturumda yöneltildi. Yanıtlar resmi cevap anahtarıyla karşılaştırıldı. Birincil sonuç, modellerin her dildeki doğruluğu; ikincil sonuç, aynı soruların iki dildeki yanıt tutarlılığıydı. Karşılaştırmalarda Pearson ki-kare, eşleştirilmiş dil içi değerlendirmede McNemar testi kullanıldı.
Bulgular: İngilizce sorularda doğruluk oranları sırasıyla ChatGPT-5 %71,6 (63/88), Copilot %70,5 (62/88) ve Gemini %65,9 (58/88) olup modeller arasında fark yoktu (p = 0,354). Türkçe sürümlerde doğruluk oranları ChatGPT-5 %68,2 (60/88), Copilot %69,3 (61/88) ve Gemini %63,6 (56/88) olarak bulundu; aradaki fark istatistiksel olarak anlamlı değildi (p = 0,126). Dil içi tutarlılık yüksek seyretti: ChatGPT-5 aynı soruların %93,2'sinde (82/88), Copilot %95,5'inde (84/88) ve Gemini %92,1'inde (81/88) aynı seçeneği verdi; hiçbir modelde Türkçe–İngilizce doğruluk farkı anlamlı değildi (ChatGPT-5 p = 0,435; Copilot p = 0,674; Gemini p = 0,481).
Sonuç: Üç BDM, klinik optik sorularında orta düzey doğruluk (%63–%72) ve iki dil arasında büyük ölçüde benzer performans sergilemiştir. Bu durum, klinik optik eğitiminde BDM'lerin dil açısından tutarlı bir yardımcı araç olabileceğini düşündürmekle birlikte, mutlak doğruluğun sınırlı oluşu modellerin tek başına otorite olarak kullanılmaması gerektiğine işaret etmektedir.

ARTICLE INFO

Keywords:
ChatGPT-5;
Clinical optics;
Conversational AI chatbots; Copilot; Gemini;
Large language models;
Ophthalmology education

ABSTRACT

Objective: To evaluate the accuracy of large language models (LLMs) on multiple-choice questions (MCQs) in clinical optics and to assess the effect of language (Turkish vs English).
Materials and Methods: We selected 88 four-option items from the American Academy of Ophthalmology (AAO) 2024–2025 Basic and Clinical Science Course (BCSC) Clinical Optics and Vision Rehabilitation book; the original English items were translated into Turkish by a certified translator. The same questions, in both languages, were posed to ChatGPT-5, Copilot, and Gemini using a standardized prompt, with a fresh session for each question. Model outputs were compared with the official answer key. The primary outcome was model accuracy in each language; the secondary outcome was within-question agreement across the two languages. Group comparisons used the Pearson chi-square test, and within-question language comparisons used the McNemar test.
Results: On the English set, accuracies were 71.6% (63/88) for ChatGPT-5, 70.5% (62/88) for Copilot, and 65.9% (58/88) for Gemini, with no significant difference among models (p = 0.354). On the Turkish set, accuracies were 68.2% (60/88), 69.3% (61/88), and 63.6% (56/88), respectively; differences were not significant (p = 0.126). Cross-language consistency was high: ChatGPT-5 gave the same option on 93.2% (82/88) of paired questions, Copilot on 95.5% (84/88), and Gemini on 92.1% (81/88). For none of the models was the English–Turkish accuracy difference significant (ChatGPT-5 p = 0.435; Copilot p = 0.674; Gemini p = 0.481).
Conclusion: Across clinical optics MCQs, the three LLMs achieved moderate accuracy (63%–72%) and largely comparable performance in Turkish and English. While this cross-lingual stability supports their use as language-consistent adjuncts in clinical optics education, the limited absolute accuracy argues against relying on them as stand-alone authorities.

Makale Bilgisi | Article Information

Makale Türü | Article Type: Araştırma Makalesi | Research Article
Geliş Tarihi | Received: 05.01.2026
Kabul Tarihi | Accepted: 16.03.2026
Yayım Tarihi | Published: 27.04.2026

Çıkar çatışması
Yazarlar herhangi bir çıkar çatışması veya finansal destek bildirmemişlerdir.
Finansman
Yazarlar bu çalışma için hiçbir finansal destek almadıklarını beyan eder.
Teşekkürler
Çalışmanın kongrede sunulduğuna dair bilgi
Çalışma herhangi bir kongrede/sempozyumda sunulmamıştır.
Etik Onay
Çalışmada insan katılımcılara/insana ait verilere ya da hayvan deneylerine ilişkin herhangi bir veri kullanılmamıştır. Bu nedenle etik kurul onayı ve bilgilendirilmiş onam gerekmemektedir.

Sorumlu Yazar | Correspondence Author:
Yusuf Samet Atlıhan

Address for Correspondence: Göz Hastalıkları Kliniği, Silopi Devlet Hastanesi, Şırnak, Türkiye
Mail: Ysf20smt@gmail.com

Yazar Katkıları
Motivasyon / Konsept: YSA
Çalışma Tasarımı: YSA, TSB, OÖ
Kontrol / Gözetim: OÖ
Veri Toplanması ve / veya İşlemesi: YSA, TSB, OÖ
Analiz ve / veya Yorum: TSB
Literatür incelemesi: YSA
Makalenin Yazılması: YSA, TSB, OÖ
Eleştirel İnceleme: TSB, OÖ

GİRİŞ

Büyük dil modeli (BDM) sohbet botları, son yıllarda gündelik ve akademik hayata girmiş; geniş veri kümeleriyle eğitilerek kavramlar arasındaki örüntüleri yakalayan ve olası devamları kestirebilen, insan düşünme süreçlerinin belli yönlerini taklit etmeyi hedefleyen bir teknoloji olarak öne çıkmıştır.1 Yapay zeka (YZ) tabanlı büyük dil modellerinin ortaya çıkmasıyla klinik karar desteği ve tıp eğitimi alanlarında potansiyel kullanım alanları giderek daha fazla araştırılmaktadır. 2-4 BDM kullanarak oftalmoloji sınavları stilinde çoktan seçmeli sorulara cevap bulmak mümkün olabilmektedir.5,6

Oftalmoloji, temel hedefi daha iyi görme olan bir uzmanlık alanı olduğundan bu hedefe ulaşmak için, gözün optik fonksiyonunun doğru anlaşılması gerekir.7 Klinik optik, oftalmoloji uzmanlık eğitiminin temel bir bileşenidir ve hem öğretimi hem de öğrenimi açısından, öğrenciler ve öğretim üyeleri için zorlayıcı olabilir.8

Bu çalışmanın amacı, klinik optik alanında bilgi edinmek ve sınavlara hazırlanmak isteyen hekimlerin BDM ile doğru bilgiye erişimini değerlendirmek ve güncel, yaygın kullanılan ve üretici tarafından ücretsiz olarak piyasaya sürülmüş olan YZ sohbet botlarının (ChatGPT-5, Copilot, Gemini) klinik optikle ilgili aynı sorulara İngilizce ve Türkçe yanıt verirken gösterdikleri performans üzerinde dil farklılığının etkisini araştırmaktır.

GEREÇ VE YÖNTEM

Araştırmanın Tasarımı

Amerikan Oftalmoloji Akademisi'nin (AAO) 2024–2025 Basic and Clinical Science Course (BCSC) Clinical Optics and Vision Rehabilitation (Klinik Optik ve Görme Rehabilitasyonu) kitabının çalışma soruları bölümünde yer alan, dört seçenekli toplam 88 çoktan seçmeli soru çalışmaya dahil edildi.9 İngilizce maddeler sertifikalı bir çevirmen tarafından Türkçeye çevrildi ve her iki dildeki sürümler Microsoft Word belgesinde düzenlendi. Aynı soruların İngilizce ve Türkçe sürümleri, ücretsiz erişilebilen üç yapay zekâ sohbet botuna uygulandı: ChatGPT-5 (OpenAI, San Francisco, CA, ABD), Copilot (Microsoft, Redmond, WA, ABD) ve Gemini (Google, Mountain View, CA, ABD) Gemini modeli Google'ın web tabanlı kullanıcı arayüzü üzerinden varsayılan model yapılandırması kullanılarak çalıştırıldı. Modeller, çevrimiçi kullanıcı arayüzleri üzerinden varsayılan ayarlarla kullanıldı. ChatGPT ve Gemini için ek bir özel çalışma modu (örneğin hızlı mod veya "düşünen" mod) seçilmedi ve sistemlerin standart/otomatik çalışma modu tercih edildi. Her soru için ayrı bir oturum açıldı ve yanıt alındıktan sonra oturum sonlandırılarak bir sonraki soru yeni bir oturumda soruldu. Modellerden elde edilen yanıtlar kitapta yer alan resmî cevap anahtarıyla karşılaştırıldı. Çalışmada kullanılan yapay zekâ sistemlerine erişim tarihi Kasım 2025 olarak kaydedildi. Uygulamaya başlamadan önce tüm modellere standart bir yönerge verildi: "Sana çoktan seçmeli sorular soracağım. Lütfen bana en doğru cevap seçeneğini ver." Her madde bağımsız bir oturumda yürütüldü; her yanıtın ardından oturum sonlandırılarak bir sonraki madde için yeni bir oturum başlatıldı. Modellerin verdiği yanıtlar, kitapta yer alan resmî cevap anahtarıyla karşılaştırılarak doğru/yanlış olarak sınıflandırıldı; ayrıca her maddenin İngilizce ve Türkçe sürümlerine verilen yanıtlar madde düzeyinde eşleştirilip kaydedildi ve tüm kayıtlar Word dosyasında tutuldu. Birincil sonuç, her modelin İngilizce ve Türkçe maddelerdeki doğru yanıtlama oranları; ikincil sonuçlar ise her bir maddenin iki dil sürümünde elde edilen doğruluk durumlarının karşılaştırılması olarak tanımlandı. Çalışmaya dahil edilen tüm çoktan seçmeli sorular yalnızca metin içerikli olup herhangi bir görsel, şekil veya diyagram içermemektedir.

Etik Onay

Bu çalışma, etik kurul onayı gerektiren araştırma kapsamına girmemektedir. Çalışmada insan katılımcılara/insana ait verilere ya da hayvan deneylerine ilişkin herhangi bir veri kullanılmamıştır. Bu nedenle etik kurul onayı ve bilgilendirilmiş onam gerekmemektedir.

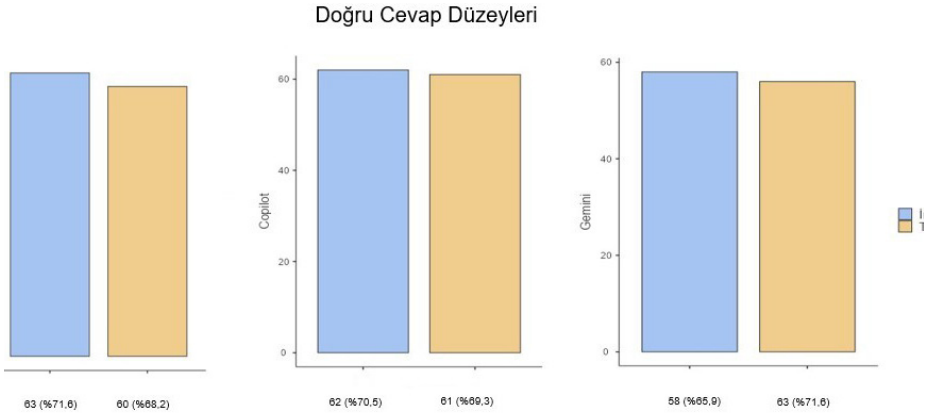
İstatistiksel Analiz

İstatistiksel analizler SPSS sürüm 23 (IBM Corp., Armonk, NY, USA) yazılımıyla gerçekleştirildi. Tanımlayıcı veriler yüzde (%) olarak sunuldu. Nominal değişkenlerde bağımsız grupların karşılaştırılması için Pearson ki-kare testi, bağımlı/eşleştirilmiş nominal verilerin karşılaştırılması için McNemar testi uygulandı. İki yönlü anlamlılık düzeyi $p < 0,05$ olarak kabul edildi.

BULGULAR

Klinik optik alanına ilişkin 88 tane çoktan seçmeli soru, YZ sohbet botlarına İngilizce olarak yöneltildi. ChatGPT-5, 63 soruyu doğru (%71,6) ve 25'ini yanlış (%28,4) yanıtladı; Copilot, 62 doğru (%70,5) ve 26 yanlış (%29,5) verdi; Gemini ise 58 doğru (%65,9) ve 30 yanlış (%34,1) yanıt verdi (Şekil 1). İngilizce sorularda üç modelin doğruluk oranları karşılaştırıldığında istatistiksel olarak anlamlı bir fark saptanmadı ($p = 0,354$).

Klinik optik alanına ilişkin 88 çoktan seçmeli sorunun Türkçe çevirileri YZ sohbet botlarına aynı protokolle soruldu. ChatGPT-5, 60 soruyu doğru (%68,2) ve 28'ini yanlış (%31,8) yanıtladı; Copilot, 61 soruyu doğru (%69,3) ve 27 soruyu yanlış (%30,7) yanıtladı; Gemini ise 56 soruyu doğru (%63,6) ve 32 soruyu yanlış (%36,4) yanıtladı (Şekil 1). Soruların Türkçe çevirilerinde üç modelin doğruluk oranları karşılaştırıldığında istatistiksel olarak anlamlı fark saptanmadı ($p = 0,126$).



Şekil 1. ChatGPT-5, Copilot ve Gemini'nin klinik optik konusunda İngilizce ve Türkçe soruları doğru cevaplama düzeyleri

Şekil 1 Açıklaması: Bu şekil, ChatGPT-5, Copilot ve Gemini'nin klinik optik alanında yöneltilen İngilizce ve Türkçe sorulara verdikleri doğru cevap sayılarını ve yüzde oranlarını göstermektedir. Her bir yapay zekâ modeli için İngilizce ve Türkçe performanslar ayrı sütunlar halinde sunulmuş olup, modellerin her iki dildeki başarı düzeyleri karşılaştırmalı olarak değerlendirilmiştir. Sütunlar üzerindeki değerler, doğru cevap sayısını ve buna karşılık gelen yüzde oranını ifade etmektedir.

ChatGPT-5, İngilizce ve Türkçe'ye çevirisi yapılmış 88 sorunun 82'sinde (%93,2) aynı yanıtı verdi; 6 soruda (%6,8) ise farklı yanıtlar verdi. Farklı yanıt verdiği 6 sorunun 3'ünde (%50) sorular Türkçe sorulduğunda doğru cevaplanırken, diğer 3'ünde (%50) ise sorular Türkçe sorulduğunda yanlış olarak cevaplandı. ChatGPT-5 için İngilizce ve Türkçe sorulardaki doğruluk oranları karşılaştırıldığında istatistiksel olarak anlamlı bir fark görülmedi ($p = 0,435$) (Tablo 1).

Tablo 1. Büyük dil modellerinin aynı sorulara verdikleri cevaplar ve değişimleri

Cevaplar	ChatGPT-5 (İngilizce)	ChatGPT-5 (Türkçe)	Copilot (İngilizce)	Copilot (Türkçe)	Gemini (İngilizce)	Gemini (Türkçe)
Doğru cevap	63 (%71,6)	60 (%68,2)	62 (%70,5)	61 (%69,3)	58 (%65,9)	56 (%63,6)
Yanlış cevap	25 (%28,4)	28 (%31,8)	26 (%29,5)	27 (%30,7)	30 (%34,1)	32 (%36,4)
p	*0.435**		*0.674**		*0.481**	
Aynı cevabı üretme	82 (%93,2)		84 (%95,5)		81 (%92,1)	
Farklı cevabı üretme	6 (%6,8)		4 (%4,5)		7 (%7,9)	
Doğru-yanlış değişimi	3 (%50)		3 (%75)		4 (%57,1)	
Yanlış-doğru değişimi	3 (%50)		1 (%25)		3 (%42,9)	

*McNemar testi

Copilot, İngilizce ve Türkçe'ye çevirisi yapılmış 88 sorunun 84'ünde (%95,5) aynı yanıtı verdi; 4 soruda (%4,5) ise farklı yanıtlar verdi. Farklı yanıt verdiği 4 sorunun 3'ünde (%75) sorular Türkçe sorulduğunda doğru cevaplanırken, 1 (%25) soruda ise Türkçe sorulduğunda yanlış olarak cevaplandı. Copilot için İngilizce ve Türkçe sorulardaki doğruluk oranları karşılaştırıldığında istatistiksel olarak anlamlı bir fark görülmedi (p = 0,674) (Tablo 1).

Gemini, İngilizce ve Türkçe'ye çevirisi yapılmış 88 sorunun 81'inde (%92,1) aynı yanıtı verdi; 7 soruda (%7,9) ise farklı yanıtlar verdi. Farklı yanıt verdiği 7 sorunun 4'ünde (%57,1) sorular Türkçe sorulduğunda doğru cevaplanırken, diğer 3'ünde (%42,9) ise sorular Türkçe sorulduğunda yanlış olarak cevaplandı. Gemini için İngilizce ve Türkçe sorulardaki doğruluk oranları karşılaştırıldığında istatistiksel olarak anlamlı bir fark görülmedi (p = 0,481) (Tablo 1).

TARTIŞMA

Oftalmoloji alanında BDM'ler hızla kendilerine yer edinirken, verdikleri bilgilerin doğruluklarının oftalmolojinin alt alanları arasında karşılaştırıldığı çalışmaların sayısı da artmaktadır.10-14 ChatGPT ve Gemini gibi BDM tabanlı sohbet botları tıp eğitiminde öğretim aracı olarak yoğun ilgi görmektedir,2, 15-17 BDM'lerin ilk sürümleri, Türkiye Tıpta Uzmanlık Sınavı (TUS), Amerika Birleşik Devletleri Tıp Lisanslama Sınavı (USMLE) ve oftalmolojiye özgü soru bankaları gibi farklı sınav bağlamlarında doğruluk oranlarının çoğunlukla %50–%70 aralığında olduğu rapor edilmiştir;18, 19 bu tablo temkinli bir iyimserliği desteklemekle birlikte, muhakeme derinliği, alan-özgül duyarlılık ve çok dilli tutarlılık açısından belirgin sınırlılıklara da işaret etmektedir.

Klinik optik, oftalmoloji eğitiminde bilgi açıklarının sık bildirildiği ve birçok göz hekimi için zorlayıcı bir konudur.8 BDM klinik optik sorularındaki performansı ise şimdiye dek sınırlı olarak incelenmiştir; bu durum çalışmamızın özgün değerini artırmaktadır. Ayrıca konunun karmaşıklığı yalnızca öğrenenler için değil, YZ modelleri için de güçlük yaratmaktadır: nitekim, oftalmolojinin farklı alt alanlarından çoktan seçmeli sorularla yapılan bir değerlendirmede ChatGPT-5, oftalmolojik patoloji sorularında en yüksek doğruluğa (%66,7) ulaşırken, klinik optik sorularında en düşük doğruluk oranı gözlenmiştir (%32,6).20 Bu tablo, klinik optikte BDM performansını iyileştirmek için hedefe özgü veri setleri ve değerlendirme yöntemlerine ihtiyaç olduğunu düşündürmektedir.

Bizim çalışmamızda da gözlendiği gibi farklı dillerde benzer doğruluk oranlarının elde edilmesi, büyük dil modellerinin diller arasında benzer bir performans sergileyebileceğini düşündürmektedir.

Bununla birlikte, bu bulgunun farklı veri setleri ve daha geniş örneklemlerle yapılacak çalışmalarla desteklenmesi gerekmektedir.21, 22

Bu çalışma, klinik optik alanında çoktan seçmeli sorulara verilen yanıtlar üzerinden üç yaygın BDM sohbet botunun (ChatGPT-5, Copilot, Gemini) dil bağımlı performansını sınadı ve hem İngilizce hem de Türkçe'de benzer doğruluk düzeyleri sergilediklerini; diller arasında istatistiksel olarak anlamlı fark oluşmadığını gösterdi. İngilizce ve Türkçe sorular arasında her üç modelin de yüksek oranda aynı yanıtı üretmesi (%92–96) dil farklılığının pratik etkisinin sınırlı kaldığını düşündürüyor. Bununla birlikte, doğruluk oranlarının %63–72 arasında olması, modellerin yardımcı eğitim aracı olarak güçlü yanlarını gösterirken, tek başına otorite olarak kullanılmaması gerektiğini de hatırlatıyor.

Bu çalışmanın güçlü yönü, aynı 88 adet çoktan seçmeli soruların iki dilde ve üç farklı modelle, standartlaştırılmış yönergeler altında sınanmış olmasıdır. Bu çalışmanın bazı sınırlılıkları bulunmaktadır. Öncelikle, değerlendirme yalnızca 88 adet çoktan seçmeli soru ile sınırlı bir veri seti üzerinden gerçekleştirilmiştir. Daha geniş soru bankaları ve farklı soru formatlarıyla yapılacak çalışmalar, büyük dil modellerinin klinik optik alanındaki performansının daha kapsamlı değerlendirilmesine ve elde edilen bulguların genellenebilirliğinin artırılmasına katkı sağlayabilir. Çalışmada yalnızca klinik optiğe odaklanılması ve BDM'lerin oftalmolojinin diğer alt alanları arasında karşılaştırılmamış olması, soru biçiminin çoktan seçmeliyle sınırlı kalması, modellerin sürüm/güncelleme değişkenliğine ve ücretsiz sürümlerin kısıtlarına duyarlılık ile çeviri sürecinden kaynaklanabilecek kaçınılmaz mikro-nüanslar yer alır. Bu etmenler, bulguların genellenebilirliğini kısmen sınırlar. Gelecek çalışmaların, konu yelpazesini genişleterek kornea, retina ve nöro-oftalmoloji gibi oftalmolojinin farklı alt alanlarını kapsaması; ayrıca açık uçlu ve vaka temelli soruları içeren değerlendirme tasarımlarını kullanması yararlı olabilir. Bunun yanında, modellerin yalnızca doğru yanıt oranlarının değil, aynı zamanda gerekçeli yanıtların kalitesi ve kaynak gösterme davranışlarının da sistematik olarak değerlendirilmesi önerilmektedir. Ayrıca, dil kapsamının Türkçe'nin ötesine genişletilmesi, model performansının zaman içindeki değişiminin (sürüm güncellemeleri) izlenmesi ve hata türlerinin (bilgi eksikliği, hesaplama hatası, terim kullanımı farklılıkları vb.) ayrıntılı biçimde analiz edilmesi, bu alandaki bulguların daha kapsamlı biçimde anlaşılmasına katkı sağlayacaktır.

Bulgular, büyük dil modellerinin klinik optik öğreniminde yardımcı bir araç olarak kullanılabileceğini ve farklı diller arasında büyük ölçüde benzer bir performans sergilediğini düşündürmektedir. Bununla birlikte, mutlak doğruluk oranlarının sınırlı olması, bu sistemlerin tek başına güvenilir bir bilgi kaynağı olarak kullanılmaması gerektiğine işaret etmektedir. Bu bağlamda, geleneksel eğitim kaynaklarının temel rolünü koruduğu ve büyük dil modellerinin eleştirel bir yaklaşımla tamamlayıcı bir araç olarak kullanıldığı bir eğitim yaklaşımının daha uygun olabileceği değerlendirilmektedir.

Çıkar Çatışması ve Finansman Bildirimi

Yazar bildirecek bir çıkar çatışması olmadığını beyan eder. Yazar bu çalışma için hiçbir finansal destek almadığını da beyan eder.

REFERANSLAR

1. 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):1233-1239. doi:10.1056/NEJMsr2214184.
2. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2(2):e0000198. doi:10.1371/journal.pdig.0000198.
3. Mihalache A, Popovic MM, Muni RH. Performance of an artificial intelligence chatbot in ophthalmic knowledge assessment. JAMA Ophthalmol. 2023;141(6):589-597. doi:10.1001/jamaophthalmol.2023.1144.
4. Vrdoljak J, Boban Z, Vilović M, et al. A review of large language models in medical education, clinical decision support, and healthcare administration. Healthcare (Basel). 2025;13(6):603. doi:10.3390/healthcare13060603.

5. 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(5):100106. doi:10.1016/j.apjo.2024.100106.
6. Antaki F, Touma S, Milad D, et al. Evaluating the performance of ChatGPT in ophthalmology: An analysis of its successes and shortcomings. Ophthalmol Sci. 2023;3(4):100324. doi:10.1016/j.xops.2023.100324.
7. Alipour F, Samavati K, Abdi P, et al. An innovative optics lab design for residency training in ophthalmology. Sci Rep. 2023;13(1):21844. doi:10.1038/s41598-023-44688-1.
8. Tu DC, Park DW, Krisciunas T, et al. Teaching ophthalmology residents clinical optics via a flipped classroom curriculum. J Acad Ophthalmol. 2019;11(1):e16-e21.
9. American Academy of Ophthalmology. Clinical Optics and Vision Rehabilitation. Basic and Clinical Science Course, Section 3, 2024–2025. American Academy of Ophthalmology; 2024.
10. Şensoy E, Şensoy M, Çıtırık M. Dil farklılıklarının ChatGPT-3.5, Copilot ve Gemini'nin pediatrik oftalmoloji ve şaşılık çoktan seçmeli sorulardaki etkinliğinin değerlendirilmesi. Anadolu Klin. 2025;30(2):262-267.
11. Dhiman S, Mishra C, Kaur S, et al. Comparative evaluation of five AI chatbots in pediatric ophthalmology: A multidomain expert-based appraisal. AJO Int. 2025;100188.
12. Cohen SA, Fisher AC, Xu BY, et al. Comparing the accuracy and readability of glaucoma-related question responses and educational materials by Google and ChatGPT. J Curr Glaucoma Pract. 2024;18(3):110-116. doi:10.5005/jp-journals-10078-1448.
13. Balas M, Mandelcorn ED, Yan P, et al. ChatGPT and retinal disease: A cross-sectional study on AI comprehension of clinical guidelines. Can J Ophthalmol. 2025;60(1):e117-e123. doi:10.1016/j.jcjo.2024.06.001.
14. Das D, Narayan A, Mishra V, et al. AI chatbots in answering questions related to ocular oncology: A comparative study between DeepSeek v3, ChatGPT-4o, and Gemini 2.0. Cureus. 2025;17(8):e90773. doi:10.7759/cureus.90773.
15. 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(9):1016-1019. doi:10.1111/ceo.14451.
16. David D, Zloto O, Katz G, et al. The use of artificial intelligence-based chatbots in ophthalmology triage. Eye (Lond). 2025;39(4):785-789. doi:10.1038/s41433-024-03488-1.
17. Carlà MM, Gambini G, Baldascino A, et al. Exploring AI-chatbots' capability to suggest surgical planning in ophthalmology: ChatGPT versus Google Gemini analysis of retinal detachment cases. Br J Ophthalmol. 2024;108(10):1457-1469. doi:10.1136/bjo-2023-325143.
18. 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(4):177-185. doi:10.4274/tjo.galenos.2025.27895.
19. Oztermeli AD, Oztermeli A. ChatGPT performance in the medical specialty exam: An observational study. Medicine (Baltimore). 2023;102(32):e34673. doi:10.1097/MD.0000000000034673.
20. Yavrum F, Kocabas Ozkoyuncu D. Is ChatGPT a useful tool for ophthalmology practice? Acta Med Alanya. 2024;8(3):221-227. doi:10.30565/medalanya.1531790.
21. 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(7):706-711. doi:10.1016/j.jfo.2023.05.006.
22. Alonso I, Oronoz M, Agerri R. MedExpQA: Multilingual benchmarking of large language models for medical question answering. Artif Intell Med. 2024;155:102938. doi:10.1016/j.artmed.2024.102938.