DOI: 10.24938/kutfd.1784362 Kırıkkale Üniversitesi Tıp Fakültesi Dergisi 2026;28(1):49-54
Özgün Araştırma *Original Article*

PEDODONTİ SORULARININ YANITLANMASINDA YAPAY ZEKÂ PERFORMANSINA DİLİN ETKİSİ: CHATGPT-4.0 VE DEEPSEEK-R1 İLE TÜRKÇE VE İNGİLİZCE KARŞILAŞTIRMASI

The Effect of Language on Artificial Intelligence Performance in Answering Pedodontics Questions: Comparison of Turkish and English with ChatGPT-4.0 and DeepSeek-R1

Esra HATO[1] **Koray PEKER**[1]

[1] *Kırıkkale Üniversitesi, Diş Hekimliği Fakültesi, Çocuk Diş Hekimliği ABD, KIRIKKALE, TÜRKİYE*

ÖZ

Amaç: Çalışmamızda, *ChatGPT-4.0* ve *DeepSeek R1* adlı iki farklı sohbet robotunun pedodonti alanındaki çoktan seçmeli sorulardaki gösterdiği başarının ve kullanılan dilin bu başarı üzerindeki etkisinin değerlendirilmesi amaçlanmıştır.

Gereç ve Yöntemler: 2 farklı yapay zekâ sohbet robotunun cevaplandırması için pedodonti konularından 20 soru oluşturuldu. Bu soruların İngilizce ve Türkçe versiyonları *ChatGPT-4.0* ve *DeepSeek R1* yapay zekâ sohbet robotlarına cevaplandırılması için soruldu. Alınan cevaplar doğru ve yanlış olarak kaydedildi. Analizler *IBM Statistical Package for Social Sciences* (SPSS) *Windows version* 27 (SPSS Inc. Chicago, IL, USA) programında gerçekleştirilmiştir. Çalışmada anlamlılık düzeyi $p<0,05$ olarak alınmıştır.

Bulgular: *ChatGPT*'nin her iki dilde de doğruluk oranı %89 olarak bulunmuş; *Deepseek*'in Türkçe dilindeki doğruluk oranı %90, İngilizce dilindeki doğruluk oranı ise %92,5 olarak bulunmuştur. Model ve dil değişkeni açısından doğruluk oranı arasında istatistiksel olarak anlamlı bir fark elde edilememiştir.

Sonuç: Çalışmamızda iki modelin de Türkçe ve İngilizce cevap oranlarının benzer olduğu görülmüştür. *DeepSeek* İngilizce dilinde daha iyi performans göstermiştir. Kullanımlarının kolay olması ve metin tabanlı çoktan seçmeli soruları cevaplamadaki güçlü performansları göz önüne alındığında, *ChatGPT-4.0* ve *DeepSeek* gibi büyük dil modelleri diş hekimliği eğitiminde öğrenmeyi desteklemek için kullanılabilir bir araç olarak değerlendirilebilir.

Anahtar Kelimeler: *Çocuk diş hekimliği, yapay zekâ, diş hekimliği eğitimi*

ABSTRACT

Objective: Our study aimed to evaluate the performance of two different chatbots, ChatGPT-4.0 and DeepSeek R1, on multiple-choice questions in the field of pedodontics, as well as the impact of the language used on this performance.

Material and Methods: Twenty questions on pedodontics topics were created for the two different AI chatbots to answer. English and Turkish versions of these questions were presented to the ChatGPT 4.0 and DeepSeek R1 AI chatbots. The answers were recorded as either correct or incorrect. Analyses were performed using IBM Statistical Package for Social Sciences (SPSS) Windows version 27 (SPSS Inc. Chicago, IL, USA). The significance level in this study was set at $p<0.05$.

Results: ChatGPT's accuracy rate was 89% in both languages, whereas DeepSeek achieved 90% in Turkish and 92.5% in English. There was no statistically significant difference in accuracy rates across models or languages.

Conclusion: Our study found that the response rates for both models in Turkish and English were similar. DeepSeek performed better with the English language. Due to their user-friendly nature and impressive ability to answer text-based, multiple-choice questions, large language models such as ChatGPT-4.0 and DeepSeek are viable tools for supporting learning in dental education.

Keywords: *Pediatric dentistry, artificial intelligence, dental education*

Yazışma Adresi/Correspondence: **Dr. Esra HATO**
Kırıkkale Üniversitesi, Diş Hekimliği Fakültesi, Çocuk Diş Hekimliği ABD, KIRIKKALE, TÜRKİYE
Tel/Phone: +905446054829 **E-posta**/E-mail: esrahato@gmail.com
Geliş Tarihi/Received: 15.09.2025 **Kabul Tarihi**/Accepted: 23.01.2026

KÜ Tıp Fak Derg 2026;28(1):49-54
Doi: 10.24938/kutfd.1784362

GİRİŞ

Büyük dil modelleri, insan benzeri metinler üretmek için büyük veri kümeleri üzerinde eğitilen çok katmanlı ve tekrarlayan sinir ağı yapılarına dayalı yapay zekâ sistemleridir.[1] Günümüzde yapay zekâ destekli modeller, kapsamlı veri kümeleri üzerinde eğitilerek doğal, anlamlı ve insan benzeri etkileşimlerde bulunabilme kapasitesine ulaşmıştır.[2] Bu sistemler; yazma, analiz etme ve problem çözme gibi karmaşık görevlerde kullanılabilmektedir ve kısa sürede sağlık, eğitim, araştırma ve endüstri gibi alanlarda yaygın şekilde kullanılmaya başlanmıştır.[3,4]

Diş hekimliği alanına yapay zekâ teknolojisinin entegre edilmesi, diş hekimliği eğitiminin ve klinik uygulamalarının gelişmesine katkı sağlamaktadır.[5,6] Klinik uygulamalarda yapay zekâ teknolojisinin kullanımı tanı doğruluğunun artırılması ve tedavi prognozunun daha iyi tahmin edilmesi ile tedavi planlamasının iyileştirilmesini sağlayabilmektedir. Çocuk diş hekimliği alanında yapay zekâ; erken çocukluk çağı çürüklerinin, dental plakların, sürnümerer diş varlığının, ektopik erüpsiyonun tespiti; diş yaşı tahmini, süt ve daimî dişlerin ayırt edilmesi ile fissür örtücü sınıflaması gibi alanlarda kullanılabilmektedir.[7,8] Yapay zekâ sistemleri hem öğrencilerin öğrenme deneyimlerinde hem eğiticilerin eğitsel değerlendirme süreçlerinde köklü değişiklikler yapabilecek potansiyele sahiptir.[9] Bu sistemler öğrencilerin sorularını yanıtlama, kavramların anlaşılmasına yardımcı olma ve öğrenci performansını değerlendirme gibi amaçlarla kullanılabilmektedir.[10]

Önceki çalışmalar yapay zekâ sohbet robotlarının diş hekimliği uzmanlık sınavlarındaki performanslarını değerlendirmiştir ve bazı çalışmalarda sohbet robotlarının sınavdan geçer not alamadığını ve henüz insan seviyesinde bir performansa ulaşamadığı bildirilmişken; bazı çalışmalarda ise başarılı performans gösterdiğin bildirilmiştir. [11-15]

Sohbet robotlarının yeterliliği ve performansları, eğitildikleri eğitim verilerinin dilsel çeşitliği ve kalitesiyle yakından ilişkilidir. Her ne kadar birçok model çok dilli sürümler sunsa da eğitim korpuslarının çoğunluğunun İngilizce odaklı olması, İngilizce dışındaki dillerde düşük performansa sebep olabilmektedir.[16,17] Bu nedenle, yapay zekâ tabanlı sohbet robotlarının Türkçe dilindeki akademik ve mesleki yeterliliğin değerlendirilmesine yönelik çalışmalara ihtiyaç vardır.

Çalışmamızda, *ChatGPT-4.0* ve *DeepSeek R1* adlı iki farklı sohbet robotunun çocuk diş hekimliği alanındaki çoktan seçmeli sorulardaki gösterdiği başarının ve kullanılan dilin bu başarı üzerindeki etkisinin değerlendirilmesi amaçlanmıştır. Çalışmamızın sıfır hipotezi, '*ChatGPT-4.0* ve *DeepSeek R1* adlı yapay zekâ destekli sohbet robotlarının, pedodonti alanındaki çoktan seçmeli soruları cevaplama performansları arasında fark yoktur ve kullanılan dilin cevaplama performansları üzerinde belirleyici bir etkisi yoktur' şeklindedir.

GEREÇ VE YÖNTEM

Çalışmada, insan ve hayvan konuları ele alınmadığı için etik kurul onayına ve Helsinki Deklarasyon prensipleri uyumuna gerek duyulmamıştır.

Bu çalışma için sohbet robotlarına sorulmak üzere çocuk diş hekimliği konularına ilişkin teorik bilgiyi ölçen, metin tabanlı, çoktan seçmeli 20 adet soru hazırlandı (Ek-1). Hazırlanan sorular, çocuk diş hekimliği alanında deneyimli iki uzman akademisyen tarafından içerik uygunluğu, bilimsel doğruluk ve dil ifadesi açısından bağımsız olarak (uygun, uygun değil ve düzeltilmeli şeklinde) değerlendirilmiştir. Geri bildirimler doğrultusunda gerekli düzenlemeler yapılmış ve soruların son hali onaylanmıştır. Soruların İngilizce çevirileri profesyonel bir çevirmen tarafından yapılmıştır.

Tüm sohbet robotu yanıtları Haziran 2025'te alınmıştır. Bu çalışmada *ChatGPT-4.0* (OpenAI, San Francisco, CA, ABD) ve *DeepSeek R1* (Hangzhou DeepSeek Yapay Zekâ Temel Teknoloji Araştırma Şirketi; Pekin, Çin) ücretsiz sürümleri kullanılmıştır.

Her modele soru sorulmadan önce geçmiş ve çerezler temizlenmiştir. Ayrıca, yeni bir tarayıcı sekmesi açılmış ve etkileşimleri sıfırlamak için her yanıttan önce "yeni konuşma" seçeneği kullanılmıştır. Tüm sorular, ifade, noktalama veya sözdiziminde herhangi bir değişiklik yapılmadan girilmiştir. Prosedürel tutarlılığı sağlamak ve insan hatası riskini en aza indirmek için tek bir araştırmacı (K.P.) tarafından, her oturumda sistematik olarak iki sohbet robotuna da aynı sırayla, aynı soru setinin İngilizce ve Türkçe versiyonlarını yöneltmiştir. 2 farklı sohbet robotuna Türkçe ve İngilizce 20 sorunun 10 kez tekrarlanarak yöneltilmesi ile toplamda 800 yanıt elde edilmiştir.

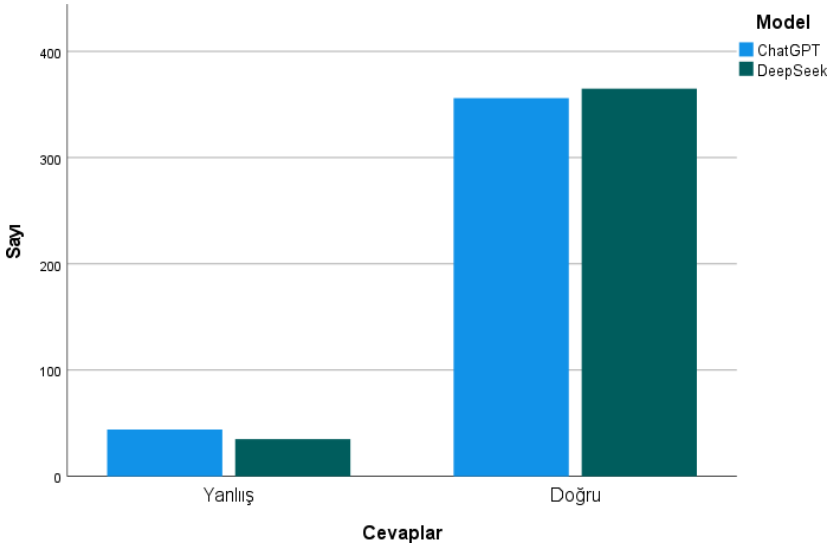
İstatistiksel analiz

Bu çalışmada verilerin tanımlayıcı istatistikleri (sayı, yüzde, ortalama, standart sapma, medyan, minimum ve maksimum) verilmiştir. Normal dağılım varsayımı *Shapiro-Wilk* testi ile kontrol edilmiştir. Normallik varsayımının karşılanmadığı durumlarda bağımsız iki grubun karşılaştırılmasında *Mann-Whitney U* testi uygulanmıştır. Kategorik değişkenleri arasındaki ilişkinin test edilmesinde örneklem boyutu varsayımı (beklenen değer >5) karşılandığı durumlarda *Pearson* Ki-Kare testi uygulanmıştır. Analizler *IBM Statistical Package for Social Sciences* (SPSS) *Windows version 27* (SPSS Inc. Chicago, IL, USA) programında gerçekleştirilmiştir.

KÜ Tıp Fak Derg 2026;28(1):49-54
Doi: 10.24938/kutfd.1784362

BULGULAR

Hazırlanan 20 sorunun Türkçe ve İngilizce versiyonları her iki sohbet robotuna 10 farklı zamanda yöneltilmiştir ve 800 cevap elde edilmiştir. *ChatGPT* ve *DeepSeek*'in Türkçe ve İngilizce dilindeki toplam doğru-yanlış cevap dağılımı Şekil 1'de gösterilmiştir.



Şekil 1: Modellere göre toplam cevapların dağılımlarına ait çubuk grafiği

Tablo 1'de yapay zekâ sohbet robotlarının modellere ve dillere göre cevap dağılımları verilmiştir. *ChatGPT* Türkçe ve İngilizce sorularda 178 doğru cevap verirken; *DeepSeek* Türkçe sorularda toplam 180, İngilizce sorularda ise 185 doğru cevap vermiştir. Yapılan analizler sonucunda doğru/yanlış cevap dağılımlarında hem model hem dil değişkenleri açısından her iki dil arasında istatistiksel olarak anlamlı bir fark bulunmamıştır.

Tablo 1: Çalışmada kullanılan yapay zekâ sohbet robotlarının cevap dağılımları ve karşılaştırılması

		Yanlış			Doğru				
		n	**%**	**%S.**	**n**	**%**	**%S.**	**Test İstatistiği***	**p**
Türkçe	*ChatGPT*	22	11,0	52,4	178	89,0	49,7	0,106	0,744
	DeepSeek	20	10,0	47,6	180	90,0	50,3		
İngilizce	*ChatGPT*	22	11,0	59,5	178	89,0	49,0	1,459	0,227
	DeepSeek	15	7,5	40,5	185	92,5	51,0		
Toplam	*ChatGPT*	44	11,0	55,7	356	89,0	49,4	1,138	0,286
	DeepSeek	35	8,8	44,3	365	91,3	50,6		
ChatGPT	Türkçe	22	11,0	50,0	178	89,0	50,0	0,000	1,000
	İngilizce	22	11,0	50,0	178	89,0	50,0		
DeepSeek	Türkçe	20	10,0	57,1	180	90,0	49,3	0,783	0,376
	İngilizce	15	7,5	42,9	185	92,5	50,7		
Toplam	Türkçe	42	10,5	53,2	358	89,5	49,7	0,351	0,553
	İngilizce	37	9,3	46,8	363	90,8	50,3		

%: Satır yüzdesi, %S.: Sütun yüzdesi,*Pearson Ki-kare testi

Dil ve model değişkenlerine ait doğruluk oranlarının dağılımları Tablo 2'de verilmiştir. *ChatGPT*'nin her iki dilde de doğruluk oranı %89; *Deepseek*'in Türkçe dilindeki doğruluk oranı %90, İngilizce dilindeki doğruluk oranı ise %92,5 olarak bulunmuştur. Model ve dil değişkeni açısından doğruluk oranında her iki dil açısından istatistiksel olarak anlamlı bir fark elde edilememiştir.

Tablo 2: Dil ve modellere için doğruluk oranlarının dağılımları ve karşılaştırılması

		Min.-Maks.	**Ort.±S.S. (Medyan)**	**Test İstatistiği***	**p**
ChatGPT	Türkçe	80-95	89±5,16(90)	-0,662	0,529
	İngilizce	60-100	89±11,5(90)		
DeepSeek	Türkçe	80-95	90±4,71(90)	-1,258	0,280
	İngilizce	90-95	92,5±2,64(92,5)		
Türkçe	*ChatGPT*	80-95	89±5,16(90)	-0,477	0,684
	DeepSeek	80-95	90±4,71(90)		
İngilizce	*ChatGPT*	60-100	89±11,5(90)	-0,600	0,579
	DeepSeek	90-95	92,5±2,64(92,5)		

Min: Minimum, Max, Maksimum, Ort: Ortalama, SS: Standart sapma, * Mann Whitney U testi

KÜ Tıp Fak Derg 2026;28(1):49-54
Doi: 10.24938/kutfd.1784362

Tablo 3'te modellerin dillere göre cevap tutarlılığının dağılımı verilmiştir. *ChatGPT* Türkçe ve İngilizce soruların 168'ine aynı cevabı, 32'sine farklı cevaplar vermiştir. *ChatGPT* her iki dilde de 6 soruya yanlış, 162 soruya doğru cevap vermiştir. *DeepSeek* ise Türkçe ve İngilizce soruların 175'ine aynı cevabı, 25'ine ise farklı cevapları vermiştir. Modellerin doğru cevap tutarlılıkları yüksek, yanlış cevap tutarlılıkları ise düşük bulunmuştur. Genel olarak *DeepSeek* modelinin %87,5 tutarlılık oranına, *ChatGPT* modelinin ise %84 oranında tutarlılık oranına sahip olduğu görülmüştür. Dillere göre cevapların tutarlılığının incelenmesinde Kappa istatistikleri kontrol edilmiştir. Analizler sonucunda her iki modelde ve toplam cevaplarda dillerdeki uyumun çok düşük düzeyde olduğu görülmüştür (p<0,05).

Tablo 3: Modellerin Dillere göre cevap tutarlılığının dağılımı ve aralarındaki ilişkiler

		Türkçe							
		Yanlış			Doğru				
	İngilizce	n	%	%S.	n	%	%S.	Kappa İstatistiği	p
ChatGPT	Yanlış	6	27,3	27,3	16	72,7	9,0	0,183	0,010*
	Doğru	16	9,0	72,7	162	91,0	91,0		
DeepSeek	Yanlış	5	33,3	25,0	10	66,7	5,6	0,219	0,002*
	Doğru	15	8,1	75,0	170	91,9	94,4		
Total	Yanlış	11	29,7	26,2	26	70,3	7,3	0,200	<0,001*
	Doğru	31	8,5	73,8	332	91,5	92,7		
Model		Tutarlı cevap (Doğru+Yanlış) (n)					Toplam (N)		Tutarlılık oranı
ChatGPT		168					200		84,0
DeepSeek		175					200		87,5
Toplam		343					400		85,75
		Doğru cevap tutarlılığı (n)					Toplam (N)		Tutarlılık oranı
ChatGPT		162					178		91,0
DeepSeek		170					185		91,9
Genel Toplam		332					363		91,5
		Yanlış cevap tutarlılığı (n)					Toplam (N)		Tutarlılık oranı
ChatGPT		6					22		27,3
DeepSeek		5					15		33,3
Genel Toplam		11					37		29,7

*p<0,05, %: Satır yüzdesi, %S.: Sütun yüzdesi

TARTIŞMA

Bu çalışmanın amacı çocuk diş hekimliğine ait soruları cevaplamada *ChatGPT-4.0* ve *DeepSeek R1* adlı iki farklı sohbet robotunun performanslarını 2 farklı dilde değerlendirmek ve karşılaştırmaktır. Çalışmamızın sıfır hipotezi, her iki modelin soruları cevaplama performansları arasında fark olmadığı ve kullanılan dilin cevaplama performansları üzerinde belirleyici bir etkisinin olmadığı yönündedir.

Genel doğruluk oranları *DeepSeek* için %91,25, *ChatGPT* için ise %89 olarak bulunmuştur. *DeepSeek* her iki dilde de en yüksek doğruluk oranına ulaşmış olsa da model veya dil değişkeni arasında istatistiksel anlamlı fark bulunmamıştır. Bu durum modellerin çok dilli yanıt üretme performanslarının benzer olduğunu ve kullanılan dilin cevaplama performansı üzerinde etkisinin belirleyici olmadığını göstermektedir. Bu doğrultuda çalışmamızın sıfır hipotezi kabul edilmiştir. Ayrıca *DeepSeek* modelinin diller arasında doğru-yanlış cevap değişiminin (n=25), *ChatGPT*'ye (n=32) göre daha az gözlenmesi yanıt tutarlılığı açısından önemli bir bulgu olarak kabul edilebilir. Bu farklılığın ortaya çıkmasında yapay zekâ uygulamalarındaki halüsinasyon eğilimlerinden, diller arası geçiş, modellerin geliştirilmesinde kullanılan veri kümelerindeki dil dağılım dengesizliği sebep olabilir. Genel olarak her iki modelde de Türkçe ve İngilizce dillerindeki doğru cevap oranlarının yüksek olması, çok dilli kullanım açısından olumlu olarak değerlendirilebilir.

Çoğu sohbet robotunun eğitim verilerinin ağırlıklı olarak İngilizce olması, İngilizce dışındaki bağlamlardaki yeterlilikleri konusunda endişelere sebep olmuştur.[18,19] Ancak büyük dil modellerinin düzenli olarak güncellenmesi ve yeni versiyonlarının sunulması, çeviri ve çok dilli giriş yeteneklerinin de gelişmesini olumlu yönde etkileyecektir. Böylece İngilizce dışındaki dillerde ortaya çıkabilecek sorunların azalmasına katkı sağlayacaktır.

Literatürde genel olarak tıp ve diş hekimliği alanlarına ait soruların çeşitli yapay zekâ sohbet robotları tarafından yanıtlama performanslarının değerlendirildiği çalışmalar mevcuttur ve çoğunlukla *ChatGPT*'nin farklı sürümleri kullanılmıştır.[2,17,20-22] Bildiğimiz kadarıyla çocuk diş hekimliği alanında Türkçe ve İngilizce dilindeki *DeepSeek* ve *ChatGPT* sohbet robotlarının yanıt performanslarının değerlendirildiği ve karşılaştırıldığı bir çalışma bulunmamaktadır.

Büyük dil modellerinin Türkçe ve İngilizce dillerindeki yanıt performanslarını değerlendiren sınırlı sayıda çalışma vardır.[17,23,24] ChatGPT'nin iki sürümünün (3.5 ve 4.0) genel cerrahi uzmanlık sorularındaki

performansının karşılaştırıldığı çalışmada: Türkçe sorularda, *ChatGPT-3.5*'in %66,66, *ChatGPT-4*'ınise %69,41 doğruluk oranına ulaştığı ve *ChatGPT-4*'ün Türkçe dilinde daha iyi performans gösterdiği bildirilmiştir ($p<0,05$).[17] İngilizce sorularda ise *ChatGPT-3.5*'un %67,05, *ChatGPT-4*'ün ise %70,19 doğruluk oranına ulaştığı; *ChatGPT-4*'ün İngilizce dilinde daha iyi performans gösterdiği bulunmuştur ($p<0,05$). *ChatGPT*'nin aynı sürümlerinin Türkçe ve İngilizce performansları karşılaştırıldığında, her iki model de İngilizce'de biraz daha yüksek doğruluk göstermesine rağmen, bu farklılıklar istatistiksel olarak anlamlı bulunmamıştır ($p>0,05$).[17] Üç farklı yapay zekâ sohbet robotunun okülofasiyal plastik ve orbita cerrahisi sorularındaki performanslarının karşılaştırıldığı çalışmada Türkçe'de *ChatGPT-3.5* %23,3, *Copilot* %63,3, *Gemini* %33,3 doğruluk oranına ulaşmıştır. İngilizce'de ise, doğruluk oranları *ChatGPT-3.5* için %43,3, *Copilot* için %73,3, *Gemini* için %46,7 olarak bildirilmiştir.[23] *ChatGPT-3,5*, İngilizce ve Türkçe'de benzer performanslar göstermiştir.[23] *ChatGPT-40* (GPT), *Claude 3.5 Sonnet (Claude*), *Grok* 2 (Grok) ve *Gemini 2.0 Advanced* (Gemini) sohbet robotlarının rejeneratif endodonti alanındaki Türkçe ve İngilizce açık uçlu sorulara verdiği yanıtların değerlendirildiği çalışmada ise İngilizce'de Türkçe'ye göre önemli ölçüde daha yüksek doğruluk oranları elde edildiği bildirilmiştir.[24] Çalışmamızdaki sohbet robotlarının Türkçe ve İngilizce'deki performanslarının benzer olması önceki çalışmaların sonuçları ile uyumludur. Aynı zamanda çalışmamızda *ChatGPT* ve *DeepSeek* modellerinin her iki dilde de doğruluk oranları bahsedilen çalışmalardan yüksek bulunmuştur. Bu farklılıkların ortaya çıkmasında kullanılan büyük dil modellerinin ve versiyonlarının farklılık göstermesi, soruların ait olduğu bilimsel alanların ve içerik özelliklerinin değişkenlik göstermesi gibi sebeplerin yanı sıra, sürekli güncellemelerle modellerin geliştirilmesinin de saptanan yüksekliğe katkı sağlamış olabileceğini düşünmekteyiz.

Aşık ve Kuru'nun diş hekimliği uzmanlık eğitimi giriş sınavında sorulan çocuk diş hekimliğine ait sorular üzerinden *ChatGPT* sohbet robotunun yanıtlama performansını değerlendirdikleri çalışmada, doğruluk oranın yıllara göre en düşük %30, en yüksek %90 olduğu bildirilmiştir.[25] Bizim sonuçlarımızda *ChatGPT* için doğruluk oranları en düşük %60, en yüksek %100 olarak bulunmuştur. Bu durum, bilim alanı aynı olmasına rağmen soruların ve kullanılan versiyonların farklı olmasının model performansını etkileyebileceğini göstermektedir.

Wu ve ark. *ChatGPT-4.0*'nun Tayvan Ulusal Diş Hekimliği Lisanslama Sınavlarındaki oral patoloji sorularını yanıtlama performansı üzerinde dil ve soru tiplerinin etkisini incelemiş ve İngilizce çevirinin, başarıyı anlamlı şekilde artırdığını bildirmiştir.[22] Fang ve ark. Çince tıbbi lisanslama sınav sorularının İngilizce'ye çevrilmesinin, *ChatGPT-4* kullanıldığında doğru yanıt sayısını 256'dan 260'a çıkardığını, ancak farkın anlamlı düzeye ulaşmadığını bildirmiştir.[26] Çalışmamızda Türkçe ve İngilizce dili arasında *ChatGPT* modelinde doğru sayılarında bir farklılık gözlenmemiştir. Bu durum Türkçe ve Çince dil yapı farklılıklarından kaynaklanmış olabilir.

Çalışmamızdaki her iki modellin de yüksek doğruluk oranına sahip olması görsel tabanlı sorular yerine metin tabanlı soruların kullanılmasından kaynaklanmış olabilir. Literatürde soru setinde görsel tabanlı soruların sayısının artması ile modellerin doğruluk oranlarının düşmesine sebep olduğu ve görsel soruların İngilizce olarak sorulmasında bile doğruluk oranın belirgin derecede artmadığı bildirilmiştir.[22] Bu durum yapay zekâ modellerinin görsel veri işlemesinin zor olması ile ilişkilendirilebilir.

Çalışmamızda soru sayısının görece olarak az olması, soruların yalnızca tek tip (çoktan seçmeli) ve metin tabanlı olması önemli bir sınırlılık oluşturmaktadır. Çalışmamızda soruların konularına göre alt başlıklara ayrılarak sohbet robotlarının cevaplama performansının değerlendirilmesi ve karşılaştırılması yapılmamıştır. Ayrıca soruların zorluk seviyelerine göre bir gruplama yapılmamıştır. İngilizce çeviride profesyonel çevirmen tarafından çeviri yapılmış olup geri-çeviri yöntemi uygulanmaması da sınırlılıklardan biridir. İleri araştırmalarda farklı soru formatlarıyla ve daha geniş veri setleriyle ve farklı diş hekimliği branşlarında değerlendirme yapılması faydalı olacaktır.

Sonuç olarak çalışmamızda iki model de pedodonti sorularını yüksek doğruluk oranıyla cevaplamış ve soruların İngilizceye çevrilmesi sonucunda *ChatGPT* doğruluk oranı aynı kalmışken *DeepSeek* doğruluk oranı artmıştır. Kullanımlarının kolay olması ve metin tabanlı çoktan seçmeli soruları cevaplamadaki güçlü performansları göz önüne alındığında, *ChatGPT-4.0* ve *DeepSeek* gibi büyük dil modelleri diş hekimliği eğitiminde öğrenmeyi desteklemek için kullanılabilir bir araç olarak değerlendirilebilir. Sağlık profesyonelleri yapay zekâ teknolojisindeki son gelişmelerden haberdar olmalı ve bunların uygulama ve araştırmalar üzerindeki potansiyel etkilerinin farkında olmalıdır. Aynı zamanda yapay zekâ sohbet robotlarını kullanırken cevapların değerlendirilmesinde temkinli bir yaklaşım benimsenmesi gerektiği unutulmamalıdır.

Çıkar Çatışması Beyanı: Yazarlar herhangi bir çıkar çatışması olmadığını beyan ederler.

Katkı Oranı Beyanı: Anafikir/Planlama: EH, KP; Analiz/Yorum: EH, KP; Veri Sağlama: KP; Yazım: EH; Gözden Geçirme ve Düzeltme: EH, KP; Onaylama: EH, KP.

Destek / Teşekkür Beyanı: Herhangi bir kurum veya kişiden finansal destek alınmamıştır.

Etik Kurul Onayı: Çalışma, insan ve hayvan konularını ele almadığı için etik kurul onayına ve Helsinki Deklarasyon prensipleri uyumuna gerek duyulmamıştır. Bu özgün araştırma 31. Uluslararası Türk Pedodonti Derneği Kongresi'nde (4-7 Ekim 2025) 'Pedodonti Sorularının Yanıtlanması Yapay Zekâ Performansına Dilin Etkisi: ChatGPT-4.0 ve DeepSeek-R1 ile Türkçe ve İngilizce Karşılaştırması' adı altında sözlü sunum olarak sunulmuştur.

KAYNAKLAR

1. Kaygisiz ÖF, Teke MT. Can Deepseek and ChatGPT be used in the diagnosis of oral pathologies? *BMC Oral Health*. 2025;25(1):638.
2. Sarı MBD, Sezer B. ChatGPT-4 Omni's accuracy in multiple-choice dentistry questions: A multidisciplinary and bilingual assessment. *Essent Dent*. 2025;4:1-9.
3. Mohammad-Rahimi H, Setzer FC, Aminoshariae A, Dummer PMH, Duncan HF, Nosrat A. Artificial intelligence chatbots in endodontic education-Concepts and potential applications. *Int Endod J*. 2025;00:1-14.
4. Choudhury A, Shahsavar Y, Shamszare H. User intent to use DeepSeek for healthcare purposes and their trust in the large language model: Multinational survey study. *JMIR Hum Factors*. 2025;12:e72867.
5. Kusaka S, Akitomo T, Hamada M, et al. Usefulness of generative artificial intelligence (AI) tools in pediatric dentistry. *Diagnostics*. 2024;14(24):2818.
6. Kukreja P. Integration of artificial intelligence in dentistry: A systematic review of educational and clinical implications. *Cureus*. 2025;17(2):e79350-e79350.
7. Vishwanathaiah S, Fageeh HN, Khanagar SB, Maganur PC. Artificial intelligence its uses and application in pediatric dentistry: A review. *Biomedicines*. 2023;11(3):788.
8. Alessa N. Application of artificial intelligence in pediatric dentistry: A literature review. *J Pharm Bioallied Sci*. 2024;16(Suppl 3):S1938-S1940.
9. Ahmed WM, Azhari AA, Alfaraj A, Alhamadani A, Zhang M, Lu CT. The quality of AI-generated dental caries multiple choice questions: A comparative analysis of ChatGPT and Google Bard language models. *Heliyon*. 2024;10(7):e28198.
10. Fang Q, Reynaldi R, Araminta AS, et al. Artificial Intelligence (AI)-driven dental education: Exploring the role of chatbots in a clinical learning environment. *J Prosthet Dent*. 2024;134(4):1296-1303.
11. Lin CCC, Sun JS, Chang CH, Chang YH, Chang JZC. Performance of artificial intelligence chatbots in national dental licensing examination. *J Dent Sci*. 2025;20:2307-2314.
12. Kim W, Kim BC, Yeom HG. Performance of large language models on the Korean dental licensing examination: A comparative study. *Int Dent J*. 2025;75(1):176-184.
13. Chau RCW, Thu KM, Yu OY, Hsung RTC, Lo ECM, Lam WYH. Performance of generative artificial intelligence in dental licensing examinations. *Int Dent J*. 2024;74(3):616-621.
14. Oğuzman RT, Yurdabakan ZZ. Performance of chat generative pretrained transformer and bard on the questions asked in the dental specialty entrance examination in Turkey regarding bloom's revised taxonomy. *Curr Res Dent Sci*. 2024;34(1):25-34.
15. Sismanoglu S, Capan BS. Performance of artificial intelligence on Turkish dental specialization exam: Can ChatGPT-4.0 and gemini advanced achieve comparable results to humans? *BMC Med Educ*. 2025;25(1):214.
16. Yao Z, Duan L, Xu S, Chi L, Sheng D. Performance of large language models in the non-English context: Qualitative study of models trained on different languages in Chinese medical examinations. *JMIR Med Inform*. 2025;13(1):e69485.
17. Orman S. ChatGPT's medical exam performance: Version and language analysis in general surgery fellowship exam. *Eur J Hum Health*. 2025;4(1):1-9.
18. Kim MG, Hwang G, Chang J, Chang S, Roh HW, Park RW. Performance of open-source large language models in psychiatry: Usability study through comparative analysis of non-English records and English translations. *J Med Internet Res*. 2025;27:e69857.
19. Liu X, Wu J, Shao A, et al. Uncovering language disparity of ChatGPT on retinal vascular disease classification: Cross-sectional study. *J Med Internet Res*. 2024;26:e51926.
20. Bilgin Avşar D, Ertan AA. Diş hekimliğinde uzmanlık sınavında sorulan protetik diş tedavisi sorularının ChatGPT-3.5 ve Gemini tarafından cevaplanma performanslarının karşılaştırmalı olarak incelenmesi: Kesitsel araştırma. *Türkiye Klinikleri Diş Hekimliği Bilimleri Derg*. 2024;30(4):668-673.
21. Mahmoud R, Shuster A, Kleinman S, Arbel S, Ianculovici C, Peleg O. Evaluating artificial intelligence chatbots in oral and maxillofacial surgery board exams: Performance and potential. *J Oral Maxillofac Surg*. 2025;83(3):382-389.
22. Wu YH, Tso KY, Chiang CP. Impact of language and question types on ChatGPT-4o's performance in answering oral pathology questions from Taiwan National Dental Licensing Examinations. *J Dent Sci*. 2025;20(4):2176-2180.
23. Şensoy E, Çıtırık M. Okülofasiyal plastik ve orbital cerrahide İngilizce ve Türkçe dil çeşitliliğinin yapay zekâ chatbot performansına etkisi: ChatGPT-3.5, Copilot ve Gemini üzerine bir çalışma. *Osman Tıp Derg*. 2024;46(5):781-786.
24. Büyüközer Özkan H, Doğan Çankaya T, Kölüş T. The impact of language variability on artificial intelligence performance in regenerative endodontics. *Healthcare*. 2025;13(10):1190.
25. Aşık A, Kuru E. Diş hekimliğinde uzmanlık eğitim giriş sınavında sorulan çocuk diş hekimliği sorularına ChatGPT'nin verdiği cevapların analizi: Kesitsel araştırma. *Türkiye Klinikleri Diş Hekimliği Bilimleri Derg*. 2025;31(3):401-406.
26. Fang C, Wu Y, Fu W, et al. How does ChatGPT-4 preform on non-English national medical licensing examination? An evaluation in Chinese language. *PLOS Digit Health*. 2023;2(12):e0000397.