



# Refraktif Cerrahide Sık Sorulan Sorular için ChatGPT 3.5, ChatGPT 4.0, Gemini ve Microsoft Copilot Tarafından Oluşturulan Yanıtların Okunabilirliği ve Uygunluğu

## Readability and Appropriateness of Responses Generated by ChatGPT 3.5, ChatGPT 4.0, Gemini, and Microsoft Copilot for FAQs in Refractive Surgery

✉ Fahri Onur Aydın, ✉ Burakhan Kürşat Aksoy, ✉ Ali Ceylan, ✉ Yusuf Berk Akbaş, ✉ Serhat Ermiş, ✉ Burçin Kepez Yıldız, ✉ Yusuf Yıldırım

Sağlık Bilimleri Üniversitesi, Başakşehir Çam ve Sakura Şehir Hastanesi, Göz Hastalıkları Kliniği, İstanbul, Türkiye

### Öz

**Amaç:** Refraktif cerrahi ile ilgili sıkça sorulan sorulara (SSS) büyük dil modeli (BDM) sohbet robotlarının verdiği cevapların uygunluğunu ve okunabilirliğini değerlendirmek.

**Gereç ve Yöntem:** Yaygın olarak kullanılan 4 BDM sohbet robotuna hastalar tarafından sıkça sorulan refraktif cerrahi ile ilgili 40 soru soruldu. Cevapların uygunluğu, deneyimli 2 refraktif cerrah tarafından değerlendirildi. Okunabilirlik ise 5 farklı indeks ile değerlendirildi.

**Bulgular:** BDM sohbet robotları tarafından üretilen cevaplara dayanarak, ChatGPT 3.5 tarafından verilen cevapların %45'i (n=18) doğruyken, ChatGPT 4.0'ın %52,5'i (n=21), Gemini'nin %87,5'i (n=35) ve Copilot'un %60'ı (n=24) doğruydu. Okunabilirlik açısından, tüm BDM sohbet robotlarının çok zor okunduğu ve yanıtların üniversite düzeyinde olduğu gözlemlendi.

**Sonuç:** Günlük yaşamımızda yer edinmeye başlayan bu BDM sohbet robotları nadiren uygunsuz cevaplar verebilmektedir. Aynı zamanda okunması oldukça zor olan bu BDM sohbet robotları arasında, uygun cevap üretmede en başarılı olanı Gemini olup, okunabilirlik açısından diğerlerine göre göreceli olarak daha iyi bulunmuştur.

**Anahtar Kelimeler:** Yapay zeka, sohbet robotu, refraktif cerrahide SSS, ChatGPT, Gemini, Copilot

### Abstract

**Objectives:** To assess the appropriateness and readability of large language model (LLM) chatbots' answers to frequently asked questions about refractive surgery.

**Materials and Methods:** Four commonly used LLM chatbots were asked 40 questions frequently asked by patients about refractive surgery. The appropriateness of the answers was evaluated by 2 experienced refractive surgeons. Readability was evaluated with 5 different indexes.

**Results:** Based on the responses generated by the LLM chatbots, 45% (n=18) of the answers given by ChatGPT 3.5 were correct, while this rate was 52.5% (n=21) for ChatGPT 4.0, 87.5% (n=35) for Gemini, and 60% (n=24) for Copilot. In terms of readability, it was observed that all LLM chatbots were very difficult to read and required a university degree.

**Conclusion:** These LLM chatbots, which are finding a place in our daily lives, can occasionally provide inappropriate answers. Although all were difficult to read, Gemini was the most successful LLM chatbot in terms of generating appropriate answers and was relatively better in terms of readability.

**Keywords:** Artificial intelligence, chatbots, refractive surgery FAQs, ChatGPT, Gemini, Copilot

### Giriş

Yapay zekanın ("artificial intelligence", AI) sağlık hizmetlerine hızlı entegrasyonu, hasta katılımını ve bilgi yayılımında dönüşüme neden olmuştur. AI modelleri giderek primer tıbbi bilgi kaynağı haline geldiğinden, tıbbi sorulara verdikleri yanıtların fizibilitesini ve doğruluğunu değerlendirmek önemlidir.<sup>1,2</sup> Doğal dil işlemedeki gelişmelerin yönlendirdiği konuşan robotlarının yükselişi, sağlık sektöründe umut verici yeni bir döneme işaret ediyor. Bu robotlar, hastalıkları önleme, tanı, tedavi, izlem ve hasta desteği gibi çeşitli tıbbi alanlarda dikkate değer bir potansiyel göstermektedir.<sup>3</sup>

OpenAI'nın ChatGPT, Google'ın Gemini ve Microsoft'un Copilot gibi büyük dil modeli (BDM) sohbet robotları, AI teknolojisinde önemli bir sıçramayı temsil etmektedir. Bu

**Cite this article as:** 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

**Yazışma Adresi/Address for Correspondence:** Fahri Onur Aydın, Sağlık Bilimleri Üniversitesi, Başakşehir Çam ve Sakura Şehir Hastanesi, Göz Hastalıkları Kliniği, İstanbul, Türkiye

**E-posta:** onuraydin90@hotmail.com **ORCID-ID:** orcid.org/0000-0001-9414-6788

**Geliş Tarihi/Received:** 16.07.2024 **Kabul Tarihi/Accepted:** 11.11.2024

DOI: 10.4274/tjo.galenos.2024.28234



modeller, büyük eğitim veriler setleri ve sofistike algoritmalarından yararlanarak çeşitli metin tabanlı sorgulara insan benzeri yanıtlar üretmek için tasarlanmıştır.<sup>4</sup> Kendinden denetimli öğrenme ve geniş metinsel verilerle eğitim ile karakterize BDM sohbet robotlarının evrimi, insan etkileşimlerini yakından taklit eden yanıtlar üretmelerini sağlamıştır. Ayrıntılı ve konu ile ilgili bilgi sağlayabilmeleri, onları tıbbi uygulamalar için özellikle değerli kılar.<sup>5,6</sup>

Oftalmoloji alanında, özellikle refraktif cerrahide, hastalar durumları ve tedavi seçenekleri hakkında bilgi edinmek için sıklıkla internete başvurmaktadır. Bu bilgilerin kalitesi ve okunabilirliği, hastanın konuyu kavrayışını ve karar vermelerini doğrudan etkilediği için çok önemlidir. BDM sohbet robotlarının tıbbi tavsiyelerde bulunmak açısından olası yararlarına rağmen, doğru ve anlaşılır bilgi sunmadaki etkinlikleri hala kapsamlı bir değerlendirme gerektirmektedir.

Bu çalışmada, refraktif cerrahi hakkında güvenilir ve erişilebilir bilgi sağlamada farklı BDM sohbet robotlarının güçlü yönlerini ve kısıtlılıklarının araştırılması amaçlanmıştır. Bu araştırma, yanıtlarının uygunluğunu ve okunabilirliğini değerlendirerek, AI odaklı hasta eğitimini geliştirmeyi ve böylece hastaların göz sağlığı hakkında bilinçli kararlar vermeleri için doğru ve anlaşılır bilgiler almasını sağlamayı amaçlamaktadır.

## Gereç ve Yöntem

Çalışmamıza herhangi bir hasta dahil edilmediğinden etik kurul onayı gerekmemiştir.

Bu çalışma, BDM sohbet robotları tarafından sağlanan bilgilerin uygunluğunu ve okunabilirliğini araştırmak için tasarlanmıştır. Çalışma için yeni geliştirilen ve sık kullanılan dört BDM sohbet robotu (ChatGPT 3.5, ChatGPT 4.0, Google Gemini ve Microsoft Copilot) seçildi. Refraktif cerrahlardan, hastaların refraktif cerrahi ile ilgili sıklıkla hasta portalı üzerinden veya klinikte sordukları 40 sorudan oluşan bir liste hazırlamaları istendi. Bu sorular daha sonra 3 Temmuz 2024'te BDM sohbet robotları tarafından yanıtlandı. Cevaplar iki deneyimli refraktif cerrah (Y.Y., B.K.Y.) tarafından uygunluk ve yeterlilik açısından değerlendirildi. Cevaplar "uygun", "eksik" ve "uygunsuz" olarak kategorize edildi. Uygun bir yanıt, yanıtları değerlendiren hekimin hastalara vereceği önerilere benzer doğru bir yanıt olarak tanımlandı. Uygunsuz bir yanıt ya yanlış ya da klinik açıdan yanıtları değerlendiren hekimin önerisinden farklıydı. Eksik bir yanıt konu ile ilgili ve doğruydur, ancak yeterli bilgi içermiyordu.

Ortalama bir bireyin her bir cevabı okuma kolaylığını değerlendirmek için, cevaplar Readable adlı çevrimiçi bir okunabilirlik uygulamasına girildi (<https://app.readable.com/text/>).<sup>7</sup> Çalışmada kullandığımız okunabilirlik ve anlaşılabilirlik kriterleri ve standardizasyon İngilizce temelli idi. Çalışmamızda soruları İngilizce olarak sorduk ve cevapları İngilizce olarak aldık. Her cevabın okunabilirliğini değerlendirmek için beş farklı endeks kullanıldı. Bu endeksler Gunning Fog Endeksi, Coleman-Liau Endeksi, Flesch Okuma Kolaylığı Skoru, Flesch-Kincaid Sınıf Düzeyi ve SMOG Endeksi ("Simple Measure

of Gobbledygook") idi.<sup>8</sup> Flesch okuma testinde kullanılan matematiksel formüller kelime karmaşıklığı ve cümle uzunluğuna dayanmaktadır. Flesch Okuma Kolaylığı skoru 1 ile 100 arasında sayısal bir değerdir. Skorun büyük olmasını okunabilirliğin yüksek olduğunu gösterir ve 70 ile 80 arasındaki bir skor 8. sınıf seviyesine karşılık gelir.<sup>7</sup> Gunning Fog Endeksi, ortalama cümle uzunluğu ile birlikte çok heceli kelimelerin sıklığını değerlendirir.<sup>9</sup> Endeks skoru 0 ile 20 arasında değişir ve metnin basitliği ve anlaşılabilirliğini derecelendirir.<sup>7</sup> Coleman-Liau Endeksi, tıbbi verilerin değerlendirilmesine yardımcı olur ve tipik olarak diğer endekslerle birlikte kullanılır.<sup>1</sup> Ortalama cümle uzunluğuna ve her yüz kelimedeki ortalama harf sayısına odaklanır.<sup>9</sup> SMOG Endeksi, cümle örneklerinde çok heceli kelimelerin sıklığını kullanır.<sup>9</sup> Yaygın olarak kullanılmasına rağmen, SMOG en sık sağlık hizmetlerinde tercih edilmektedir.<sup>10</sup> Son üç endeksin sonuçları, bir öğrencinin yazılı materyali anlamak için Amerika Birleşik Devletleri'nde okuması gereken sınıf seviyesine karşılık gelir. Bu nedenle, Gunning Fog, Coleman-Liau ve SMOG endex değerleri daha düşük olan metinlerin okunması ve anlaşılması daha kolay olmalıdır.<sup>11</sup>

## İstatistiksel Analiz

İstatistiksel analizler SPSS programı (IBM SPSS Statistics, sürüm 25; IBM Corp., Armonk, NY, ABD) kullanılarak yapıldı. Tanımlayıcı analizler yapıldı ve normal dağılıma uygunluk (Shapiro-Wilk) test edildi. Verilerin normal dağılım göstermediği göz önüne alınarak, dört BDM sohbet robotunun ortalama puanlarını karşılaştırmak için parametrik olmayan bir Kruskal-Wallis testi ve Bonferroni düzeltmesi yapıldı. İstatistiksel açıdan düzeltilmiş p değerinin 0,05'ten küçük olması anlamlı kabul edildi.

## Bulgular

### Uygunluk

BDM sohbet robotları tarafından oluşturulan cevaplar değerlendirildiğinde ChatGPT 3.5 tarafından verilen cevapların %45'i (n=18) doğru iken, ChatGPT 4.0'ın %52,5'i (n=21), Gemini'nin %87,5'i (n=35) ve Copilot cevaplarının %60'ı (n=24) doğruydur. ChatGPT 3.5, ChatGPT 4.0 ve Copilot'tan her biri bir soruya uygunsuz cevaplar verirken, Gemini hiçbir soruya uygunsuz cevap vermedi ([Şekil 1](#)).

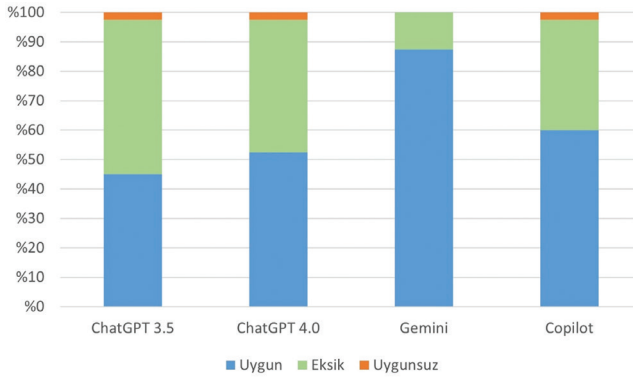
BDM sohbet robotları uygunluk açısından karşılaştırıldığında istatistiksel olarak anlamlı farklılık olduğu bulundu (p=0,001). Alt grup analizi yapıldığında Gemini ile ChatGPT 3.5 ve 4.0 arasında bu fark gözlemlendi (sırasıyla p=0,001, p=0,008) ([Tablo 1](#)).

### Okunabilirlik

Okunabilirlik endeksleri [Tablo 2](#)'de özetlenmiştir. Flesch-Kincaid Sınıf Düzeyi'ne göre karşılaştırıldığında BDM sohbet robotları arasında anlamlı farklılık gözlemlendi (p=0,003). İkili karşılaştırmalarda, bu farkın ChatGPT 3.5 ile Gemini ve ChatGPT 3.5 ile Copilot arasında olduğu ve ChatGPT 3.5'in değerlerinin anlamlı düzeyde daha yüksek olduğu bulundu (sırasıyla p=0,017 ve p=0,008; [Şekil 2a](#)). Diğer sohbet robotları arasında anlamlı bir fark izlenmedi. Sohbet robotları arasında

Flesch Okuma Kolaylığı skorları ( $p=0,534$ ; Şekil 2b) veya Coleman-Liau skoru ( $p=0,867$ ; Şekil 2c) açısından anlamlı bir fark yoktu. SMOG endeksi sonuçları karşılaştırıldığında sohbet robotları arasında anlamlı fark gözlemlendi ( $p=0,012$ ). Bu farkın, Copilot SMOG skorunun ChatGPT 3.5'e kıyasla anlamlı düzeyde düşük olmasından kaynaklandığı bulundu (Şekil 2d). Gunning Fog skorları değerlendirildiğinde yine gruplar arasında anlamlı fark gözlemlendi ( $p=0,001$ ). İkili karşılaştırmalar, Copilot skorlarının hem ChatGPT 3.5 hem de ChatGPT 4.0'dan anlamlı düzeyde düşük olduğunu gösterdi (sırasıyla  $p=0,003$  ve  $0,021$ ) (Şekil 2e).

Kelime ve karakter sayıları karşılaştırıldığında, Gemini'nin diğer BDM sohbet robotlarından anlamlı düzeyde daha uzun yanıtlar verdiği görüldü (her ikisi için de  $p<0,001$ ). Gemini'nin verdiği yanıtların kelime ve karakter sayıları, Copilot'a (her ikisi için de  $p=0,001$ ) ve ChatGPT 3.5 ve 4.0'a kıyasla (hepsi için  $p<0,001$ ) anlamlı düzeyde daha yüksekti. ChatGPT sürümlerinin verdiği yanıtların kelime ve karakter sayıları benzerdi (Tablo 1).



**Şekil 1.** İki deneyimli refraktif cerrah tarafından belirlenen refraktif cerrahi ile ilgili sorulara büyük dil modeli sohbet robotlarının yanıtlarının fikir birliğine dayalı doğruluk derecelendirmeleri

	ChatGPT 3.5	ChatGPT 4.0	Gemini	Copilot	p değeri
<b>Uygunluk</b>	2,42±0,54	2,50±0,55	2,87±0,33	2,57±0,54	0,001
<b>Kelime sayısı</b>	21,15±4,40	21,67±6,24	318,62±73,98	103,90±46,44	<0,001
<b>Karakter sayısı</b>	115,00±24,33	118,65±32,97	1767,02±450,00	587,15±260,76	<0,001

	ChatGPT 3.5	ChatGPT 4.0	Gemini	Copilot	p değeri
<b>Coleman-Liau</b>	14,86±3,90	14,99±5,11	14,60±1,58	14,88±2,51	0,867
<b>Flesch Okuma Kolaylığı Skoru</b>	30,97±22,49	31,79±26,72	37,39±7,98	32,76±12,88	0,534
<b>Flesch-Kincaid Sınıf Düzeyi</b>	13,95±3,38	13,49±3,92	11,76±1,35	11,71±1,89	0,003
<b>SMOG Endeksi</b>	15,03±2,80	14,50±3,07	13,93±1,35	13,34±1,28	0,012
<b>Gunning Fog</b>	16,30±3,96	15,74±4,73	14,09±1,65	13,48±2,08	0,001

SMOG: Gobbledygook'un Basit Ölçümü

## Tartışma

AI kullanımı dünya çapında giderek yaygınlaşmaktadır. Kullanımın artması ile birçok yeni AI modeli geliştirilmektedir. Bu modeller arasında, öğrenilen verileri kullanarak internette gezinen ve sohbet robotu konuşmalarında anında yanıtlar üreten eğitilmiş dil modelleri yer alır.<sup>12</sup> Bu makale, bu varyasyonun BDM sohbet robot performansını ve yanıt kalitesini nasıl etkilediğine dair derinlemesine bir analiz sunarak, farklı BDM sohbet robotlarının yanıtları arasındaki farklılıkların esas olarak kullanılan algoritmalarındaki farklılıklardan kaynaklandığını vurgulamaktadır.

Günümüzde birçok kişi BDM sohbet robotlarını çeşitli amaçlarla kullanmaktadır. Bunlardan biri sağlık hakkındaki sorularına cevap almaktır. Ancak, sağlıkla ilgili bilgi edinmek için AI'yı kullanmak bazı sorunlar neden olabilir. Bu sorunlar arasında, güncel olmayan veya yanlış bilgilerin elde edilmesi ve karmaşık bir şekilde sunulan doğru bilgilerin yanlış anlaşılması yer alır. Bu nedenle bu bilgilerin hem doğru hem de herkes tarafından anlaşılır olması çok önemlidir.

Çalışmamızda, sohbet robotların yanıtlarının uygunluğu değerlendirildiğinde, Gemini'nin soruları diğer BDM sohbet robotlarına göre anlamlı derecede daha yüksek oranda doğru yanıtladığı görüldü. Bulgularımızın aksine Tepe ve Emekli<sup>13</sup> ChatGPT 4.0, Gemini ve Copilot'u karşılaştırdıkları bir çalışmada meme görüntüleme ile ilgili sorulara en uygun cevapları ChatGPT 4.0'ın verdiğini bildirmişlerdir. Başka bir çalışmada Lee ve ark.<sup>14</sup>, hipertansiyon eğitimi için kaynak olarak Gemini ve ChatGPT 3.5'i karşılaştırmış ve benzer sonuçlar elde edildiğini belirlemişlerdir.

Çalışmamızda, kapsamlı sonuçlar elde etmek için beş farklı geçerliliği kabul edilen okunabilirlik endeksi kullanılmıştır. Bu endekslere göre, BDM sohbet robotlarının ürettiği yanıtlar düşük okunabilirlik skorlarına sahipti. Flesch Okuma Kolaylığı skorları 30 ila 50 arasında değişirken Gemini en yüksek puanı (yani en okunabilir cevaplar) aldı. Bu, metinlerin üniversite öğrencileri

tarafından anlaşılabilirliğini ve zorluk seviyesinin yetişkinlerin sadece %33'ü için uygun olduğunu düşündürmektedir.<sup>15</sup> Flesch-Kincaid Sınıf Düzeyi sonuçları değerlendirildiğinde, ChatGPT yanıtlarının 14. sınıf ve üstü kişiler için uygun olduğu, Gemini ve Copilot'un ise 12. sınıf ve üstü kişiler için uygun olduğu bulunmuştur. Bu sonuçlar Gemini ve Copilot'un okunabilirliğinin biraz daha iyi olduğunu düşündürmektedir.<sup>16</sup> Gunning Fog Endeksi'ne göre de tüm BDM sohbet robotlarının düzeyi üniversite seviyesindedir. Bununla birlikte, ChatGPT yanıtlarının lisans son sınıf öğrencileri düzeyinde, Gemini ve Copilot'un ise lisans birinci sınıf düzeyinde olduğu görülmüştür. Coleman-Liau Endeksi sonuçları tüm BDM sohbet robotları için benzer bulunurken lisans seviyesinde ve okunması zordu.<sup>17</sup> Benzer şekilde, tüm BDM sohbet robotlarının SMOG Endeksi değerleri, yanıtlarının lisans düzeyinde olduğunu ve genel çoğunluk için okunmasının zor olduğunu gösterdi.<sup>10</sup>

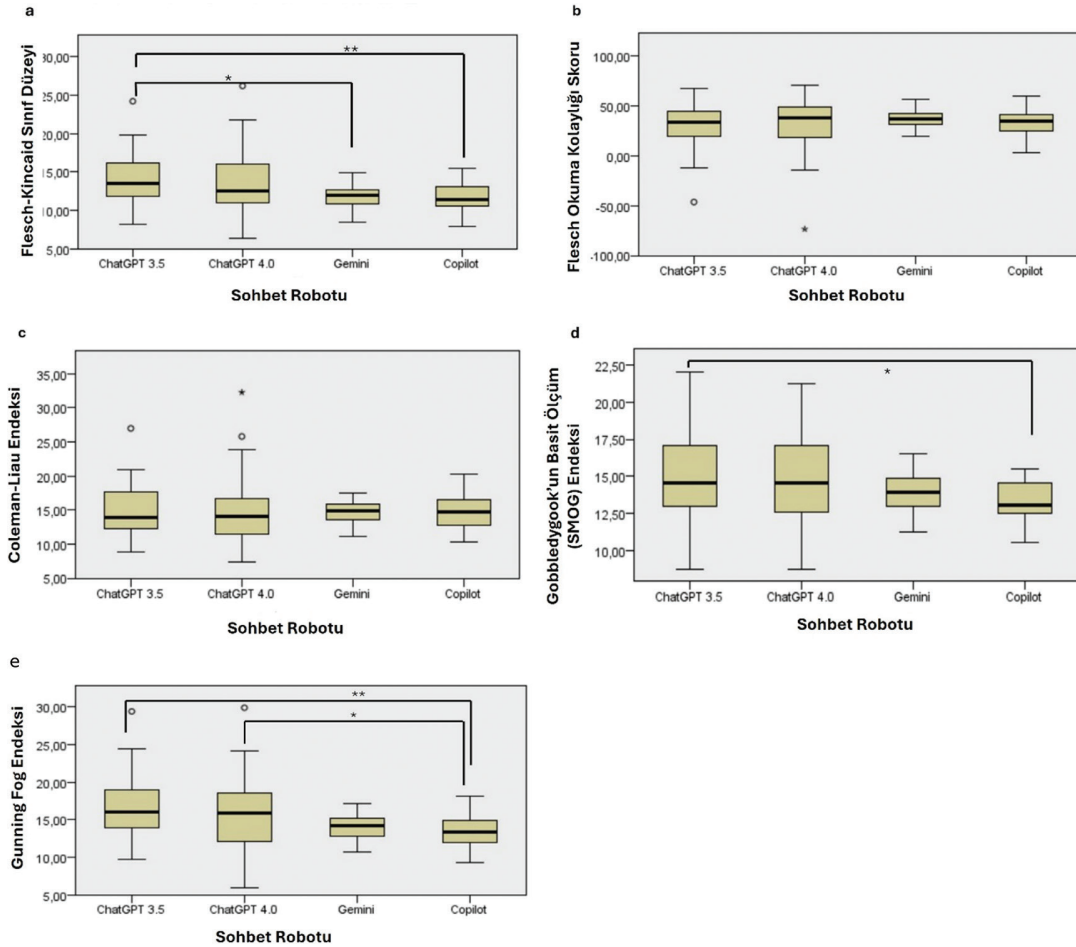
ChatGPT 4.0 ile yapılan bir çalışmada, okunabilirlik analizlerinin sonuçları bizim çalışmamızda elde edilen sonuçlara benzerdi, bu da lisans veya yüksek lisans seviyesine işaret ediyordu ve okunması zordu.<sup>18</sup> Tepe ve Emekli<sup>13</sup> tarafından yapılan

çalışmada, ChatGPT 4.0, Gemini ve Copilot okunabilirlik açısından karşılaştırılmış, ChatGPT 4.0'in en zor ve Gemini'nin nispeten daha kolay olduğu, ancak hepsinin okunabilirliğinin düşük olduğu bildirilmiştir.

Kelime ve karakter sayısı değerlendirildiğinde, her iki ChatGPT sürümünün de diğer BDM sohbet robotlarına göre anlamlı düzeyde daha az kelime ve karakter kullandığı bulundu. En fazla sayıda kelime ve karakteri Gemini kullandı. Yanıtları anlamlı düzeyde daha uzun olmasına rağmen, Gemini'nin yanıtlarının okunabilirlik ve doğruluğu daha iyi bulunmuştur.

Literatürde BDM sohbet robotları ile ilgili yapılan diğer çalışmalarda da benzer metodoloji kullanılmış olsa da hasta kohortu kullanılması durumunda daha bütüncül bir değerlendirme mümkün olabilir. Bu fikir gelecekteki araştırmalara yön verebilir.

Sonuçların gösterdiği gibi, BDM sohbet robotları eksik veya bazen yanlış bilgiler verebilir. Ayrıca, verdikleri bilgiler doğru olsa bile, okunabilirliğin nispeten düşük olması nedeniyle hastaları yanıltma olasılığı da vardır. Bu durum hastalar için bir risk oluşturmaktadır. Bu olasılıkları azaltmak için, yeni BDM



**Şekil 2.** Büyük dil modeli sohbet robotlarının okunabilirlik puanları. a) Flesch-Kincaid Sınıf Düzeyi, b) Flesch Okuma Kolaylığı skoru, c) Coleman-Liau Endeksi, d) Gobbledygook'un Basit Ölçüm (SMOG) Endeksi, e) Gunning Fog Endeksi



sohbet robotlarının özellikle sağlıkla ilgili konularda sağlık çalışanları ile işbirliği yapılarak geliştirilmesi, doğruluğu ve erişilebilirliği iyileştirmede faydalı olabilir.

### Çalışmanın Kısıtlılıkları

Bu çalışmanın bazı kısıtlılıkları bulunmaktadır. İlk olarak, araştırma 40 soru ile sınırlandırılmıştır, bu da bulguların genellenebilirliğini sınırlayabilir. Ayrıca, BDM sohbet robotları ile etkileşime girerken girdilerin formülasyonu, üretilen yanıtların kalitesini ve doğasını önemli ölçüde etkileyebilmektedir. BDM sohbet robotlarının tekrarlanabilirliği de tartışmalıdır. Bu çalışmada, her soru BDM sohbet robotlarına yalnızca bir kez gönderilmiştir. Ayrıca cevapların okunabilirliğinin gerçek hastalar tarafından değerlendirilmemiş olması çalışmanın bir diğer kısıtlılığıdır.

### Sonuç

Sonuç olarak, Gemini'nin refraktif cerrahi ile ilgili sorulara uygun cevaplar vermede diğer BDM sohbet robotlarından daha iyi olduğunu gözlemledik. Okunabilirlik açısından, tüm sohbet robotların yanıtlarının okunmasının zor olduğunu, ancak Gemini ve Copilot'un nispeten daha okunabilir olduğunu gözlemledik. Sonuç olarak, BDM sohbet robotlarının yanıtları karşılaştırıldığında, Gemini'nin hem uygunluk düzeyi hem de okunabilirlik açısından en iyi olduğu, ChatGPT 3.5'in ise en kötü yanıtları verdiği görüldü. Hastalarımıza bu BDM sohbet robotlarının nadiren de olsa uygunsuz cevaplar verebileceğini hatırlatmakta yarar vardır.

### Etik

**Etik Kurul Onayı:** Gerekmez.

**Hasta Onayı:** Gerekmez.

### Beyan

### Yazarlık Katkıları

Konsept: F.O.A., S.E., Y.B.A., B.K.Y., Dizayn: F.O.A., A.C., Y.B.A., Y.Y., Veri Toplama veya İşleme: B.K.A., A.C., S.E., Analiz veya Yorumlama: A.C., B.K.Y., Y.Y., Literatür Arama: F.O.A., B.K.A., B.K.Y., Yazan: F.O.A., B.K.A., A.C., Y.B.A., S.E., B.K.Y., Y.Y.

**Çıkar Çatışması:** Yazarlar bu makale ile ilgili olarak herhangi bir çıkar çatışması bildirmemiştir.

**Finansal Destek:** Çalışmamız için hiçbir kurum ya da kişiden finansal destek alınmamıştır.

### Kaynaklar

1. Xu L, Sanders L, Li K, Chow JC. Chatbot for health care and oncology applications using artificial intelligence and machine learning: systematic review. *JMIR Cancer*. 2021;7:27850.

- Moor M, Banerjee O, Abad ZSH, Krumholz HM, Leskovec J, Topol EJ, Rajpurkar P. Foundation models for generalist medical artificial intelligence. *Nature*. 2023;616:259-265.
- De Angelis L, Baglivo F, Arzilli G, Privitera GP, Ferragina P, Tozzi AE, Rizzo C. ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. *Front Public Health*. 2023;11:1166120.
- Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *N Engl J Med*. 2023;388:1233-1239.
- Ali R, Tang OY, Connolly ID, Fridley JS, Shin JH, Zadnik Sullivan PL, Cielo D, Oyelese AA, Doberstein CE, Telfeian AE, Gokaslan ZL, Asaad WE. Performance of ChatGPT, GPT-4, and Google Bard on a neurosurgery oral boards preparation question bank. *Neurosurgery*. 2023;93:1090-1098.
- Bernstein IA, Zhang YV, Govil D, Majid I, Chang RT, Sun Y, Shue A, Chou JC, Schehlein E, Christopher KL, Groth SL, Ludwig C, Wang SY. Comparison of ophthalmologist and large language model chatbot responses to online patient eye care questions. *JAMA Netw Open*. 2023;6:2330320.
- Readability is an essential content marketing tool. *Readable*. Available from: <https://readable.com/readability/#goodscore>. Accessed April 26, 2023.
- Patel AJ, Kloosterboer A, Yannuzzi NA, Venkateswaran N, Sridhar J. Evaluation of the content, quality, and readability of patient accessible online resources regarding cataracts. *Semin Ophthalmol*. 2021;36:384-391.
- Basch CH, Mohlman J, Hillyer GC, Garcia P. Public health communication in time of crisis: readability of on-line COVID-19 information. *Disaster Med Public Health Prep*. 2020;14:635-637.
- Hedman AS. Using the SMOG formula to revise a health-related document. *Am J Health Educ*. 2008;39:61-64.
- Robinson E, McMenemy D. 'To be understood as to understand': a readability analysis of public library acceptable use policies. *J Librarians Inf Sci*. 2020;52:713-725.
- 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;2:0000198.
- Tepe M, Emekli E. Assessing the responses of large language models (ChatGPT-4, Gemini, and Microsoft Copilot) to frequently asked questions in breast imaging: a study on readability and accuracy. *Cureus*. 2024;16:59960.
- Lee TJ, Campbell DJ, Patel S, Hossain A, Radfar N, Siddiqui E, Gardin JM. Unlocking health literacy: the ultimate guide to hypertension education from ChatGPT versus Google Gemini. *Cureus*. 2024;16:59898.
- DuBay WH. The principles of readability. 2014. Available from: <https://files.eric.ed.gov/fulltext/ED490073.pdf>
- Flesch R. How To Write Plain English: A Book For La Wyers And Consumers. 2014. Available from: [https://books.google.com.tr/books/about/How\\_to\\_Write\\_Plain\\_English.html?id=-kpZAAAAMAAJ&redir\\_esc=y](https://books.google.com.tr/books/about/How_to_Write_Plain_English.html?id=-kpZAAAAMAAJ&redir_esc=y)
- Coleman M, Liao TL. A computer readability formula designed for machine scoring. *J Appl Psychol*. 1975;60:283-284.
- Momenaei B, Wakabayashi T, Shahlaee A, Durrani AF, Pandit SA, Wang K, Mansour HA, Abishek RM, Xu D, Sridhar J, Yonekawa Y, Kuriyan AE. Appropriateness and readability of ChatGPT-4-generated responses for surgical treatment of retinal diseases. *Ophthalmol Retina*. 2023;7:862-868.