Araştırma Makalesi / Research Article

Oftalmik Patolojiler ve Göz İçi Tümörlerinde Dil Farklılıklarının Yapay Zeka Chatbot Performansı Üzerindeki Etkisinin Değerlendirilmesi: ChatGPT-3.5, Copilot ve Gemini Üzerine Bir Çalışma

Assessing the Impact of Language Differences on Artificial Intelligence Chatbot Performance in Ophthalmic Pathologies and Intraocular Tumors: A Study of ChatGPT-3.5, Copilot, and Gemini

Eyüpcan ŞENSOY [1], Mehmet ÇITIRIK [1]

[1]Sağlık Bilimleri Üniversitesi, Ankara Etlik Şehir Hastanesi, Göz Hastalıkları Anabilim Dalı, Ankara, TÜRKİYE

Öz

Amaç: ChatGPT-3.5, Copilot ve Gemini yapay zeka sohbet botlarının oftalmik patolojiler ve intraoküler tümörlerle ilişkili çoktan seçmeli sorularda ki başarısına dil farklılığının etkisini araştırmak
Materyal ve metod: Oftalmik patolojiler ve intraoküler tümörlerle ilgili bilgi düzeyini test eden 36 İngilizce soru çalışmaya dahil edildi. Sertifikasyonlu çevirmen (native speaker) tarafından Türkçe çevirilerinin gerçekleştirilmesi sonrasında bu soruların hem İngilizce hem de Türkçe olarak ChatGPT-3.5, Copilot ve Gemini sohbet botlarına soruldu. Verilen cevaplar cevap anahtarı ile karşılaştırılıp doğru ve yanlış olarak gruplandırıldı.
Bulgular: ChatGPT-3.5, Copilot ve Gemini İngilizce sorulara sırası ile %75, %66,7 ve %63,9 oranında doğru cevap verdi. Bu programlar Türkçe sorulara ise sırası ile %63,9, %66,7 ve %69,4 oranında doğru cevap verdi. Sohbet botları arasında soruların Türkçe hallerini cevaplama da farklı oranda doğru cevap görüldüğü halde, istatistiksel olarak anlamlı bir fark tespit edilmedi (p>0,05).
Sonuç: Yapay zeka sohbet botlarının bilgi dağarcığının geliştirilmesinin yanında farklı dillerde aynı algıyı oluşturabilmek ve tek doğruya erişimi sağlayabilmek için farklı dilleri anlama, çevirebilme ve fikir üretebilme özelliklerinin de geliştirilmeye ihtiyacı vardır.

Anahtar Kelimeler: ChatGPT-3.5, Copilot, Gemini, Göz içi tümörler ve oftalmik patolojiler, İngilizce ve Türkçe

Abstract

Background: To investigate the effect of language differences on the success of ChatGPT-3.5, Copilot, and Gemini artificial intelligence chatbots in multiple-choice questions related to ophthalmic pathologies and intraocular tumors.
Materials and Methods: Thirty-six English questions testing knowledge about ophthalmic pathologies and intraocular tumors were included in the study. These questions were asked to ChatGPT-3.5, Copilot, and Gemini chatbots in both English and Turkish after the Turkish translations were realized by a certified translator (native speaker). The answers given were compared with the answer key and grouped as correct and incorrect.
Results: ChatGPT-3.5, Copilot, and Gemini answered the questions in English correctly at a rate of 75%, 66.7%, and 63.9%, respectively. These programs answered the Turkish questions correctly at a rate of 63.9%, 66.7%, and 69.4%, respectively. Although there were different rates of correct answers between chatbots in answering the Turkish versions of the questions, no statistically significant difference was detected (p>0.05).
Conclusions: In addition to improving the knowledge of artificial intelligence chatbots, their ability to understand different languages, translate, and generate ideas needs to be improved to create the same perception in different languages and provide access to a single truth.

Keywords: ChatGPT-3.5, Copilot, Gemini, English and Turkish, Intraocular tumors and ophthalmic pathologies

Sorumlu Yazar / Corresponding Author

Dr. Eyüpcan ŞENSOY
Sağlık Bilimleri Üniversitesi, Ankara Etlik Şehir Hastanesi, Göz Hastalıkları Anabilim Dalı, Ankara, TÜRKİYE

E-mail: dreyupcansensoy@yahoo.com

Geliş tarihi / Received: 26.07.2024

Kabul tarihi / Accepted: 24.01.2025

DOI: 10.35440/hutfd.1522631

Giriş

Yapay zeka uygulamalarının yaygınlaşması ile birlikte etkileri tıbbın tüm alanlarında izlenmeye başlanmış ve bu alanlara çok çeşitli katkılar sağlamıştır (1). Oftalmoloji alanında yaygınlaşmaya başlaması ise yaklaşık 2015 yılından sonra gerçekleşmiş, özellikle derin öğrenme tabanlı yapay zeka programlarının diyabetik retinopati, glokom gibi retinal ve optik sinir patolojilerinin tanı ve takibinde kullanılabileceğinin anlaşılması ile birlikte popülerlik kazanmıştır (2-4). Yapay zeka uygulamalarının bir diğer kolu da Büyük Dil Modeli (BDM) tabanlı programlardır. Bu programlar derin öğrenme tabanlı programlar gibi insanların öğrenme şeklini taklit etmek yerine kavramlar arasında ilişki kurma, istatistiksel modellerle analiz etme ve bunları yorumlayabilme kabiliyetine sahiptir. BDM tabanlı programlar kullanıcılar tarafından tanımlanmış olan bilgileri anlayabilir, onlarla ilgili özetler çıkarabilir ve sorulan sorulara uygun cevaplar üretebilirler (5).
Çalışmamızın amacı BDM temelli yapay zeka programları olan ChatGPT-3.5 (OpenAI), Copilot (Microsoft) ve Gemini (Google) sohbet botlarının oftalmik patolojiler ve intraoküler tümörler ile ilişkili olan çoktan seçmeli farklı dillerdeki (İngilizce ve Türkçe) aynı soruları doğru cevaplama potansiyelini değerlendirmek ve farklı dillerdeki soruların bu başarıya olan etkisini incelemektir.

Materyal ve Metod

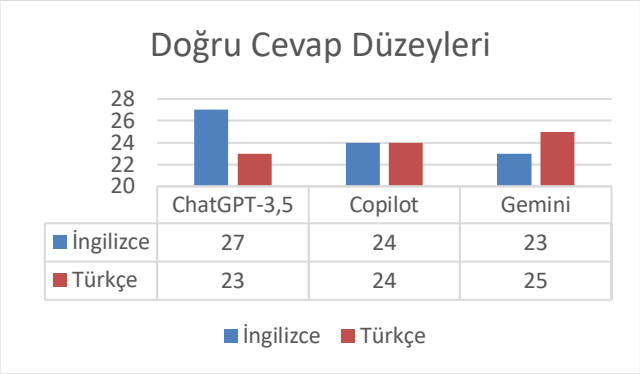
Amerikan Akademi ve Oftalmoloji 2023-2024 Basic and Clinical Science Course (BCSC) Oftalmik Patolojiler ve İntraoküler Tümörler kitabı çalışma soruları kısmında yer alan 36 sorunun tamamı çalışmaya dahil edildi (6). İngilizce olan bu sorular daha sonra sertifikasyonlu çevirmen (native speaker) tarafından Türkçeye çevrildi. Soruların İngilizce ve Türkçe versiyonları ücretsiz olarak erişim sağlanabilen ChatGPT-3.5 (OpenAI; San Francisco, CA), Copilot (Microsoft, Redmond,WA) ve Gemini (Google, Mountain View, California, United States) yapay zeka sohbet botlarına 7 Temmuz 2024 tarihinde uygulandı. Sorular sorulmadan önce yapay zeka programlarına 'Sana çoktan seçmeli sorular soracağım. Lütfen bana doğru cevap şıkkını ver.' komutu verildi. Her sorunun cevaplanmasından sonra oturum sonlandırıldı. Sohbet botlarının verdikleri cevaplar kitap arkasında yer alan cevap anahtarı ile karşılaştırılarak doğru ve yanlış olarak iki grup altında incelendi. Çalışmamızdaki veriler insan ve hayvan denekleri kaynaklı olmadığı için etik kurul onayı gerektirmemektedir.

İstatistiksel Analiz

Verilerin istatistiksel analizinde Statistical Package fort he Social Sciences sürüm 23 (SPSS Inc., Chicago, IL, USA) programı kullanıldı. Yüzdelik değerler hesaplandı. Bağımsız gruplarda nominal verilerin istatistiksel karşılaştırması için Pearson Ki-kare testi kullanıldı. Bağımlı gruplarda nominal verilerin istatistiksel karşılaştırması için McNemar testi kullanıldı. P değerinin 0,05'in altında olması anlamlılık düzeyi olarak değerlendirildi.

Bulgular

Oftalmik patolojiler ve intraoküler tümörler ile ilgili bilgi düzeyini sorgulayan 36 adet İngilizce çoktan seçmeli soru ChatGPT-3.5, Copilot ve Gemini yapay zeka sohbet programlarına uygulandı ChatGPT-3.5, sorulan soruların 27'sine (%75) doğru cevap, 9'una (%25) yanlış cevap verdi. Copilot sorulan soruların 24'üne (%66,7) doğru cevap, 12'sine (%35,3) yanlış cevap verdi. Gemini sorulan soruların 23'üne (%63,9) doğru cevap, 13'üne (%36,1) yanlış cevap verdi (Şekil 1). İngilizce soruları doğru cevaplamada her üç yapay zeka programı arasında istatistiksel anlamlı düzeyde bir fark tespit edilmedi (p=0,572 Pearson Ki-kare testi).



Şekil 1. ChatGPT-3,5, Copilot ve Gemini'nin oftalmik patolojiler ve intraoküler tümörler hakkındaki İngilizce ve Türkçe soruları doğru cevaplama düzeyleri

Aynı soruların Türkçe versiyonları ChatGPT-3.5, Copilot ve Gemini sohbet botlarına uygulandı. ChatGPT-3.5, sorulan soruların 23'üne (%63,9) doğru cevap, 13'üne (%36,1) yanlış cevap verdi. Copilot sorulan soruların 24'üne (%66,7) doğru cevap, 12'sine (%35,3) yanlış cevap verdi. Gemini sorulan soruların 25'ine (%69,4) doğru cevap, 11'ine (%30,6) yanlış cevap verdi (Şekil 1). Türkçe soruları doğru cevaplamada her üç yapay zeka programı arasında istatistiksel anlamlı düzeyde bir fark tespit edilmedi (p=0,882 Pearson Ki-kare testi).
ChatGPT-3.5 sorulan İngilizce ve Türkçe soruların 24'üne (%66,7) aynı cevabı verirken, 12'sine (%33,3) farklı cevaplar verdi. Farklı cevaplar verdiği soruların sekizi (%66,7) Türkçe sorulduğunda yanlış cevaplanırken, dördü (%33,3) Türkçe sorulduğunda doğru cevaplandı. ChatGPT-3.5'in aynı soruları İngilizce ve Türkçe olarak doğru olarak cevaplamasında istatistiksel olarak anlamlı düzeyde bir fark izlenmedi (p=0,388 McNemar testi) (Tablo 1).
Copilot sorulan İngilizce ve Türkçe soruların 26'sına (%72,2) aynı cevabı verirken, 10'una (%27,8) farklı cevaplar verdi. Farklı cevaplar verdiği soruların beşi (%50) Türkçe sorulduğunda yanlış cevaplanırken, beşi (%50) Türkçe sorulduğunda doğru cevaplandı. Copilot'un aynı soruları İngilizce ve Türkçe olarak doğru olarak cevaplamasında istatistiksel olarak anlamlı düzeyde bir fark izlenmedi (p=1,0 McNemar testi) (Tablo 1).

Gemini sorulan İngilizce ve Türkçe soruların 28'ine (%77,8) aynı cevabı verirken, sekizine (%22,2) farklı cevaplar verdi. Farklı cevaplar verdiği soruların dördü (%50) Türkçe sorulduğunda yanlış cevaplanırken, dördü (%50) Türkçe sorulduğunda doğru cevaplandı. Gemini'nin Aynı soruları İngilizce ve Türkçe olarak doğru olarak cevaplamasında istatistiksel olarak anlamlı düzeyde bir fark izlenmedi (p=0,727 McNemar testi) (Tablo 1).

Tablo 1. Yapay zeka sohbet botlarının aynı sorulara verdikleri cevaplar ve değişimleri

Cevaplar	ChatGPT-3,5 (İngilizce)	ChatGPT-3,5 (Türkçe)	Copilot (İngilizce)	Copilot (Türkçe)	Gemini (İngilizce)	Gemini (Türkçe)
Doğru	27 (%75)	23 (%63,9)	24 (%66,7)	24 (%66,7)	23 (%63,9)	25 (%69,4)
Yanlış	9 (%25)	13 (%36,1)	12 (%35,3)	12 (%35,3)	13 (%36,1)	11 (%30,6)
P değeri	0,388*		1,0*		0,727*	
Aynı cevabı üretme	24 (%66,7)		26 (%72,2)		28 (%77,8)	
Farklı cevabı üretme	12 (%33,3)		10 (%27,8)		8 (%22,2)	
Doğru-yanlış değişimi	8 (%66,7)		5 (%50)		4 (%50)	
Yanlış-doğru değişimi	4 (%33,3)		5 (%50)		4 (%50)	

**: McNemar testi*

Tartışma

ChatGPT insan düşünce yapısını taklit ederek insan benzeri cevaplar oluşturmayı amaçlayan 175 milyar gibi çok geniş bir bilgi ağı ile eğitilmiş, BDM temelli yapay zeka programları arasında bu özellikleri ile öne çıkmış ve kendisine has bir yer oluşturmuş bir yapay zeka sohbet botudur (7). En son Eylül 2021 de güncelleme alan ve internet erişimi olmayan ChatGPT-3.5 ise bu gelişim zincirinin son ücretsiz halkasını oluşturmaktadır (8). BDM grubunun diğer önemli bir temsilcisi olan ve ücretsiz erişim sağlanabilen Copilot ise Şubat 2023 tarihinde GPT-4'ün entegrasyonu ile birlikte sürekli gelişiminin devam ettiği bir yapay zeka sohbet botudur (8). Copilot sohbet botu aynı zamanda aldığı kaynakları alıntılayarak okuyucunun daha ayrıntılı bilgi edinebilmesi için bir yol gösterici olabilmektedir (8 ,9). Gemini ise BDM grubunun diğer önemli ücretsiz erişim sağlanabilen bir üyesidir. Copilotla çevrimiçi internet erişiminin olması ve sürekli güncellenmesi gibi ortak özelliklerinin yanında karşılaştığı çeşitli sorunlar karşısında kesin yanıtlar sunabilme özelliği ile benzerlerinden ayrılmaktadır (10, 11). Yapay zeka sohbet botları bu çok çeşitli avantajları göz önüne alındığında tıp eğitiminde de kendisine çok çeşitli yerler bulmuş, bilgiye hızlı erişmek isteyen, literatür taramak veya dil çevirisi yapmak isteyen tıp fakülte öğrencilerinden sağlık profesyonellerine kadar çok geniş bir alanı etkisi altında bırakmıştır (12, 13). Fakat bu programların bazı kısıtlılıklarının olması programlarının etkinliğini ve güvenilirliğini değiştirebilmektedir. Bu kısıtlılığın bir örneği çevrimiçi internet erişimi olan bu sohbet botlarının ücretli sitelere girişlerinin kısıtlılığıdır. Bu da doğru bilgi için güncel olan bilgiye erişimin aksamasına neden olacağı düşüncesidir (14, 15). Yapay zeka sohbet botlarının bu ve benzeri kısıtlılıkları düşünüldüğünde bu konuda daha ayrıntılı bilgi edinmek ve bu programlarının performanslarını incelemek amacı ile çok çeşitli konularda araştırmalar yapılmıştır. Bu konulardan biri bu programların çoktan seçmeli sorulardaki başarısını ve yeterliliğinin araştırılmasıdır. Bu konu oftalmoloji alanında sık olarak gündeme gelmiş ve yoğun bir şekilde araştırılmıştır. ChatGPT-3.5 ve ChatGPT-4.0'ın 380 oftalmoloji sorusuna verdikleri cevapları inceleyen bir araştırmada ChatGPT-3.5 soruların %55'ine doğru cevap verirken, ChatGPT-4.0 soruların %70'ine doğru cevap vermiştir. Retina ve oküler onkoloji alanındaki bilgileri sorgulayan sorulara ise ChatGPT-3.5 %54, ChatGPT-4.0 %73 oranında doğru cevap vermiştir. Araştırıcılar bu çalışmasının sonunda her ne kadar bu programların şu an için yeterli olmadıklarını belirtseler de gelecekteki gelişmelerle tıp eğitiminde önemli bir rol üstlenebileceklerini belirtmektedirler (16). ChatGPT-3.5, ChatGPT-4.0 ve insan katılımcıların 467 oftalmoloji sorusunu cevapladığı ve başarılarının değerlendirildiği çalışmada ChatGPT-3.5 %55,46, ChatGPT-4.0 %73,2 ve insan katılımcılar %58,1 oranında başarı göstermişlerdir. Oküler patolojiler ve tümörler ile ilgili sorularda ise sırası ile başarıları %45, %70 ve %58 düzeylerinde olduğu belirtilmiştir. Yazarlar sonuç olarak da yapay zeka programlarının tıp eğitimi ve uygulamalarında önemli bir araç olarak kullanılabileceği görüşünü savunmuşlardır (17). Tüm bunlara ek olarak bir başka araştırma ise Türkçe oftalmoloji sorularında ChatGPT-3.5, ChatGPT-4.0, Bing ve Bard'ın etkinliklerini araştırmışlar ve bu programların sırası ile %51, %77,5, %63 ve %45,5 doğrululuk oranlarına sahip olduklarını belirtmişlerdir ve sonuç olarak bu programların hala geliştirilmeye ihtiyacının olduğunu vurgulamışlardır (18). Yapay zeka sohbet botlarının performansının incelendiği bir farklı konu ise soruların sorulduğu bölgeye göre doğru cevaplanma düzeylerinde bir farklılığın oluşup olmadığının araştırılmasıdır. Gemini'nin farklı ülkelerin internet ağından aynı sorulardaki performansı değerlendirilmiştir. Programın performansı her ne kadar kabul edilebilir olduğu belirtilse de farklı ülkelerden erişimin başarı düzeylerini etkilediği ifade edilmiştir (19).

Bu çalışmada yapay zeka sohbet botlarının aynı soruların farklı dillerde sorulmasının başarılarını etkileyip etkilemediğini araştırmak istedik. Kendi araştırmamızda herkesin rahat erişim sağlayabileceği ve ücretsiz olduğu için gündelik hayatta daha kolay ve sık kullanılabileceğini düşündüğümüz ChatGPT-3,5, Copilot ve Gemini sohbet botlarını çalışmamıza dahil ettik. Kendi serimizde soruların İngilizce ve Türkçe cevaplamasındaki başarı oranları makul görülse bile eşit dü-

zeyde değildi. Her ne kadar farklı dillerdeki başarı oranlarında istatistiksel düzeyde anlamlı fark mevcut olmasa da ChatGPT ve Gemini'nin başarı düzeyleri birbirlerinden farklıydı. Ayrıca Copilot'un her ne kadar başarı düzeyleri aynı gibi gözükse de soru bazlı incelendiğinde aynı sorulara farklı yanıtlar ürettiklerini gözlemledik. Biz bu farklılığın sebebinin programların dil çeviri yeteneklerinin sınırlı olması ve soruları doğru yorumlama ya da erişilen bilgiyi uygun bir şekilde anlayıp, değerlendirme yeteneğinin kısıtlı olmasına bağlı olabileceğini düşünmekteyiz. Her ne kadar başarı oranları makul gözükse de dil farklılığına bağlı cevapların değişmesi araştırıcıyı bu programlara danışma sırasında kullanması gereken dil konusunda ikileme götürecek ve doğru bilgiye erişim konusunda bu programlara olan güveni zedeleyecektir.

Oftalmoloji 2023-2024 Basic and Clinical Science Course (BCSC) Oftalmik Patolojiler ve İntraoküler Tümörler kitabının çalışma soruları 36 soruyu içeriyordu ve sohbet robotlarına bunları sorduk. Soru sayısının az olmasının istatistiksel sonucun anlamlı çıkıp çıkmamasına etki edebileceğini düşünmekteyiz fakat temel bilgileri ölçen bu kitabın sorularına ilave soru eklemenin yanlış olacağı kanaatine vardık. Daha fazla soru içeren testlerde farklı değerlerin çıkıp çıkmayacağının araştırılması gerektiğini düşünmekteyiz.

Soru sayımızın azlığı, buna bağlı olarak soruların alt konu dallarına ayrılmayışı, soruların zorluk ve karmaşıklık düzeylerine göre değerlendirilmemesi, sorulara verilen cevapların sadece doğru ve yanlış olarak kategorize edilmesi ve yapay zeka programları ile insan performansının karşılaştırılıp incelenmemiş olması çalışmamızın en önemli kısıtlı yönlerini oluşturmaktadır.

Sonuç olarak çalışmamız üç farklı üretici tarafından piyasaya sürülmüş olan ChatGPT-3.5, Copilot ve Gemini adlı yapay zeka sohbet botlarının oküler patolojiler ve intraoküler tümörler ile ilgili çoktan seçmeli farklı dillerde aynı soruları cevaplamadaki performansının değerlendirildiği ilk araştırmadır. Bu programların geniş bilgilerle beslenmesi ve geliştirilmesi sıklıkla gündemde olan önemli bir konu olmakla birlikte farklı dillerde aynı soruları cevaplamadaki bu farklılıkta düzeltilmesi ve geliştirilmesi gereken diğer önemli bir hususu oluşturmaktadır. Kullanıcılar açısından tutarlı ve objektif bir ortamın sağlanması, bu programların oftalmoloji eğitiminde daha önemli sorumluluklar üstlenmesine aracılık edecektir.

Etik onam: *Çalışmamızdaki veriler, insan ve hayvan denekleri kaynaklı olmadığı için etik kurul onayı gerektirmemektedir.*

Yazar Katkıları:
Konsept: E.Ş.
Literatür Tarama: E.Ş.
Tasarım: E.Ş., M.Ç.
Veri toplama: E.Ş.
Analiz ve yorum: E.Ş., M.Ç.
Makale yazımı: E.Ş.
Eleştirel incelenmesi: E.Ş., M.Ç.

Çıkar Çatışması*: Yazarlar çıkar çatışması olmadığını bildirmişlerdir.*

Finansal Destek: *Bu çalışma herhangi bir fon tarafından desteklenmemiştir.*

Kaynaklar

1. Rahimy E. Deep learning applications in ophthalmology. Curr Opin Ophthalmol. 2018;29(3):254-260.
2. Ting DSW, Pasquale LR, Peng L, Campbell JP, Lee AY, Raman R, et al. Artificial intelligence and deep learning in ophthalmology. Br J Ophthalmol. 2019;103(2):167-175.
3. Antaki F, Coussa RG, Kahwati G, Hammamji K, Sebag M, Duval R. Accuracy of automated machine learning in classifying retinal pathologies from ultra-widefield pseudocolour fundus images. Br J Ophthalmol. 2023;107(1):90-95.
4. Schmidt-Erfurth U, Sadeghipour A, Gerendas BS, Waldstein SM, Bogunović H. Artificial intelligence in retina. Prog Retin Eye Res. 2018;67:1-29.
5. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language Models are Unsupervised Multitask Learners. Accessed June 26, 2023. https://github.com/codelucas/newspaper
6. Syed NA, Berry JL, Heegaard S, Kramer TR, Lee V, Raparia K, et al., eds. Ophthalmic Pathology and Intraocular Tumors. American Academy of Ophthalmology; 2023.
7. Wen J, Wang W. The future of ChatGPT in academic research and publishing: A commentary for clinical and translational medicine. Clin Transl Med. 2023;13(3): e1207
8. Tao BKL, Hua N, Milkovich J, Micieli JA. ChatGPT-3.5 and Bing Chat in ophthalmology: an updated evaluation of performance, readability, and informative sources. Eye 2024. Published online March 20, 2024:1-6.
9. Bing Chat | Microsoft Edge. Accessed July 4, 2024. https://www.microsoft.com/en-us/edge/features/bing-chat?form=MT00D8
10. Waisberg E, Ong J, Masalkhi M, Zaman N, Sarker P, Lee AG, et al. Google's AI chatbot "Bard": a side-by-side comparison with ChatGPT and its utilization in ophthalmology. Eye 2023 38:4. 2023;38(4):642-645.
11. Google AI updates: Bard and new AI features in Search. Accessed July 4, 2024. https://blog.google/technology/ai/bard-google-ai-search-updates/
12. Khan RA, Jawaid M, Khan AR, Sajjad M. ChatGPT - Reshaping medical education and clinical management. Pak J Med Sci. 2023;39(2):605.
13. Jeblick K, Schachtner B, Dexl J, Mittermeier A, Stüber AT, Topalis J, et al. ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. Eur Radiol. 2024;34(5):2817-2825.
14. Korngiebel DM, Mooney SD. Considering the possibilities and pitfalls of Generative Pre-trained Transformer 3 (GPT-3) in healthcare delivery. npj Digital Medicine 2021 *4:1*. 2021;4(1):1-3.
15. Nath S, Marie A, Ellershaw S, Korot E, Keane PA. New meaning for NLP: the trials and tribulations of natural language processing with GPT-3 in ophthalmology. Br J Ophthalmol. 2022;106(7):889-892.
16. Haddad F, Saade JS. Performance of ChatGPT on Ophthalmology-Related Questions Across Various Examination Levels: Observational Study. JMIR Med Educ. 2024;10:e50842.
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(6):e40822.
18. Canleblebici M, Dal A, Erdağ M. Evaluation of the Performance of Large Language Models (ChatGPT-3.5, ChatGPT-4, Bing and Bard) in Turkish Ophthalmology Chief-Assistant Exams: A Comparative Study. Turkiye Klinikleri J Ophthalmol. Published online June 11, 2024.
19. Mihalache A, Grad J, Patil NS, Huang RS, Popovic MM, Mallipatna A, Kertes PJ, Muni RH. Google Gemini and Bard artificial intelligence chatbot performance in ophthalmology knowledge assessment. Eye (Lond). 2024;38(13):2530-2535..