ÖZGÜN ARAŞTIRMA

ChatGPT ve Gemini Modellerinin Ortodonti Alanındaki Türkçe Yeterliliği: Hasta Sorularına Verilen Yanıtların Doğruluk, Eksiksizlik ve Okunabilirlik Değerlendirilmesi

The Turkish Proficiency of ChatGPT and Gemini in Orthodontics: An Evaluation of the Accuracy, Completeness, and Readability of Responses to Patient Questions

Dr. Gizem Boztaş Demir
Sağlık Bilimleri Üniversitesi, Gülhane Diş Hekimliği Fakültesi, Ortodonti Anabilim Dalı, Ortodonti Anabilim Dalı, Ankara
ORCID ID: 0000-0002-9871-8226

Prof. Dr. Serkan Görgülü
Sağlık Bilimleri Üniversitesi, Gülhane Diş Hekimliği Fakültesi, Ortodonti Anabilim Dalı, Ortodonti Anabilim Dalı, Ankara
ORCID ID: 0000-0003-1617-573X

Geliş tarihi: 18.12.2024

Kabul tarihi: 20.02.2025

doi: 10.5505/yeditepe.2025.15010

Yazışma adresi:
Dr. Gizem BOZTAŞ DEMİR
Sağlık Bilimleri Üniversitesi, Gülhane Diş Hekimliği Fakültesi, Ortodonti Anabilim Dalı,
Ortodonti Anabilim Dalı, Ankara, Türkiye
Adres: Ankara, Türkiye
Tel: +90 538 672 18 99
Fax: +90 312 304 60 26
E-posta: dt.gizemdemir@gmail.com

ÖZET

Amaç: Bu çalışma, ChatGPT ve Gemini geniş dil modellerinin ortodontik tedaviye yönelik sıkça sorulan sorulara verdikleri yanıtların eksiksizlik, doğruluk ve okunabilirlik düzeylerini karşılaştırmayı amaçlamaktadır.

Gereç ve Yöntem: Sorular, genel, tedavi ile ilgili ve bakım ve hijyen olmak üzere üç kategoriye ayrılmıştır. ChatGPT ve Gemini modellerinden elde edilen yanıtlar, üçlü Likert ölçeği ile eksiksizlik, altılı Likert ölçeği ile doğruluk açısından değerlendirilmiş; okunabilirlik ise Türkçe'ye uyarlanmış Ateşman Okunabilirlik Formülü kullanılarak analiz edilmiştir. İstatistiksel analizler, açık kaynaklı Jamovi yazılımı (The Jamovi Project 2022, sürüm 2.3.21.0) ile gerçekleştirilmiştir.

Bulgular: ChatGPT, eksiksizlik açısından tüm sorularda (p=0,042) ve tedavi ile ilgili sorularda (p=0,037) Gemini'ye göre istatistiksel olarak anlamlı düzeyde üstün performans göstermiştir. Ancak, doğruluk açısından iki model arasında anlamlı bir fark tespit edilmemiştir. Okunabilirlik açısından ise Gemini, ChatGPT'ye kıyasla tüm sorular (p=0,001), genel kategori (p=0,013) ve bakım ve temizlik kategorisi (p=0,01) değerlendirmelerinde istatistiksel olarak anlamlı derecede daha yüksek skorlar elde etmiştir.

Sonuç: Bu çalışma, ChatGPT'nin eksiksizlik açısından üstün performans sergilediğini, ancak Gemini'nin okunabilirlik düzeyinde daha iyi sonuçlar verdiğini göstermiştir. Her iki model de doğruluk açısından yeterli performans sergilemiş olup, Türkçe dilinde hasta bilgilendirme süreçlerinde potansiyel araçlar olarak değerlendirilmektedir. Bununla birlikte, modellerin eksiksizlik ve okunabilirlik arasında bir denge sağlayacak şekilde geliştirilmesi, hasta iletişiminin etkinliğini artırmak için önemlidir.

Anahtar Kelimeler: Ortodonti, geniş dil modelleri, jeneratif yapay zeka.

ABSTRACT

Aim: This study aimed to compare the performance of ChatGPT and Gemini large language models (LLMs) in answering frequently asked questions regarding orthodontic treatment. The evaluation was based on the completeness, accuracy, and readability of the responses in Turkish.

Materials and Method: Frequently asked questions related to orthodontic treatment were categorized into general, treatment-related, and care and hygiene groups. Responses from ChatGPT and Gemini models were assessed for completeness using a three-point Likert scale, accuracy using a six-point Likert scale, and readability using the Turkish-adapted Ateşman Readability Formula. Statistical analyses were conducted using open-source Jamovi software (The Jamovi

Project 2022, version 2.3.21.0, www.jamovi.org).

Results: ChatGPT demonstrated statistically significant superiority over Gemini in completeness across all questions (p=0,042) and treatment-related questions (p=0,037). However, no statistically significant difference was observed between the two models in terms of accuracy. In terms of readability, Gemini achieved significantly higher scores compared to ChatGPT across all questions (p=0,001), the general category (p=0,013), and the care and hygiene category (p=0,01), indicating responses that were easier to understand.

Conclusion: This study revealed that ChatGPT outperformed Gemini in terms of completeness, particularly for all questions and treatment-related questions, while both models performed similarly in accuracy. On the other hand, Gemini provided responses with higher readability, making them more accessible for patients. Both models hold promise as patient information tools in orthodontics, but achieving a balance between completeness and readability remains essential for enhancing their effectiveness.

Keywords: Orthodontics, large language models, generative artificial intelligence.

GİRİŞ

Son yıllarda, yapay zekanın bir dalı olan Doğal Dil İşleme (Natural Language Processing- NLP), insan dilini anlamaya ve bilgisayarların bu dille etkileşim kurmasına olanak tanıyan önemli bir alan olarak öne çıkmıştır. NLP sayesinde bilgi çıkarımı, dil oluşturma ve çeviri gibi görevler, dilin yapısal ve anlamsal özellikleri analiz edilerek gerçekleştirilebilmektedir. NLP'nin son yıllardaki en dikkat çekici ilerlemelerinden biri, Büyük Dil Modelleri'nin (Large Language Models-LLM) geliştirilmesidir.[1] Gelişmiş sinir ağı mimarileri ile oluşturulan bu modeller, bağlam, nüanslar ve semantik ilişkileri başarılı bir şekilde çözümleyerek etkili ve anlaşılır çözümler sunabilmektedir. LLM'lerin akıcı ve anlamlı metinler oluşturabilme, diyaloglar üzerinden soruları yanıtlayabilme, diller arasında çeviri yapabilme ve dil temelli birçok görevi yerine getirebilme kapasitesine sahip olduğu bildirilmiştir. Bu modeller dil işleme görevlerini kolaylaştırma yetenekleri ile sağlık bilimlerinde de yenilikçi uygulamalara olanak sağlamıştır.[2]

Hasta ve hasta yakınları, tedavi süreçleriyle ilgili pek çok soru ve belirsizlikle karşılaşmakta, bu nedenle sıkça bilgi arayışı içine girmektedirler. Bu sorular için her durumda hekime ulaşmak mümkün olamayabilir.[3] LLM'ler, geniş veri setleri üzerinde eğitilmiş olup, verileri tokenize ederek dilin yapısal ve anlamsal özelliklerini analiz eder ve sorulara bağlamsal olarak tutarlı, hızlı ve akıcı yanıtlar üretebilmektedir. Bu mekanizma, geniş dil modellerinin kullanıcıların sorularına anında ve hedefe yönelik cevaplar sunmasını sağlamaktadır.[4] Bu nedenle, hasta ve hasta yakınları tarafından kolay erişilebilir bir bilgi kaynağı olarak sıklıkla tercih edilmektedir. Ortodonti, uzun süreli, aşamalı ve çeşitli tedavi seçeneklerini içeren bir alandır. Tedavi sürecince hasta ve hasta yakınlarının zihninde pek çok soru işareti oluşturmaktadır. Ancak doğrudan hekime ulaşmanın her zaman mümkün olmaması, hasta ve hasta yakınlarını geniş dil modellerine yönlendirebilir.[3,5]

ChatGPT, Gemini, Grok, BERT, CLAUDE vb. gibi geniş dil modelleri, sağlık bilimlerinde hasta sorularının yanıtlama yetenekleri açısından literatürde çeşitli çalışmalar tarafından değerlendirilmiştir.[2,3,5-9] Bu çalışmalarda modellerin hasta sorularını yanıtlama yetenekleri İngilizce dilinde değerlendirilmiştir. Ancak ülkemizde, hastalar ve hasta yakınları tarafından bu sorular modellere genellikle Türkçe olarak yöneltilmektedir. Geniş dil modelleri tokenizasyon mekanizması ile çalışarak dilin yapısal ve anlamsal özelliklerini analiz eder; yani doğrudan çeviri yapmazlar. Ancak ana eğitim dillerinin İngilizce olması, Türkçe yanıtların doğruluğu ve tutarlılığı açısından farklılıklara yol açabilir.[1] Dolayısıyla, modellerin Türkçe sorulara verdiği yanıtların yeterliliğinin değerlendirilmesi önemlidir. Bu nedenle yapılan çalışmaların Türkçe dilinde genişletilmesi, yerel kullanıcıların bilgi erişim süreçlerinin iyileştirilmesi açısından büyük önem taşımaktadır. Geniş dil modellerinin, hasta sorularına Türkçe olarak doğru, eksiksiz yanıtlar vermesi ve bu yanıtların okunabilirliği büyük önem taşımaktadır. Özellikle hasta ve hasta yakınlarının bilgiye erişim amacıyla yönelttiği sorularda, verilen yanıtların fazla teknik veya karmaşık bir dil içermesi, cevapların anlaşılmasını zorlaştırabilir. Bu nedenle, sunulan bilgilerin hastaların sağlık okuryazarlığı seviyesine uygun, sade ve anlaşılır bir dilde olması gerekmektedir.[3,7,9]

Bu çalışmanın birincil amacı, hastaların ortodontik tedavi süreçleriyle ilgili sorabilecekleri Türkçe dilindeki sorulara sıklıkla kullanılan iki geniş dil modeli tarafından verilen yanıtların eksiksizlik, doğruluk ve okunabilirlik düzeyi açısından değerlendirilmesidir. İkincil amaç, ChatGPT ve Gemini modellerinin ortodontik tedavi süreçleriyle ilgili hasta sorularını yanıtlama konusundaki yeterliliklerini ve üretilen yanıtların okunabilirlik düzeylerini karşılaştırmalı olarak değerlendirmektir. Çalışmanın sıfır hipotezi, ChatGPT ve Gemini modelleri arasında hastaların ortodontik tedavi süreçlerinde sorabilecekleri sorulara verilen yanıtların eksiksizlik, doğruluk ve okunabilirlik düzeyi açısından anlamlı bir fark olmadığıdır.

GEREÇ ve YÖNTEM

Çalışmanın akış şeması Şekil 1'de gösterilmiştir.

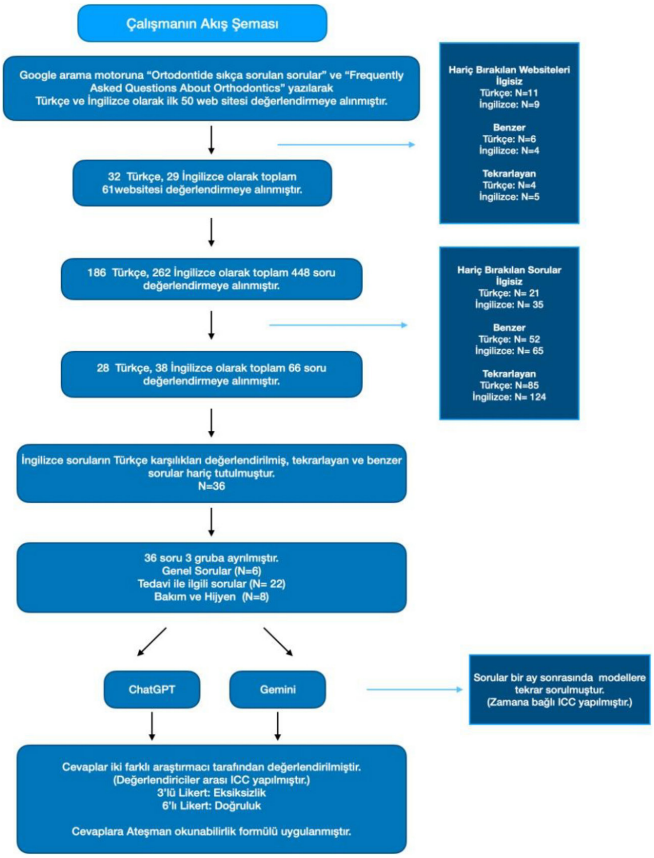
1. Etik Değerlendirme

Bu çalışma, protokolde belirtildiği üzere insan katılımcıla-

rını içermediği ve onların örneklerini kullanmadığı için etik onay gerektirmemektedir.

2. Soruların Hazırlanması

Çalışmada, "ortodonti hakkında sık sorulan sorular" arama terimi kullanılarak Google Arama Motoru üzerinden hem İngilizce hem de Türkçe web siteleri taranmıştır.[3] Türkçe ve İngilizce olarak ilk 50 web sitesi değerlendirmeye alınmıştır. İlgisiz, tekrarlayan ve benzer olan içerikler çalışma dışı bırakılmıştır. Bu kriterler uygulandıktan sonra, 32 İngilizce ve 29 Türkçe olmak üzere toplam 61 uygun web sitesi çalışmaya dahil edildi. Değerlendirme sürecinde, web sitelerinden elde edilen 262 İngilizce soru ve 188 Türkçe soru, "tekrarlayan," "benzer," ve "ilgili olmayan" içerik kriterlerine göre elenmiştir. Bu süreç sonucunda 38 İngilizce ve 28 Türkçe soru uygun bulundu. Elde edilen İngilizce soruların Türkçe karşılıkları ve Türkçe sorular birlikte tekrar gözden geçirilerek tekrarlayan ve benzer olanlar yeniden elenmiştir ve nihai olarak 34 soru belirlenmiştir (Tablo 1). Bu 34 soru, içeriklerine göre genel, tedavi ile ilgili ve bakım ve hijyen ile ilgili olmak üzere üç ana başlık altında gruplandırılmıştır (Şekil 1). Çalışmanın bu süreci, ortodonti alanında 25 yıllık deneyime sahip bir araştırmacı tarafından gerçekleştirilmiştir.



Şekil 1. Çalışmanın akış şeması.

Tablo 1. Çalışmada kullanılan sorular ve soruların gruplandırılması.

Genel Sorular
Ortodonti nedir?
Ortodontist kimdir ve ne iş yapar?
Neden bir ortodonti uzmanını tercih etmelisiniz?
Diş teli tedavisine neden ihtiyaç duyarız?
Diş çapraşıklığı neden olur?
Ortodontik tedavi ücreti ne kadardır?
Tedavi ile İlgili Sorular
Diş teli tedavisi için bir yaş sınırı var mı? Yetişkinler de tedavi olabilir mi?
Çocuklar için ortodontik tedaviye ne zaman başlanmalı?
Ortodontik tedavi sırasında herhangi bir risk veya komplikasyon var mı?
Ortodontik tedavi yüz görünümünü değiştirir mi?
Ortodontik tedavi başlamadan önce tüm süt dişlerinin kaybedilmesi gerekir mi?
Dişlerimin düzelmesi için ortodonti dışında alternatif tedaviler var mı?
Estetik braketler var mı? Bu braketler metal braketler kadar etkili mi?
Lingual ortodonti nedir? Geleneksel diş tellerinden farkı nedir?
Şeffaf plaklar ve diş telleri arasındaki farklar nelerdir?
Şeffaf plaklar ve diş teli karşılaştırıldığında hangisi daha hızlı sonuç verir?
Diş telleri dişleri nasıl hareket ettirir?
Şeffaf plaklar dişleri nasıl hareket ettirir?
Ortodontik tedavi ağrıya neden olur mu?
Ortodontik tedavi ne kadar sürer?
Diş telleri varken spor yapabilir miyim?
Diş teli takmak konuşmamı etkiler mi?
Diş telleri takıldıktan sonra yemek yerken zorlanır mıyım?
Diş telleri ile birlikte MRI çektirilebilir mi?
Pekiştirme nedir ve ne kadar süre gereklidir?
Diş telleri çıkarıldıktan sonra dişlerim tekrar hareket eder mi?
Ortodontik tedavide başarı oranı nedir?
Ağız Hijyeni ve Bakım
Ortodontik tedavi sırasında nelere dikkat etmeliyim?
Braketler neden düşer?
Diş tellerinin temizliği nasıl yapılmalı?
Şeffaf plakların temizliği nasıl yapılmalı?
Ortodontik tedavi sırasında kullanılması gereken en iyi diş fırçası hangisidir?
Ortodontik tedavi sırasında diş fırçalama nasıl olmalıdır?
Ortodontik tedavi sırasında ara yüz fırçası nasıl kullanılır?
Ortodontik tedavi sonrasında diş bakımımı nasıl yapmalıyım?

a. Modellerin Yanıtlarının Değerlendirilmesi

Çalışmamızda her iki modelin de ücretsiz sürümleri kullanılmıştır. Bu seçim, hastaların ve hasta yakınlarının genellikle ücretsiz sürümlere erişim sağlaması ve bu sürümleri kullanma olasılığının yüksek olması düşüncesiyle yapılmıştır. Bu şekilde modellerin değerlendirilmesinde genel kullanıcı deneyimini daha doğru bir şekilde yansıtılması amaçlanmıştır. Her sorunun değerlendirilmesinde, ChatGPT ve Gemini modellerinin yanıtlarının tutarlılığı ve doğruluğunu sağlamak amacıyla sorular yeni bir sohbet oturumunda yazılmıştır. Her bir soruya verilen yanıt, modelin önceki sorular ve yanıtlar nedeniyle etkilenmeden, bağımsız olarak oluşturulmuştur. Bu yöntem ile modellerin soruları yanıtlamadaki performansının daha tarafsız ve

güvenilir bir şekilde değerlendirilmesi amaçlanmıştır. Geniş dil modelleri, sürekli olarak geliştirildiğinden zamanla yanıtlarında zamana bağlı değişimler gözlenebilmektedir. Bu nedenle, sorular 1 ay sonra aynı modellerde tekrar sorulmuş ve yanıtların tutarlılığı Intraclass Correlation Coefficient (ICC) analizi ile değerlendirilmiştir.

i. Doğruluk ve Eksiksizlik Değerlendirmesi

Çalışmada, ChatGPT ve Gemini modellerinin verdiği yanıtlar doğruluk ve eksiksizlik açısından değerlendirilmiştir. Değerlendirme süreci, ortodonti alanında uzman iki araştırmacı tarafından bağımsız olarak gerçekleştirilmiş ve iki değerlendiricinin puanlarının ortalaması alınmıştır.

Yanıtların doğruluğu, altı noktalı Likert ölçeği kullanılarak değerlendirilmiştir:10

1- tamamen yanlış,
2- doğruya göre daha fazla yanlış,
3- doğru ve yanlış eşit derecede,
4- doğruya daha yakın,
5- neredeyse tamamen doğru,
6- tamamen doğru.

Yanıtların eksiksizliği, üç noktalı Likert ölçeği ile değerlendirilmiştir:

1- eksik: sorunun bazı yönlerine değinmekte ancak önemli kısımlar yetersiz veya eksik,
2- yeterli: sorunun tüm yönlerine değinmekte ve gerekli minimum bilgiyi sağlamaktadır,
3- kapsamlı: sorunun tüm yönlerine değinmekte ve beklenenden daha fazla bilgi veya bağlam sunmaktadır.

Doğruluk ve bütünlük değerlendirmelerinin güvenirliğini sağlamak amacıyla iki araştırmacının farklı zamanlardaki ölçümleri (test-tekrar güvenirliği) ve iki araştırmacı arasındaki ölçümler için Intraclass Correlation Coefficient (ICC) analizi uygulanmıştır.

ii.Okunabilirlik Değerlendirmesi

Çalışmada, modeller tarafından verilen yanıtların okunabilirlik düzeyi, Flesch-Kincaid yönteminden Türkçeye uyarlanmış olan Ateşman Okunabilirlik Formülü kullanılarak değerlendirilmiştir. Ateşman skoru, Türkçe metinlerin okunabilirliğini ölçmek amacıyla geliştirilen bir araçtır ve cümle uzunluğu ile sözcüklerin hece sayısına dayalı bir hesaplama yöntemi kullanmaktadır.[11]

Ateşman Okunabilirlik Formülü: 198.825 - 40.175 × (Toplam Kelime Sayısı / Toplam Cümle Sayısı) - 2.610 × (Toplam Hece Sayısı / Toplam Kelime Sayısı)

Değerlendirme sürecinde, modellerden elde edilen yanıtlar Microsoft Word ve Microsoft Excel kullanılarak analiz edilmiştir. Metinler, Microsoft Word yazılımının "Yazım ve Dilbilgisi" sekmesi aracılığıyla değerlendirilmiştir. Ateşman skorları Microsoft Excel yazlımı kullanılarak hesaplanmıştır. Ateşman skoru metnin okunabilirlik düzeyini belirlemek için kullanılmakta olup, farklı eğitim seviyelerine göre anlaşılırlık düzeyini sınıflandırmaktadır.[11] (Tablo 2)

Tablo 2. Ateşman Türkçe Okunabilirlik İndeksi.

İndeks	Okunabilirlik Düzeyi
90-100	4. sınıf ve altı öğrenciler tarafından kolayca anlaşılır.
80-89	5. veya 6. sınıf öğrencileri tarafından kolayca anlaşılır.
70-79	7. veya 8. sınıf öğrencileri tarafından kolayca anlaşılır.
60-69	9. veya 10. sınıf öğrencileri tarafından kolayca anlaşılır.
50-59	11. veya 12. sınıf öğrencileri tarafından kolayca anlaşılır.
40-49	Ön lisans (13. veya 14. sınıf) öğrencileri tarafından kolayca anlaşılır.
30-39	Lisans mezunları tarafından kolayca anlaşılır.
29 ve altı	Lisansüstü mezunları tarafından kolayca anlaşılır.

b. İstatistiksel Analiz

Doğruluk ve bütünlük değerlendirmelerinin güvenirliğini sağlamak amacıyla Intraclass Correlation Coefficient (ICC) analizi uygulanmıştır. Analiz kapsamında, hem aynı iki araştırmacının farklı zamanlardaki ölçümleri (test-tekrar güvenirliği) hem de iki araştırmacı arasındaki ölçümler değerlendirilmiştir. Gruplarda normal dağılım varsayımı Shapiro-Wilk testi kullanılarak değerlendirilmiştir. Normal dağılım gösteren verilerde grup karşılaştırmaları için Fisher's One-Way ANOVA testi, ikili karşılaştırmalar için ise Tukey testi uygulanmıştır. Normal dağılım göstermeyen verilerde ise grup karşılaştırmaları için Kruskal-Wallis testi, ikili karşılaştırmalar için Dwass, Steel, Critchlow-Fligner (DSCF) testi kullanılmıştır. Tüm istatistiksel analizler, açık kaynaklı Jamovi yazılımı (The Jamovi Project 2022, sürüm 2.3.21.0, www.jamovi.org) kullanılarak gerçekleştirilmiştir.

BULGULAR

Yapılan analizde, birinci araştırmacının zamana bağlı ICC değeri 0,91, ikinci araştırmacının zamana bağlı ICC değeri 0,89 olarak bulunmuştur. İki araştırmacı arasındaki ölçümlerin tutarlılığına ilişkin ICC değeri 0,93 olarak hesaplanmıştır. Bu bulgular, yapılan değerlendirmelerin yüksek düzeyde tutarlı ve güvenilir olduğunu göstermektedir.

Tablo 3'te, ChatGPT ve Gemini modelleri tarafından verilen yanıtların eksiksizlik ve doğruluk skorları karşılaştırılmıştır. Modellerin eksiksizlik açısından karşılaştırmalarında, tüm sorularda (p=0,042) ve tedavi ile ilgili sorularda (p=0,037), ChatGPT, Gemini'ye göre istatistiksel olarak anlamlı düzeyde üstün performans sergilemiştir. Ancak, doğruluk açısından yapılan değerlendirmelerde iki model arasında istatistiksel olarak anlamlı bir fark tespit edilmemiştir.

Tablo 3. ChatGPT ve Gemini geniş dil modellerinin 3'lü likert ve 6'lı likert ölçekleri ile eksiksizlik ve doğruluk açısından karşılaştırılmalı değerlendirmesi.

GRUP	TEST	MODEL	N	Ort.±SS	Medyan	Min-Maks	%25-%75	p
TÜM SORULAR	BÜTÜNLÜK	GPT	36	2,47±0,476	2,5	2-3	2-3	0,044*
		GEMINI	36	2,22±0,525	2	1-3	2-2,5	
	DOĞRULUK	GPT	36	5,53±0,476	5,5	5-6	5-6	0,904
		GEMINI	36	5,54±0,467	5,5	5-6	5-6	
GENEL SORULAR	BÜTÜNLÜK	GPT	6	2,25±0,418	2	2-3	2-2,38	0,235
		GEMINI	6	2,58±0,492	2,75	2-3	2,13-3	
	DOĞRULUK	GPT	6	5,25±0,418	5	5-6	5-5,38	1,000
		GEMINI	6	5,25±0,274	5,25	5-5,5	5-5,5	
TEDAVİ İLE İLGİLİ SORULAR	BÜTÜNLÜK	GPT	22	2,52±0,475	2,5	2-3	2-3	0,037*
		GEMINI	22	2,2±0,504	2	1-3	2-2,38	
	DOĞRULUK	GPT	22	5,55±0,486	5,75	5-6	5-6	0,879
		GEMINI	22	5,57±0,495	6	5-6	5-6	
BAKIM VE TEMİZLİK İLE İLGİLİ SORULAR	BÜTÜNLÜK	GPT	8	2,67±0,516	3	2-3	2,25-3	0,12
		GEMINI	8	2,08±0,665	2	1-3	2-2,38	
	DOĞRULUK	GPT	8	5,75±0,418	6	5-6	5,63-6	1,000
		GEMINI	8	5,75±0,418	6	5-6	5,63-6	

N: Örnek sayısı, **Ort.:**Ortalama, **SS:** Standart sapma, **Min.:** En düşük değer, **Max.:** En yüksek değer **%25:** Birinci çeyreklik değeri, **%75:** Üçüncü çeyreklik değeri **p < 0,05** olduğunda sonuçlar istatistiksel olarak anlamlıdır. Grup karşılaştırmaları için Kruskal-Wallis testi kullanılmıştır.

Tablo 4'te, her bir model içerisinde soru gruplarının (genel, tedavi ile ilgili ve bakım-hijyen) eksiksizlik ve doğruluk skorları açısından karşılaştırılması yapılmıştır. Yapılan analizler sonucunda, soru grupları arasında istatistiksel olarak anlamlı bir fark tespit edilmemiştir.

Tablo 4. Soru gruplarının 3'lü ve 6'lı likert ölçekleri puanlarına göre karşılaştırılması.

TEST	GRUP	N	Ort.±SS	Medyan	Min-Maks	%25-%75	p
GPT BÜTÜNLÜK	GENEL	6	2,25±0,418	2	2-3	2-2,38	
	TEDAVİ	22	2,32±0,451	2	2-3	2-2,88	0,326
	BAKIM VE TEMİZLİK	8	2,67±0,516	3	2-3	2,25-3	
GPT DOĞRULUK	GENEL	6	5,25±0,418	5	5-6	5-5,38	
	TEDAVİ	22	5,55±0,486	5,75	5-6	5-6	0,177
	BAKIM VE TEMİZLİK	6	5,75±0,418	6	5-6	5,63-6	
GEMINI BÜTÜNLÜK	GENEL	6	2,58±0,492	2,75	2-3	2,13-3	
	TEDAVİ	22	2,2±0,504	2	1-3	2-2,38	0,267
	BAKIM VE TEMİZLİK	8	2,08±0,665	2	1-3	2-2,38	
GEMINI DOĞRULUK	GENEL	6	5,25±0,274	5,25	5-5,5	5-5,5	
	TEDAVİ	22	5,57±0,495	6	5-6	5-6	0,064
	BAKIM VE TEMİZLİK	8	5,75±0,418	6	5-6	5,63-6	

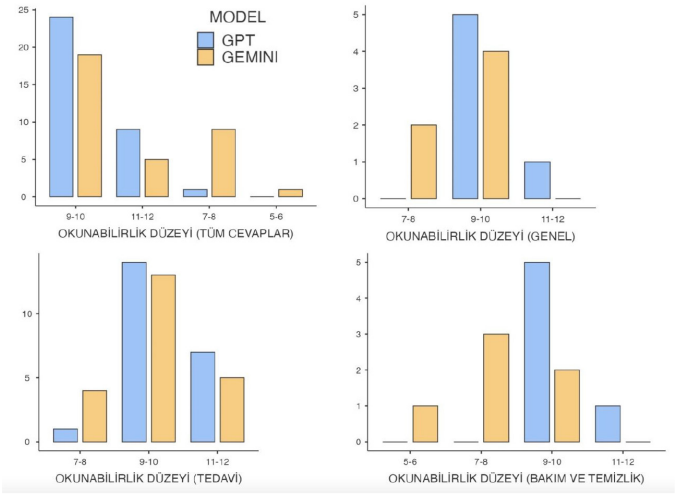
N: Örnek sayısı, **Ort.:**Ortalama, **SS:** Standart sapma, **Min:** En düşük değer, **Max.:** En yüksek değer **%25:** Birinci çeyreklik değeri, **%75:** Üçüncü çeyreklik değeri, **p<0,05** olduğunda sonuçlar istatistiksel olarak anlamlıdır. Grup karşılaştırmaları için Kruskal-Wallis testi, ikili karşılaştırmalar için Dwass, Steel, Critchlow-Fligner (DSCF) testi kullanılmıştır. Farklı harflere sahip gruplar birbirinden önemli ölçüde farklıdır.

Tablo 5'te, ChatGPT ve Gemini modellerinin okunabilirlik skorları karşılaştırılmıştır (Şekil 2). Yapılan analizlerde, tüm sorular genelinde (p=0,001), genel kategoride (p=0,013) ve bakım ve temizlik kategorisinde (p=0,01), Gemini'nin okunabilirlik skorları, ChatGPT'ye kıyasla istatistiksel olarak anlamlı derecede yüksek bulunmuştur (Şekil 3).

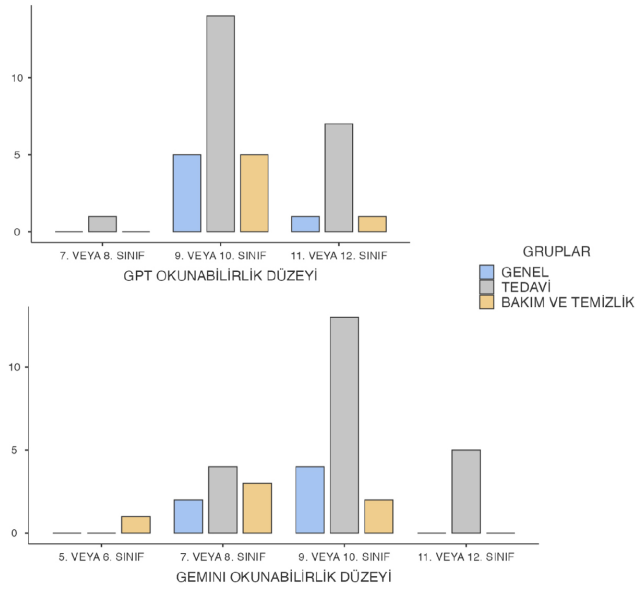
Tablo 5. Ateşman Türkçe okunabilirlik indeksine göre geniş dil modellerinin karşılaştırılması.

GRUP	MODEL	N	Ort.±SS	Medyan	Min-Maks	%25 -%75	p
TÜM SORULAR	GPT	34	63,1±4,92	63,9	51,6-71,6	58,6-67	0,001*
	GEMINI	34	67,7±5,59	68,3	56-84	65,2-70,6	
GENEL	GPT	6	63,3±3,67	63	58,4-69,2	61,5-64,7	0,013*
	GEMINI	6	69,5±1,36	69,3	67,9-71,7	68,6-70	
TEDAVİ	GPT	22	62,8±5,48	64,3	51,6-71,6	57,9-67,1	0,072
	GEMİNİ	22	65,7±5,14	66,3	56-74,3	62,8-69,3	
BAKIM VE TEMİZLİK	GPT	8	63,8±4,4	64,3	56,8-69,4	62-66,2	0,01*
	GEMINI	8	73±6,2	72	66,5-84	69,2-74,5	

N: Örnek sayısı, **Ort.:**Ortalama, **SS:** Standart sapma, **Min.:** En düşük değer, **Max.:** En yüksek değer **%25:** Birinci çeyreklik değeri, **%75:** Üçüncü çeyreklik değeri **p< 0,05** olduğunda sonuçlar istatistiksel olarak anlamlıdır. Grup karşılaştırmaları için Kruskal-Wallis testi kullanılmıştır.



Şekil 2. Soru gruplarının ve modellerin Ateşman okunabilirlik düzeyleri.



Şekil 3. Modellerin okunabilirlik düzeyinin karşılaştırılması.

TARTIŞMA

Bu çalışmada, ortodonti tedavi süreçlerinde hasta ve hasta yakınlarının sorabilecekleri sorulara ChatGPT ve Gemini geniş dil modellerinin verdikleri yanıtlar doğruluk, eksiksizlik ve okunabilirlik düzeyi açısından karşılaştırmalı olarak değerlendirilmiştir. Araştırmalarımıza göre bu çalışma, sık kullanılan ChatGPT ve Gemini modellerini Türkçe dilinde ve ortodonti alanında doğruluk, eksiksizlik ve okunabilirlik düzeyi açısından karşılaştırmalı olarak inceleyen ilk çalışma olma özelliğini taşımaktadır. Bu yönüyle literatürdeki mevcut boşluğu doldurmakta ve geniş dil modellerinin Türkçe dilindeki performansına ilişkin önemli bulgular sunmaktadır. Çalışma bulgularımıza göre, her iki modelin de doğru ve eksiksiz yanıtlar verme konusunda yeterli olduğu söylenebilir. Ancak yanıtların okunabilirlik düzeyleri, 51-69 arasında değişmekte olup, bu skorlar orta düzeyde okunabilirlik seviyesine işaret etmektedir. Bu değerlere göre, yanıtların okunabilirlik düzeyi 50-60 arasında 11.-12. sınıf, 60-70 arasında ise 9.-10. sınıf öğrencileri tarafından kolayca anlaşılabilecek seviyededir. Bu bulgular, ortodonti alanında sağlık iletişiminde hasta ve

hasta yakınlarının bilgiye erişimini iyileştirebilmek için modeller tarafından üretilen yanıtların okunabilirlik düzeylerinin sadeleştirilmesi gerektiğini ortaya koymaktadır.

Geniş dil modellerinin (LLM) performans değerlendirmelerinde Likert ölçeklerinin kullanımı, literatürde yaygın bir uygulama olup özellikle içerik analizi çalışmalarında sıkça tercih edilmektedir.[2,4,6,12,13] Johnson ve ark.[2] çalışmasında, ChatGPT tarafından verilen tıbbi sorulara yönelik yanıtlar 6'lı Likert ölçeği kullanılarak doğruluk açısından, 3'lü Likert ölçeği kullanılarak ise eksiksizlik açısından değerlendirilmiştir. Çalışmanın bulguları, modelin genellikle yüksek doğrulukta ve eksiksiz yanıtlar verdiğini ortaya koymaktadır ancak bazı sınırlamaların devam ettiği belirtilmiştir. Demir ve ark.[4], geniş dil modellerinin PICO tabanlı sorguları oluşturmadaki performansını değerlendirmek için ChatGPT 3.5 ve 4 modellerini kullanmış, doğruluk ve eksiksizlik bağımsız araştırmacılar tarafından 3'lü ve 6'lı Likert ölçekleri ile analiz etmişlerdir. Dursun ve ark.[12] çalışmasında ise ChatGPT, Gemini ve Copilot modellerinin, şeffaf plak tedavisi hakkında sık sorulan sorulara verdikleri yanıtlar değerlendirilmiştir. Bu çalışmada doğruluk 5'li Likert ölçeği ile ölçülmüştür. Literatürdeki bu çalışmalar, geniş dil modellerinin sağlık alanında doğruluk ve eksiksizlik kriterleri açısından değerlendirilmesinde Likert ölçeklerinin etkin bir yöntem olduğunu ortaya koymaktadır. Bizim çalışmamızda da bu yöntem takip edilerek, ChatGPT ve Gemini modellerinin ortodonti alanında Türkçe dilindeki performansı karşılaştırmalı olarak analiz edilmiştir.

Okunabilirlik, bir metnin hedef kitle tarafından ne derece kolay anlaşıldığını belirleyen önemli bir ölçüttür ve sağlık iletişiminde hasta ve hasta yakınlarının bilgilendirilmesi açısından kritik bir role sahiptir. Literatürde okunabilirlik değerlendirmelerinde sıklıkla Flesch Reading Ease ve Flesch-Kincaid Grade Level gibi yöntemler kullanılmaktadır. Flesch-Kincaid yöntemi, bir metnin cümle uzunluğu ve sözcüklerin hece sayısına dayalı olarak hesaplanır ve okuyucunun metni anlamak için ihtiyaç duyduğu eğitim seviyesini ortaya koyar.[14] Türkçe dilinde okunabilirlik değerlendirmesi için, Flesch-Kincaid yönteminden uyarlanmış olan Ateşman Okunabilirlik Formülü kullanılmaktadır. Ateşman skoru, cümle uzunluğu ve kelimelerin hece sayısına dayalı olarak hesaplanır. Bu yöntem, Türkçe metinlerin hedef kitleye uygunluk düzeyini belirlemek için geliştirilmiş güvenilir bir araçtır.[11] Ortodonti ve diş hekimliği literatüründe, geniş dil modellerinin yanıtlarının okunabilirlik düzeyini değerlendirmek amacıyla Flesch-Kincaid skoru, İngilizce dilindeki çalışmalar için yaygın olarak kullanılmaktadır.[3,6,7] Türkçe literatürü incelediğimizde, sağlık bilimlerinde içerik okunabilirliğinin değerlendirilmesinde, Ateşman Okunabilirlik Skorunun sıklıkla kullanıldığı görülmektedir.[15–18] Çalışmamızda, modeller tarafından verilen Türkçe yanıtların okunabilirlik düzeyini değerlendirmek için Ateşman Okunabilirlik Skoru kullanılmıştır. Bu yöntemle, her iki modelin ürettiği yanıtların hasta ve hasta yakınları tarafından ne derece anlaşılabilir olduğu objektif olarak değerlendirilmesi amaçlanmıştır.

Çalışmamızda, ChatGPT'nin eksiksizlik açısından Gemini'ye kıyasla tüm sorularda ve özellikle tedavi ile ilgili sorularda daha iyi performans sergilediği görülmüştür. ChatGPT, verdiği yanıtlarda soruların gerektirdiği bilgiyi sunmuştur ve konu ile bağlantılı önemli ek bilgilendirmeler de sağlamıştır. Özellikle tedavi ile ilgili sorularda, ChatGPT'nin yanıtlarının daha yüksek ayrıntı düzeyine sahip olduğu gözlemlenmiştir. ChatGPT'nin eksiksizlik açısından Gemini'ye kıyasla daha yüksek performans göstermesi, verdiği yanıtların yalnızca istenilen bilgiyi sunmakla kalmayıp, ek bağlam ve detaylı açıklamalar içermesinden kaynaklanmaktadır. Gemini modeli de sorulara istenilen düzeyde doğru ve yeterli yanıtlar vermiştir ancak ek bilgilendirme ve ayrıntı düzeyi açısından ChatGPT'nin gerisinde kalmaktadır. ChatGPT'nin yanıtlarının eksiksizlik açısından daha detaylı olmasının, modelin eğitim veri setindeki Türkçe bilgi kaynakları ve kullanılan derin öğrenme teknikleri ile ilişkili olabilir. ChatGPT ile eksiksizlik açısından daha iyi sonuçlar elde edilse de doğruluk ve eksiksizlik açısından her iki modelin de Türkçe dilinde yeterli olduğu görülmüştür. Dursun ve ark.[12] çalışmasında, geniş dil modellerinin doğruluk, güvenilirlik, kalite ve okunabilirlik açısından performansı değerlendirilmiştir. Çalışmada ChatGPT-4, diğer modellere kıyasla daha yüksek doğruluk skorları gösterse de bu farkın istatistiksel olarak anlamlı olmadığı belirtilmiştir. Bizim çalışmamızda da modellerin genel olarak doğruluk açısından birbirine benzer ve başarılı sonuçlar verdiği görülmüştür. Meyer ve ark.[6] çalışmasında, ChatGPT, Gemini ve Claude modellerinin rhinoplasti ile ilgili sorulara verdikleri yanıtlar doğruluk, kalite, eksiksizlik ve okunabilirlik açısından değerlendirilmiştir. Çalışmada ChatGPT, doğruluk ve genel kalite açısından daha yüksek skorlar alırken, eksiksizlik açısından Gemini ve Claude'a göre daha düşük bulunmuştur. Meyer ve ark. çalışması ile bizim çalışmamızın sonuçları arasındaki farklılıkların, modellerin veri setlerinin içeriği ve eğitildiği bilgilere erişim düzeyiyle ilgili olabileceği düşünülmektedir. Sonuçlar arasındaki bu farklılık geniş dil modellerinin konuya özgü performansının değişkenlik gösterebileceğini ortaya koymaktadır.

Çalışmamızda, ChatGPT ve Gemini modellerinin verdiği yanıtların okunabilirlik düzeyleri karşılaştırılmış ve Gemini'nin skorlarının daha yüksek olduğu belirlenmiştir. Bu durum, Gemini'nin daha sade, kısa cümlelerle oluşturduğu yanıtlardan kaynaklanmış olabilir. Buna karşılık, ChatGPT'nin eksiksizlik açısından daha yüksek performans göstermesi ve yanıtlara ek bilgiler ile detaylı açıklamalar eklemesi, cümle yapılarını karmaşıklaştırarak okunabilirliği olumsuz yönde etkilemiş olabilir. Benzer şekilde, Dursun ve ark.[12] çalışmasında da Gemini'nin İngilizce dilinde

okunabilirlik açısından diğer geniş dil modellerine göre daha iyi performans sergilediği rapor edilmiştir. Çalışmamızda, her iki modelin de tedavi ile ilgili sorularda okunabilirlik düzeyinin diğer kategorilere kıyasla daha düşük olduğu belirlenmiştir. Modellerin tedavi ile ilgili sorulara yanıt verirken tedavi süreçlerine ilişkin detaylı açıklamalar yapması, tıbbi terminoloji ve uzun cümle yapıları kullanması, yanıtların okunabilirliğini olumsuz yönde etkilemiş olabileceğini düşünüyoruz. Çalışmamızın bulguları doğrultusunda, modellerin tedaviyle ilgili konularda bilgi aktarımının, hasta ve hasta yakınlarının sağlık okuryazarlığı seviyesine uygun olarak sadeleştirilmesi gerektiği çıkarımı yapılabilir. Kılınç ve ark.[19] çalışmasında, ortodonti alanında sıkça sorulan sorular genel ve tedavi ile ilgili olarak kategorize edilmiş ve sorulara verilen yanıtların okunabilirliği değerlendirilmiştir. Bu çalışmada, genel sorulara modelin verdiği yanıtların okunabilirliği, tedavi ile ilgili sorulara verdiği yanıtlara kıyasla daha düşük bulunmuştur. Buna karşılık, bizim çalışmamızda ise tedavi ile ilgili soruların okunabilirlik düzeyi daha düşük olarak tespit edilmiştir. Bu farklılık, soru setlerinin içeriğindeki farklılıklardan ve dilin yapısal özelliklerinden kaynaklanabilir. Çalışmamızda Türkçe dilinde özellikle teknik ve tedaviye yönelik açıklamalar yaparken geniş dil modellerinin daha karmaşık cümle yapıları kullanma eğiliminde olduğu gözlemlenmiştir. Bu durum, modellerin tokenizasyon temelli çalışma prensibiyle ilişkili olabilir. Türkçe, yapısal olarak eklemeli bir dil olduğundan teknik ifadeler açıklanırken daha uzun ve karmaşık cümleler oluşabilmektedir. Türkçede kelime köklerine eklenen çekim ekleri ve yapım ekleri, kelimenin hece sayısını artırır.[20] Kelimelerin ve cümlelerin uzunluğu metnin okunabilirlik skorunu doğrudan etkiler.[11] Bu bulgular, geniş dil modellerinin dilin yapısal özelliklerine bağlı olarak farklı konularda değişken performans gösterebileceğini ve okunabilirlik değerlendirmelerinde dil özelinde dikkatli yorumlamaların yapılması gerektiğini göstermektedir.

SONUÇ

Bu çalışmada, ChatGPT ve Gemini geniş dil modellerinin, hastaların ortodontik tedavi süreçleriyle ilgili sorulara verdikleri yanıtlar doğruluk, eksiksizlik ve okunabilirlik açısından değerlendirilmiştir. Bulgular, her iki modelin de hasta sorularını yanıtlama konusunda genel olarak yeterli performans gösterdiğini ortaya koymuştur. ChatGPT, eksiksizlik açısından daha yüksek performans sergileyerek yanıtlarında istenilen bilgiyi sunmanın yanı sıra ek açıklamalarla içerik zenginliği sağlamıştır. Buna karşılık, Gemini modeli, daha sade ve anlaşılır yanıtlar sunarak okunabilirlik açısından üstün performans göstermiştir. Sıklıkla kullanılan ChatGPT ve Gemini geniş dil modelleri Türkçe dilinde ortodonti alanında hasta ve hasta yakınlarının bilgi ihtiyacını karşılamada önemli bir potansiyele sahiptir. Ancak, daha etkili bir hasta bilgilendirme süreci için modellerin yanıtlarının daha sade ve anlaşılır hale getirilmesi gerektiği sonucuna ulaşılmıştır.

KAYNAKLAR

1. Tamkin A, Brundage M, Clark J, Ganguli D. Understanding the Capabilities, Limitations, and Societal Impact of Large Language Models 2021; arXiv: 2102.02503

2. Johnson D, Goodman R, Patrinely J, Stone C, Zimmerman E et al. Assessing the Accuracy and Reliability of AI-Generated Medical Responses: An Evaluation of the Chat-GPT Model. Res Sq 2023; rs.3.rs-2566942.

3. Kılınç DD, Mansız D. Examination of the reliability and readability of Chatbot Generative Pretrained Transformer's (ChatGPT) responses to questions about orthodontics and the evolution of these responses in an updated version. Am J Orthod Dentofacial Orthop 2024; 165: 546-555.

4. Demir GB, Süküt Y, Duran GS, Topsakal KG, Görgülü S. Enhancing systematic reviews in orthodontics: a comparative examination of GPT-3.5 and GPT-4 for generating PICO-based queries with tailored prompts and configurations. Eur J Orthod 2024; 46.

5. Arslan C, Kahya K, Cesur E, Cakan DG. An evaluation of orthodontic information quality regarding artificial intelligence (AI) chatbot technologies: a comparison of ChatGPT and google BARD. Aust Orthod J 2024; 40: 149-157.

6. Meyer MKR, Kandathil CK, Davis SJ, Durairaj KK, Patel PN et al. Evaluation of Rhinoplasty Information from ChatGPT, Gemini, and Claude for Readability and Accuracy. Aesthetic Plast Surg 2024: 1-6.

7. Duran GS, Yurdakurban E, Topsakal KG. The Quality of CLP-Related Information for Patients Provided by ChatGPT. Cleft Palate Craniofac J 2023: 10556656231222387.

8. Tanaka OM, Gasparello GG, Hartmann GC, Casagrande FA, Pithon MM. Assessing the reliability of ChatGPT: a content analysis of self-generated and self-answered questions on clear aligners, TADs and digital imaging. Dental Press J Orthod 2023; 28: e2323183.

9. Abu Arqub S, Al-Moghrabi D, Allareddy V, Upadhyay M, Vaid N et al. Content analysis of AI-generated (ChatGPT) responses concerning orthodontic clear aligners. Angle Orthod 2024; 94: 263-272.

10. Jebb AT, Ng V, Tay L. A Review of Key Likert Scale Development Advances: 1995–2019. Front Psychol 2021; 12: 637547.

11. Ateşman E. Türkçede okunabilirliğin ölçülmesi. Dil Dergisi. 1997; 58: 71-74.

12. Dursun D, Bilici Geçer R. Can artificial intelligence models serve as patient information consultants in orthodontics? BMC Med Inform Decis Mak 2024; 24.

13. Taşkın S, Cesur MG, Uzun M. Yapay Zekâ Destekli Sohbet Robotlarının Yaygın Ortodontik Soruları Cevapla-

ma Başarısının Değerlendirilmesi. Med J SDU. 2023; 30: 680-686.
14. Ley P, Florio T. The use of readability formulas in health care. Psychol Health Med. 1996; 1: 7-28.
15. Yılancı HÖ, Akkaya N, Akçiçek G. Oral Ülser ve Rekürrent Aftöz Stomatit ile ilgili Türkçe İnternet Sitelerindeki Hasta Bilgilendirme Metinlerinin İçerik Kalitesi ve Okunabilirliği. ADO Klinik Bilimler Dergisi. 2023; 12: 266-272.
16. Taşdemir İ. İnternet Ortamındaki Dişeti Hastalığı ile İlgili Bilgilerin Okunabilirlik Analizi. Selcuk Dental Journal. 2023; 10: 89-93.
17. Akbulut AS. İnternet Ortamındaki Şeffaf Plak Tedavisi ile İlgili Bilgilerin Okunabilirlik Analizi. NEU Dent J 2022; 4: 7-11.
18. Deniz ÇD, Kozanhan B, Tutar MS, Özler S. Üçlü test ile ilgili internet bilgilendirme metinlerinin okunabilirlik ve içeriklerinin değerlendirilmesi. Mersin Univ Sağlık Bilim Derg 2020; 13: 35-44.
19. Kılınç DD, Mansız D. Examination of the reliability and readability of Chatbot Generative Pretrained Transformer's (ChatGPT) responses to questions about orthodontics and the evolution of these responses in an updated version. Am J Orthod Dentofacial Orthop 2024; 165: 546-555.
20. Kahraman M. Sözlük Bilim Kuram, İlke ve Yöntemler Üzerine. İnsan ve Toplum Bilimleri Araştırmaları Dergisi, 2016; 5.8: 3288-3312.