

Soruları Gözden Geçirmede Farklı Yapay Zekâ Uygulamalarının Karşılaştırılması: Doğru-Yanlış Önergeleri Üzerine Bir Uygulama*

Comparison of Different Artificial Intelligence Applications in Question Revision: An Application on True-False Items

Bedri Nejat Yüzüak¹, Funda Nalbantoğlu Yılmaz²

¹Arş. Gör., Nevşehir Hacı Bektaş Veli Üniversitesi, bnyuzuak@nevsehir.edu.tr,
(<https://orcid.org/0009-0006-8397-1030>)

²Sorumlu Yazar, Doç. Dr., Nevşehir Hacı Bektaş Veli Üniversitesi, fundan@nevsehir.edu.tr,
(<https://orcid.org/0000-0002-3228-4605>)

Geliş Tarihi: 13.08.2024

Kabul Tarihi: 12.01.2025

ÖZ

Araştırmada, doğru-yanlış önermesi yazmada dikkat edilecek noktaları oluşturma ve yazılan önermeleri ölçme ve değerlendirme ilkeleri açısından gözden geçirmede ChatGPT, Google Gemini ve WriteSonic yapay zekâ uygulamalarının kullanılabilirliğini incelemek amaçlanmaktadır. Araştırma, nitel araştırma yöntemlerinden durum çalışması olup 2023-2024 eğitim öğretim yılı bir devlet üniversitesinin farklı bölümlerde okuyan 3. sınıf öğrencilerinin “Eğitimde Ölçme ve Değerlendirme” dersi kapsamında hazırladığı doğru-yanlış önermelerinden seçilenlerle yürütülmüştür. Araştırma kapsamında belirlenen üç yapay zekâ uygulamasından doğru-yanlış önermeleri hazırlamada dikkat edilecek kuralları içeren kriterler oluşturma ve ele alınan önermeleri hem kendi oluşturduğu hem de uzmanlar tarafından verilen kriterlerle incelemesi istenmiştir. Bu doğrultuda elde edilen veriler içerik analizi ve betimsel analiz yöntemleriyle incelenmiştir. Sonuç olarak; üç yapay zekâ uygulaması da önermelerin anlaşılabilirliği, net cevap olması, tek bilgi ölçülmesi gibi yapıları değinirken olumsuz ifade kullanımı, doğrudan kaynaktan bilgi alma gibi özellikleri bazı uygulamaların göz ardı ettiği saptanmıştır. Uygulamaların ölçme değerlendirme ilkeleri bakımından kendilerine kriterler verildiğinde önermeler için daha uygun geri bildirimler verdiği, Türkçe-İngilizce incelemede dil farklılığının olduğu, uygulamaların İngilizce yazışmalar için daha uygun geri bildirimler verdiği sonuçlarına ulaşılmıştır.

Anahtar Kelimeler: Yapay zekâ, maddelerin gözden geçirilmesi, doğru-yanlış önermesi.

ABSTRACT

The aim of the study is to investigate the usability of ChatGPT, Google Gemini, and WriteSonic in the creation of points to be considered when writing true-false items and in the verification of the written items in terms of measurement and evaluation principles. The research was carried out the true-false items prepared by third-year students studying in different departments of a state university within the framework of the course "Measurement and Evaluation in Education" in the academic year 2023-2024. Three artificial intelligences identified within the framework of the research were asked to create criteria, including the rules to be taken into account when preparing true-false items, and to test the items with both the criteria – they created and those given by experts. The data obtained in this direction were

*Bu çalışma, 21–24 Mayıs 2024 tarihinde XI. International Eurasian Educational Research Kongresi'nde (EJER-2024) sunulan sözlü bildiriden üretilmiştir.

analyzed using content and descriptive analysis methods. As a result; it was determined that while all applications touched upon structures such as the understandability of propositions, having a clear answer, and measuring single information, some applications ignored negative expressions. It was concluded that the applications gave more appropriate feedback for items when they were given criteria in terms of assessment principles, that there was a language difference in Turkish-English analysis, and that the applications gave more appropriate feedback for English correspondence.

Keywords: Artificial intelligence, item revision, true-false item.

GİRİŞ

Dijital çağın hızlı bir şekilde ilerlemesiyle ortaya çıkan teknolojiler her alanda etkisini çok yönlü yansıtmaktadır. Teknolojik gelişmeler, eğitimde de etkisini göstermektedir. Söz konusu gelişmeler eğitimin teknolojiyle entegre olması zorunluluğunu doğurarak bireylerin gündelik hayat becerilerinin gelişimine katkı sağlamaktadır. Eğitim ve teknoloji entegrasyonunun günümüzde en önemli paydaşlarından birisi yapay zekâdır. Yapay zekâ; bir cihazın makine öğrenmesi aracılığıyla yeni durumlara uyum sağlama, problemlerle başa çıkabilme, çeşitli sorulara cevap verebilmesi gibi insan zekâsı temelli etkinlikleri statik ya da stabil şekilde gerçekleştiren sistem bütünüdür (Chen vd., 2020; McCarthy, 2004). Bu sistemlerin yüksek veri işleme hızına erişebilmesi, gelişmiş algoritmalara sahip olması, büyük veri analizi gerçekleştirebilmesi gibi işlevleri nedeniyle bireyler tercüme, veri analizi, resim oluşturma, sohbet etme gibi çeşitli amaçlarla yapay zekâyı gün geçtikçe daha fazla kullanmaktadır (Zeide, 2019). Belirtilen amaçlar doğrultusunda yaygın olarak kullanılan yapay zekâ uygulamalarına “Google Gemini”, “OpenAI ChatGPT”, “Writesonic”, “Microsoft Co-Pilot”, “DeepAI” örnek gösterilebilir (Koçyiğit & Darı, 2023).

Yapay zekânın dünya genelinde kullanımının artması sonucunda eğitim alanında da öğretmenlerin yapay zekâ kullanımına yönelik uygulamaların hayata geçirilmesinde bir artış gözlenmiştir (Çam vd., 2021). Özellikle eğitim müfredatının ana hatlarının belirlenmesi, ders planlarının ve programlarının tasarlanması, ölçme faaliyetlerine yönelik materyal hazırlanması ve değerlendirme süreçlerinin (öz değerlendirme, dönüt sağlanması vb.) gerçekleştirilmesine yönelik düşüncelerin ve uygulamaların artış gösterdiği belirlenmiştir (Montebello, 2018; Shah, 2023). Bu uygulamalarla birlikte bireyselleştirilmiş eğitim sistemlerinin daha güçlü bir yapı olmasıyla etkileşimli ve dinamik öğrenme süreçlerinin daha yaygın hale gelerek öğretmen öğrenci ilişkilerinin gelişmesi de beklenmektedir (Chatterjee & Bhattacharjee, 2020). Bu durum yapay zekâ uygulamalarının eğitimde kullanımına ve bu doğrultuda bir paradigma değişimine işaret etmektedir.

Eğitimin her alanında olduğu gibi ölçme ve değerlendirme alanında da yapay zekâ uygulamalarının etkisi, kullanılabilirliği ve sağladığı faydalar merak konusudur. İlgili alanda özellikle ÖSYM, uygulamalarında yapay zekâyı aktif şekilde kullanmayı hedeflemektedir. Yapay zekâ aracılığıyla ölçme ve değerlendirme alanına yönelik kullanım Shah (2023)’ün belirttiği üzere yaygınlaşmaktadır. Bu durumun temel sebebi ChatGPT, Google Gemini ve Writesonic gibi yapay zekâ uygulamalarının sahip olduğu etik kurallar çerçevesinde bireylerin değerlendirilmesi sürecinde adil ve etkili bir stratejinin sunulmasıdır (Gürlek vd., 2023; Montebello, 2018). Sunulan bu imkân dahilinde bireylerden elde edilen veriler yapay zekâ sistemleriyle analiz edilebilir bir yapıdadır. Böylece öğretmenlere ya da test uygulayıcılarına yönelik kullanışlılık ve puanlamada oluşabilecek ikilemleri engelleyecek nesnel, adil bir değerlendirme sistemi ile test güvenilirliğinin artması mümkündür.

Eğitimde ölçme faaliyeti kapsamında yapay zekâ ele alındığında yapay zekâ sistemlerinin uygulama alanı giderek yaygınlaşmakta ve bu duruma bağlı olarak yapay zekânın eğitimdeki rolü artmaktadır. Bu rollere test uygulayıcısının bireyleri değerlendirmek istediği yapıya yönelik ölçme uygulaması planlanması, ölçme faaliyetine yönelik kılavuz hazırlanması, geri bildirim

verme, ölçme aracı ve maddelerin tasarlanması, not verme, puanlama anahtarı oluşturma gibi örnekler verilebilir (Baidoo-Anu & Ansah, 2023; Çam vd., 2021; Popenici & Kerr, 2017; Shah, 2023; Taşçı & Çelebi, 2020). Bunlarla birlikte birçok açıdan sınıf içi ölçme ve değerlendirmelerde yapay zekâ uygulamaları eğitimcilere/öğretmenlere avantaj sağlayabilir. Örneğin, öğretmenlerin sınıf içi ölçmelerde öğrencilerin çalışmalarına geri bildirim verme süresini azaltabilir ya da soru üretme, doğru soru hazırlama noktasında öğretmene destek olabilir.

Günümüzde öğretmenler madde oluşturma konusunda yapay zekâdan faydalanabilmektedir. Bir öğretmen ya da test uygulayıcısının madde/soru hazırlamada üzerinde durması gereken önemli konulardan birisi de maddelerin gözden geçirilmesi/madde redaksiyonudur (Doğan, 2019). Madde redaksiyonu, öğretmen ya da bir test geliştiricisi tarafından oluşturulan bir maddenin dil, teknik kriterler, bilimsel uygunluk ve çeşitli yönlerden hem uzmanlar tarafından hem de istatistiksel açıdan incelenmesidir (Turgut & Baykul, 2012; Karakaya, 2022). Sınıf içi ölçme ve değerlendirme uygulamalarında özellikle başarı testi geliştirmede hem uzman görüşü hem de istatistiksel yöntemler kullanılarak maddeler üzerinde geliştirme çalışmaları yapılmaktadır. Bu süreç sonucunda oluşturulan maddelerdeki eksik ve hatalı yönler belirlenir ve madde düzeltmeleri gerçekleştirilir. Bu düzeltmelere örnek olarak; yaşa ve seviyeye uygunluk, açık-anlaşırlık, yazım ve dil kurallarına uygunluk, bilimsel açıdan doğruluk, ilgili madde/soru türündeki teknik kriterler açısından uygunluk gibi kurallar verilebilir (Baykul, 2000; Kan, 2014). Doğru madde yazma sürecinde önemli role sahip olan maddelerin gözden geçirilmesi sadece insan eliyle hazırlanan maddelerde değil aynı zamanda yapay zekâyla oluşturulmuş maddelerde de ele alınarak dilsel, bilimsel, alana uygunluk gibi açılardan incelenmesi gereken bir süreçtir.

Literatürde yapay zekâ uygulamalarının ölçme ve değerlendirme açısından, soru cevaplama performansı (Ali vd., 2023; Huang vd., 2024; Sallam & Al-Salahat, 2023; Soysal & Ardıç-Kemerkeya, 2024; Şahin vd., 2024), soru üretme/test geliştirme (Akçapınar vd., 2024; Bartoli vd., 2024; Bulut & Yıldırım-Erbaşlı, 2022; Ceylan & Gökçe, 2024; Indran vd., 2023; Kıyak vd., 2024; Liu vd., 2024; Nasution, 2023; Taşdelen-Teker vd., 2024; Wang vd., 2024), otomatik puanlama (Ercikan & McCaffrey, 2022; Gürbüz & Gürbüz, 2024) gibi durumlarda sıklıkla kullanıldığı görülmektedir. Ayrıca öğrencilere yönelik girdilerin takibi, öğrencilere uygun görevler sunma, sunulan görevlere yönelik etkili geri bildirim sağlama konularında yapay zekâ uygulamaları süreci destekleyebilmektedir (Boulay, 2016). Bu doğrultuda literatürde yapay zekânın ödev ve performans görevlerinde kullanımına (Koçak & Şimşek, 2024), psikomotor becerilerin ölçülmesinde ve/veya açık uçlu maddelerin puanlanmasında yapay zekâdan elde edilen sonuçların doğruluğuna dönük karşılaştırmalarda (Boztunç-Öztürk & Şahin, 2024; Demir, 2023; Kucam & Malakcıoğlu, 2024; Yiğiter & Boduroğlu, 2024; Vural, 2024) kullanıldığı görülmektedir. Ölçme ve değerlendirme alanı için ayrıca yapay zekâ uygulamalarını nitel/nicel veri analizi sürecinde kullanan araştırmalarda yer almaktadır (Tezci & Bıçak, 2024).

Seldon ve Abidoye (2018), eğitimde yapay zekâyâ yönelik doğan potansiyel doğrultusunda, yapay zekâyâ bağlı akıllı eğitim sistemi destekleneceği, madde hazırlama ve öğretmenlerin madde hazırlamadaki eksikliklerinin giderilebileceği belirlemiştir. Bu doğrultuda yapay zekâ uygulamalarının doğru soru hazırlama açısından eğitimciler/öğretmenler tarafından hazırlanan soruları gözden geçirmede kullanılabilirliği ve sağladığı faydalar ise merak konusudur. Sınıf içi ölçme uygulamalarında öğretmenler çoğu zaman hazırladıkları sorular için görüş alacak uzmanlara ulaşmamaktadır. Bu durumda hazırladıkları soruların gözden geçirilmesi/inceleme yapılamamaktadır. Bu açıdan yapay zekâ uygulamaları soruların doğruluğu ve kalitesi için ön inceleme görevini üstlenebilir mi?

Literatürde ölçme ve değerlendirme alanı için yapay zekâ uygulamaları sıklıkla soru üretmede kullanılmış olup öğretmenler tarafından hazırlanan soruları gözden geçirmede, hazırlanan sorulara geri bildirim vermedeki yapay zekâ uygulamalarının etkisini tartışan

çalışmalara rastlanmamıştır. Bu açıdan yapay zekâ uygulamalarının bir maddeyi gözden geçirme sürecinde hangi kriterleri göz önünde bulundurduğu, bir maddedeki ölçme ve değerlendirme ilkeleri bakımından olası hataları ne kadar belirleyebildiği ya da hangi yapay zekâ sisteminin bu konuda daha elverişli sonuç verdiği bilgisi tam olarak bilinmemektedir. Bu doğrultuda araştırmada, doğru-yanlış (D-Y) önermesi yazmada dikkat edilecek noktaları oluşturma ve yazılan önermeleri ölçme ve değerlendirme ilkeleri açısından gözden geçirmede yapay zekâ uygulamalarının kullanılabilirliğini incelemek amaçlanmaktadır. Bu amaç çerçevesinde ise araştırmada;

1. Doğru-yanlış önermelerini gözden geçirmede ChatGPT, Google Gemini ve Writesonic yapay zekâ uygulamaları hangi kriterleri kullanmaktadır?
2. ChatGPT, Google Gemini ve Writesonic yapay zekâ uygulamaları, öğretmen adaylarının hazırladığı doğru-yanlış önermelerini ölçme ve değerlendirme ilkeleri bakımından ne düzeyde gözden geçirmektedir?
3. Doğru-yanlış önermelerini gözden geçirmede ChatGPT, Google Gemini ve Writesonic yapay zekâ uygulamalarında Türkçe-İngilizce komut kullanma önermeler için belirlemeleri farklılaştırmakta mıdır? sorularına cevap aranmıştır.

YÖNTEM

2.1. Araştırma Deseni

Araştırma, nitel araştırma yöntemlerinden durum çalışmasıdır. Durum çalışmaları, yeni geliştirilmiş ya da uygulaması gerçekleştirilen bir veya birden çok sistemin işleme ve çalışma durumları hakkında veri toplayıp, çok yönlü incelenmesine dayanan araştırmalardır (Creswell, 2007; Chmiliar, 2010). Durum çalışmasında amaç, detaylı betimlemeler yapmak, durumu var olduğu biçimiyle anlamaktır (McMillan, 2004).

2.2. Araştırma Verileri

Araştırma, 2023-2024 eğitim öğretim yılı bahar döneminde, bir devlet üniversitesindeki Eğitim Fakültesi'nin farklı bölümlerinde okuyan 3. sınıf öğrencilerinden toplanmıştır. Bu doğrultuda "Eğitimde Ölçme ve Değerlendirme" dersi kapsamında öğrencilerin hazırladığı doğru-yanlış önermelerinden hatalı hazırlanan önermeler araştırma için seçilmiştir. Söz konusu ders kapsamında öncelikle doğru-yanlış önermesi hazırlama konusunda eğitim verildikten sonra öğretmen adaylarından alanları ile ilgili doğru-yanlış önermeleri hazırlamaları istenmiştir. Öğretmen adaylarının hazırladığı önermelerden alan, dil ve farklı hatalara sahip olma kriterleri dikkate alınarak iki adet doğru-yanlış önermesi araştırmada kullanılmak üzere belirlenmiştir. Seçilen önermelerin kritik ettirilmesinde ise üç farklı yapay zekâ uygulaması (ChatGPT 3.5, Google Gemini ve Writesonic) kullanılmıştır. Araştırmada kullanılan yapay zekâ uygulamalarının belirlenmesinde bilinen ve yaygın kullanım kriterlerine dikkat edilmiştir. Araştırmada yapay zekâ uygulamalarının belirlemeleri "Eğitimde Ölçme ve Değerlendirme" alanından bir doçent ve bir araştırma görevlisi tarafından incelenmiş ve kritik edilmiştir.

2.3. Verilerin Toplanması

Araştırmada öncelikle belirlenen yapay zekâ uygulamalarından doğru-yanlış önermesi hazırlamada dikkat edilecek noktaları içeren kriterler oluşturmaları istenmiştir. Yapay zekâ uygulamalarıyla gerçekleştirilen kritik etme süreci ise aşağıdaki gibi gerçekleştirilmiştir.

ChatGPT, Google Gemini ve Writesonic adlı yapay zekâ uygulamalarına ilk olarak "Merhaba Gemini/ ChatGPT/ Writesonic. Bugün seninle *"Doğru-Yanlış önermeleri"* üzerine akademik bir sohbet gerçekleştireceğiz. Benimle paylaşacağın bilgileri akademik kaynaklardan alırsan sevinirim." ifadesi sunularak ele alınacak alanın tanımlaması ve yapay zekâdan istenen cevabın kaynak tercihinin sağlanması hedeflenmiştir. Bu doğrultuda yapay zekâ

uygulamalarının doğru-yanlış önermelerine yönelik tanımlama yapmaları istenmiş olup konu hakkındaki genel bilgi çerçevesi kontrol edilmiştir. Genel konu başlığının tanımlanmasının ardından doğru-yanlış önermesi hazırlanması sürecinde dikkat edilecek noktaları oluşturmak için “*Öğretmenler için doğru-yanlış önermesi/sorusu hazırlamada dikkat edilecek noktaları içeren kriterler oluşturur musun? Bu kriterler doğru-yanlış önermesi/sorusu yazarken uymaları gereken kuralları içermelidir. Yazdığın her bir kriter için örnek verebilir misin?*” yönergesi sunulmuştur. Alınan cevaplardan sonra “*Öğretmenler doğru-yanlış önermesi yazarken başka nelere dikkat etmelidir?*” ile konuşma devam ettirilmiştir. Araştırmada daha sonra öğretmen adaylarının hazırladıkları doğru-yanlış önermelerinden seçilenler ChatGPT, Google Gemini ile Writesonic yapay zekâ uygulamalarına tek tek sunulmuş ve önce kendi kriterleri ile sonra uzmanların verdiği kriterlere göre kritik ettirilmiştir. Yapay zekâ uygulamalarının ilgili madde türünde oluşturduğu kriterler ve önermelere ait incelemeleri ilgili madde türündeki ölçme ve değerlendirme ilkeleri bakımından karşılaştırılıp tartışılmıştır. ChatGPT, Google Gemini ve Writesonic ile yapılan tüm yazışmalarda aynı ifadeler kullanılmıştır. Her üç yapay zekâ uygulamasına aynı doğru-yanlış önermeleri incelettirilmiştir (Mart-Nisan, 2024). Üç yapay zekâ uygulaması ile kriter oluşturma ve belirlenen önermelere ait gözden geçirme süreci araştırmacılar tarafından ayrı ayrı ve farklı zaman dilimlerinde tekrarlanmış, tartışılarak sonuçlandırılmıştır.

2.4. Verilerin Analizi

Araştırmada araştırma sorularına göre içerik analizi ve betimsel analiz kullanılmıştır. İçerik analizi, herhangi bir metnin, dokümanın, temanın, belgenin nesnel, ölçülebilir ve doğrulanabilir şekilde sistematik olarak incelenmesidir (Krippendorff, 2004). Betimsel analiz, doğrudan alıntılara yer vererek verinin sistematik ve açık bir biçimde betimlenmesi ve yorumlanmasıdır (Strauss & Corbin, 1990). Veri analiz sürecinde araştırma sorularına göre bazen veriler kodlanarak temalaştırılmış, bazen de alıntılara yer verilerek doğrudan yorumlama ile anlam ortaya çıkartılmıştır.

2.5. Güvenirlik ve Geçerlik

Araştırmada geçerli ve güvenilir sonuçlara ulaşabilmek için yapay zekâ uygulamalarını kullanarak D-Y önermelerini gözden geçirmede kriter oluşturma ve seçilen D-Y önermelerini gözden geçirme süreci araştırmacılar tarafından ayrı ayrı, aynı ifadelerle (prompt) birden fazla kez yürütülmüştür. Araştırmacıların her biri elde ettikleri bilgileri çözümlemiş ve çıkan sonuçlar birlikte kritik edilerek karşılaştırılmış, raporlanmıştır. Böylelikle araştırmacılar arası tutarlık ve nesnel geçerlik sağlanmıştır. Ayrıca yapay zekâ uygulamalarına sorulan sorularla ilgili görüşlerini detaylandırma fırsatı tanınmış ve araştırmada doğrudan alıntılara yer verilerek geçerlik artırılmaya çalışılmıştır.

BULGULAR

Araştırma kapsamında elde edilen bulgular bu bölüm kapsamında ele alınmıştır. Bu doğrultuda yapay zekâ uygulamalarının doğru-yanlış önermesi hazırlamaya yönelik oluşturduğu kriterler temalaştırılarak Tablo 1’de karşılaştırmalı olarak verilmiştir.

Tablo 1

Yapay Zekâ Uygulamaları Tarafından D-Y Önergelerini Gözden Geçirmede Kullanılan Kriterler

Kriter	Bulunduğu Yapay Zekâ
Öğrenme Hedefine/Kapsama Uygunluk	ChatGPT, Google Gemini, Writesonic
Öğrenci Düzeyine Uygunluk	ChatGPT, Google Gemini, Writesonic
Tek Konuya/Fikre Odaklanma	ChatGPT, Google Gemini, Writesonic
Açık ve Anlaşılabilirlik	ChatGPT, Google Gemini, Writesonic
Netlik ve Kesinlik	ChatGPT, Google Gemini, Writesonic
D/Y Seçenek Dengesi	ChatGPT, Google Gemini, Writesonic
Yanlışlığın Önemli Olması	ChatGPT, Google Gemini, Writesonic
Dilbilgisi ve Yazım Kuralları	ChatGPT, Google Gemini
Kişisel Yargılardan-Görüşlerden Uzaklık	ChatGPT, Google Gemini
Görsel Materyal Desteği	ChatGPT, Google Gemini
Bilimsel Doğruluk ve Güncellik	ChatGPT, Writesonic
Negatiflik	ChatGPT
Kaynak Kullanımı (aynen alınma- kopyalama)	ChatGPT

Not: Kriterler birden fazla kez deneme ile oluşturulmuştur.

Tablo 1’de verilen kriterlere göre her üç yapay zekâ uygulamasının hazırlanan önermelerin öğrenci düzeyine, ders hedef ve kapsamına uygun olması, önermelerin açık-anlaşılır olması, önermelerde net-kesin bir cevabın olması, önermedeki yanlışlığın önemli bir noktada olması, her bir önermenin tek bir bilgiyi/fikri ölçmesi, doğru-yanlış cevap dengesi gibi kriterleri ele aldığı belirlenmiştir.

Tablo 1’de verildiği gibi görsel materyal desteği, dilbilgisi ve yazım kuralları, önermelerin kişisel yargı-görüşlerden uzak olması gibi kriterlere ChatGPT ve Google Gemini değinmiştir. Bu kriterlere ek ChatGPT ve Writesonic bilimsel doğruluk ve güncellik gibi kriterler de belirlemiştir.

Doğru-yanlış önermesi yazarken “*değildir, yoktur vb.*” olumsuz ifadelerden kaçınılmalıdır. Bu açıdan yapay zekâ uygulamalarının oluşturduğu kriterler incelendiğinde bir tek ChatGPT olumsuz ifade kullanılmamasına değinmiştir. Ayrıca önerme ifadelerini bir kaynaktan olduğu gibi almama kriterini bir tek ChatGPT ele almıştır.

Araştırmanın ikinci kısmında öğretmen adaylarının ölçme ve değerlendirme dersi kapsamında hazırladığı ve hataları olan doğru-yanlış önermelerinden seçilenler üç yapay zekâ uygulamasına kendi oluşturdukları kriterlere göre incelettirilmiştir. İncelemelerde araştırmacılar tarafından birden fazla kez deneme yapılmıştır. Bu inceleme sonuçları, uygulamaların kendi ifadeleriyle sadeleştirilerek Tablo 2 ve Tablo 3’te her bir önerme için ayrı ayrı verilmiştir.

Tablo 2*ChatGPT, Google Gemini ve Writesonic ile Önergelerin Gözden Geçirilmesi Örnek 1***Doğru-Yanlış Önermesi: *It is not unlikely for ice not to melt at room temperature.***

ChatGPT	- Önerme karmaşık bir yapıya sahiptir ve net değildir. Önerme, belirsiz bir ifade içermektedir ve kesin bir doğruluk ifadesi değildir. İfade açık değildir. - İfade, çift olumsuzluk ve karmaşık bir cümle yapısı içermektedir. - İfade orijinal gibi görünüyor, başka kaynaklardan kopyalanmış bir ifade değil. İfade, genel bilgiye dayanmakta ve güncel görünmektedir - Önerme belirli bir konuyu ifade etmektedir, ancak karmaşık yapısı nedeniyle net bir konu belirtmemektedir. - İfade, nesnel bir ifade gibi görünmektedir, kişisel bir yargı içermemektedir. - Önerme, bilimsel bir gerçeği ifade etmiyor gibi görünmektedir, daha çok olasılıklar üzerine spekülasyon yapıyor gibi. - İfade karmaşık bir yapıya sahip olduğu için öğrenci düzeyine uygun değildir. - Önerme, sadece sözel ifadelerle belirtilmiştir, görsel destek içermez. <i>Sonuç olarak, önerme belirtilen kurallara uygun değildir. Karmaşık bir cümle yapısıyla ifade edilmiş, netlikten uzak ve belirsiz bir ifade içermektedir. Bu nedenle, iletişimde doğruluk ve etkinlik sağlamakta zorluk çekebilir.</i>
Google Gemini	- Önerme net ve anlaşılır bir şekilde ifade edilmemiştir. "Not unlikely" ifadesi belirsizdir. - Önerme tek bir fikre odaklanmamaktadır. "Not unlikely" ifadesi, buzun erime olasılığının düşük olduğunu, ancak imkânsız olmadığını gösterir. Bu, iki farklı fikri (buz erimez ve buz erime olasılığı düşüktür) tek bir önermede birleştirir. - Önerme yanlıştır. Buz, atmosferik basınç altında 0°C'nin üzerindeki herhangi bir sıcaklıkta erir. Oda sıcaklığı (genellikle 20-25°C arasında) buz erime noktasının çok üzerindedir. Önerme şu şekilde düzeltilebilir: "Buz, oda sıcaklığında erir." - Önerme dilbilgisi ve yazım kurallarına uygun değildir. "Not unlikely" ifadesi yerine "unlikely" veya "improbable" kullanılabilir. - Önerme şu şekilde düzeltilebilir: "Buz erimesi olasılıksızdır." - Önerme, doğru önermelere benzemiyor gibi görünse de yanlış bilgi içermektedir. Önerme şu şekilde düzeltilebilir: "Buz, oda sıcaklığında erir." - Önerme, fizik veya kimya gibi buz erime noktasına ilişkin bilgi gerektiren bir alana aittir. Bu alana hâkim olmayan bir kişi tarafından yazılmış olması muhtemeldir. - Önerme, bir fizik veya kimya uzmanı tarafından gözden geçirilmeli ve düzeltilmelidir. <i>"It is not unlikely for ice not to melt at room temperature" önermesi, yukarıda listelenen tüm kriterleri ihlal etmektedir. Önerme yanlış bilgi içermekte, belirsiz ve dilbilgisi kurallarına uygun değildir. Önermenin daha net, doğru ve dilbilgisi kurallarına uygun hale getirilmesi gerekir.</i> Düzeltilmiş Önerme: "Buz, oda sıcaklığında hızla erir."
Writesonic	- Önerme karmaşık bir yapıya sahip olduğu için netlik ve anlaşılabilirlik açısından zayıf bir durumdadır. - Bilimsel olarak doğru olmayan bir ifade içermektedir. Genel olarak buz oda sıcaklığında erir, bu ifade ise bu bilgiye aykırıdır. - Önerme, öğrenci seviyesine uygun bir dil ve karmaşıklık düzeyine sahip değildir. - Önerme, genel bilimsel bilgilere aykırı bir ifade içermektedir. - Önerme, ikinci anlamlılığa yer veren bir ifade içermemektedir. - Dilin doğru kullanıldığı söylenebilir ancak anlam karmaşasına yol açabilecek bir ifade içermektedir. <i>Sonuç olarak, "it is not unlikely for ice not to melt at room temperature" önermesi, bilimsel gerçeklere aykırı bir ifade içerdiği için uygun bir doğru-yanlış önermesi olmamaktadır. Bu önerme, öğrencilerin kafasını karıştırabilecek veya yanlış bilgi edinmelerine sebep olabilecek niteliktedir.</i>

Tablo 2’de ilk önerme için ChatGPT tarafından yapılan gözden geçirmeler incelendiğinde, ilgili önermenin açık-anlaşılır olmadığı, karmaşık bir yapıya sahip olduğu, belirsiz ifade içerdiği ve çifte olumsuzluk içeren cümle yapısında olduğu gibi genel olarak sorudaki ölçme ve değerlendirme uzmanı bakış açısıyla gerekli geri bildirimleri verdiği belirlenmiştir. Bunun yanı sıra önermenin başka kaynaklardan kopyalanıp kopyalanmadığı,

karmaşık ifade yapısı gereği öğrenci düzeyine uygun olmadığı gibi incelemeleri de bulunmaktadır.

Aynı önerme için Google Gemini tarafından alınan geri bildirimler incelendiğinde, önermenin *açık-anlaşılır olmadığı, belirsizlik içeren ifadenin bulunduğu (not unlikely)* gibi önerme için doğru geri bildirimleri bulunmaktadır. Bu geri bildirimler haricinde önerme için belirtilmesi gereken olumsuz ifade kullanımı ile ilgili açıklama verilmemiştir. Fakat başka bir maddede yanlış bir bilgi sunmamak adına önermeyi *"Buz, oda sıcaklığında erir."* biçiminde olumsuzlukları gidererek düzeltmiştir. Diğer uygulamalardan farklı olarak ise Google Gemini oluşturduğu kriterlere göre önermeyi inceledikten sonra "düzeltilmiş önerme" önerileri de vermektedir.

Writesonic aynı önerme için *karmaşık bir yapıya sahip olduğu, netlik ve anlaşılabilirlik açısından zayıf olduğu, anlam karmaşasına sebep olacak ifade içerdiği* geri bildirimlerini vermiştir. Writesonic ayrıca önermedeki bilginin bilimsel olarak yanlışlığı üzerinde de durmuştur.

Her üç yapay zekâ uygulaması da sunulan önermeyi kendi kriterleri ile gözden geçirdiğinde önermeyi kullanmayı genel olarak uygun bulmamaktadır. Önermedeki önemli yanlışlıklardan biri olan birden fazla olumsuzluk ifadesinin kullanılması hatasını en açık biçimde ChatGPT ele almıştır. Diğer uygulamalar doğru-yanlış önermesinde olumsuz ifade kullanmamak gerektiği ile ilgili açık bir ifade kullanmamıştır. Üç inceleme sonucunda en uygun ve gerekli açıklamayı ChatGPT'nin yaptığı söylenebilir.

Tablo 3'te ikinci önerme için yapay zekâ uygulamalarının kendi kriterlerine göre incelemeleri verilmiştir. Tablo 3'te ChatGPT tarafından yapılan gözden geçirmeler incelendiğinde *ilgili önermenin net ve basit bir ifadeyle sunulduğu, karmaşıklık içermediği, dil bilgisi ve yazım kurallarına uygun olduğu, çift anlamlılık içermediği, ders müfredatına ve öğrenci seviyesine uygun olduğu, önyargı/belirsizlik ifadeleri içermediği, mantıklı ve anlamlı olduğu, yanıltıcı olmadığı* gibi genel olarak sorudaki ölçme ve değerlendirme uzmanı bakış açısıyla gerekli geri bildirimleri verdiği belirlenmiştir. Bunların haricinde önermenin tek bir fikre/bilgiye odaklanması bakımından ChatGPT herhangi bir geri bildirim vermemiş, önermedeki bu hatayı belirleyememiştir. Birçok açıdan doğru bir geri bildirim verse de bu önemli ayrıntı belirlenememiştir. ChatGPT önerme ile ilgili sonuç cümlesinde ise ilgili önermeyi "*uygun*" olarak nitelendirmiştir.

Tablo 3*ChatGPT, Google Gemini ve Writesonic ile Önermelerin Gözden Geçirilmesi Örnek 2***Doğru-Yanlış Önermesi: Dünya haritasında sular kahverengi, ormanlar mavi, dağlar yeşil renkle gösterilir.**

ChatGPT	• Önerme oldukça net ve basit bir ifadeyle sunulmuştur. Önerme karmaşıklık içermemektedir. Somut dil kullanılmış. • Önerme, harita renklerinin genel uygulamasına dayandığı için yanıltıcı değildir. • Harita okuma becerilerini geliştirmek için temel bir önerme olduğu için öğrenci düzeyine uygun olduğunu söyleyebiliriz. • Dilbilgisi ve yazım kurallarına uygun bir şekilde yazılmıştır. • Dilbilgisel hata ve çift anlamlılık yoktur. • Coğrafya dersi müfredatında yer alan dünya haritaları konusuyla doğrudan ilgilidir. • Önerme herhangi bir önyargı ve belirsizlik içermiyor. • Önerme bağlam içinde mantıklı ve anlamlıdır. • Bu ifade öğrencilerin genellikle duyduğu bir ifade olduğu için gerçeklik sorunu yok. • Bu ifade, harita okuma becerisine dayandığı için ilgili bir kaynaktan alıntı yapılmıştır.
Google Gemini	<i>Sonuç olarak, bu önerme, hazırlanan kriterlere uygun bir şekilde oluşturulmuş gibi görünmektedir. Bu tür önerme, öğrencilerin harita okuma becerilerini geliştirmelerine yardımcı olabilir.</i> • İfade, açık ve anlaşılır bir dil ile yazılmıştır. Önerme, kolayca uygulanabilir ve çok uzun veya karmaşık değildir. • Önerme, tek bir ana fikri test etmektedir. Dünya haritasında renklerin kullanımı. • Önerme, müfredat kapsamında yer alan önemli bir bilgiyi test etmektedir. • Önermede herhangi bir dilbilgisi veya noktalama hatası bulunmamaktadır. • Önermede herhangi bir mantıksal hata bulunmamaktadır. • Öğrenciyi yanıltıcı veya çelişkili bilgiler sunulmamaktadır. • Önermede görsel materyal kullanımı gerekmemektedir. • Önerme, doğru ve yanlış ifadeleri dengeli bir şekilde sunmaktadır. • Önerme, ortaokul öğrencileri için uygun olabilir. • Herhangi bir kültürel gruba veya inanca karşı ayrımcılık içermez. • Önerme, dürüstlük ve adil değerlendirme gibi etik ilkelerle uyumludur. • Önermede tahmin kolaylaştırıcı ipucu kullanılmamıştır. <i>Önerme doğru-yanlış test maddesi hazırlama kriterlerine uygun bir şekilde hazırlanmıştır. Önerme açık, net, dengeli ve mantıklıdır. Dilbilgisi ve noktalama kurallarına uygundur. Ayrıca müfredat kapsamında yer alan önemli bir bilgiyi test etmektedir.</i>
Writesonic	Açık, anlaşılır ve basit bir dil kullanılmıştır. • Önerme net bir ifade kullanmaktadır, ancak belirtilen renklerin gerçek dünya haritasıyla uyumadığı bilinmektedir. Sular genellikle mavi, ormanlar yeşil ve dağlar kahverengi renkle gösterilir. • Önermenin uzunluğu uygundur. • Önerme öğrencilerin sorgulama yapabilecekleri bir ifade içermemektedir. Bu nedenle yanlışlanabilirlik kriterine uymamaktadır. • Önerme öğrencilerin bilgi seviyesine uygun bir konuyu ele almakta, ancak renklerin belirtilmesi gerçeklikle örtüşmediği için yanlış bir ifade içermektedir. • Önerme tek bir konuya odaklanmış gibi görünmektedir, ancak verilen bilgilerin doğruluğu konusunda sorun bulunmaktadır. <i>Sonuç olarak, önerme bazı kriterlere uysa da genel olarak doğru bir ifade değildir. Gerçek dünya haritasında suların genellikle mavi, ormanların yeşil ve dağların kahverengi renkle gösterildiği bilinmektedir. Bu nedenle önerme, bilinen gerçeklikle örtüşmediği için doğru bir ifade değildir.</i>

Tablo 3'te aynı önerme için Google Gemini tarafından alınan geri bildirimler incelendiğinde, önermenin açık-anlaşılır olduğu, karmaşık olmadığı, dil bilgisi ve noktalama hatası olmadığı, müfredat kapsamında yer alan bir bilgiyi yokladığı, mantıksal hata, yanıltıcı bilgi, çelişkili ifade, tahmini kolaylaştırıcı ipucu, ayrımcılık içermediği gibi önerme için doğru geri bildirimleri bulunmaktadır. Ayrıca önerme için doğrudan sınıf düzeyi belirtilmediğinden önermeyi ortaokul öğrencileri için uygun, lise öğrencileri için çok kolay, ilkökul öğrencileri için ise zor düzeyde bulmuştur. Bunlar haricinde görsel materyal desteği gibi bir kriter

bulunmaktadır. Önermede *görsel materyal kullanımı gerekmediğini belirtmiştir*. “*Önerme, tek bir ana fikri test etmektedir. Dünya haritasında renklerin kullanımı.*” geri bildirimi ile Google Gemini bu kritere değinmiş fakat uygun bir değerlendirme yapamamıştır. Önerme birden fazla yargının/bilginin doğru-yanlışlığını içermektedir. Bununla birlikte sonuç cümlesinde ise ilgili önemeyi “*hazırlama kriterlerine uygun*” olarak nitelendirmiştir.

Writesonic aynı önerme için *açık-anlaşılır, basit, net bir dil kullanıldığı, önerme uzunluğunun uygunluğu, bilgi seviyesinde bir önerme olduğu* gibi doğru geri bildirimler vermiştir. Önermedeki tek bir fikre/bilgiye odaklanmaması bakımından bulunan hatayı belirleyememiştir. Bu durum ile ilgili “*Önerme tek bir konuya odaklanmış gibi görünmektedir, ancak verilen bilgilerin doğruluğu konusunda sorun bulunmaktadır.*” ifadesini kullanmıştır. Ayrıca Writesonic önermede genel olarak ifadedeki bilginin bilimsel olarak yanlışlığı üzerine durmuştur. Örneğin; “*renklerin belirtilmesi gerçeklikle örtüşmediği için yanlış bir ifade içermektedir.*”, “*belirtilen renklerin gerçek dünya haritasıyla uyummadığı bilinmektedir. Sular genellikle mavi, ormanlar yeşil ve dağlar kahverengi renkle gösterilir.*” vb. Writesonic önerme ile ilgili sonuç cümlesinde ise “*verilen önerme bazı kriterlere uysa da genel olarak doğru bir ifade değildir.*” ifadesini kullanmıştır. Bunun gerekçesi olarak “*önerme, bilinen gerçeklikle örtüşmediği için doğru bir ifade değildir.*” açıklamasını yapmıştır.

ChatGPT ve Google Gemini sunulan önermeyi kendi kriterleri ile gözden geçirdiğinde önermeyi kullanmayı uygun bulmaktadır. Bir tek Writesonic önermeyi *bilinen gerçeklerle örtüşmediği için* uygun bulmamıştır. Önermede temel bir problem bulunmaktadır. Birden fazla fikir tek bir önerme içerisinde sunulmuştur. Önermede sular, ormanlar ve dağların ayrı ayrı ele alınması beklenen bir özelliktir. Bunu üç yapay zekâ uygulaması da belirleyememiştir. Konuyla ilgili Google Gemini: “*Önerme, tek bir ana fikri test etmektedir. Writesonic: “Önerme tek bir konuya odaklanmış gibi görünmektedir.”* şeklinde geri bildirimler bulunmaktadır. Üç inceleme sonucunda ChatGPT ve Google Gemini’nin benzer bildirimler verdiği, Google Gemini’nin düzeltme örnek ifadelerine de yer verdiği, Writesonic diğerlerine göre daha kısa ifadelerle, yüzeysel geri bildirimler sunduğu belirlenmiştir.

Araştırmanın üçüncü kısmında ise yapay zekâ uygulamalarına ölçme ve değerlendirme ilkeleri bakımından D-Y önermelerini hazırlamada dikkat edilecek noktaları içeren araştırmacılar tarafından hazırlanmış kriterler sunulmuş ve araştırma için seçilen önermeler tekrar incelettirilmiştir. Bu inceleme sonuçlarından örnekler ise Tablo 4 ve Tablo 5’te verilmiştir.

Tablo 4’te D-Y önermelerini araştırmacılar tarafından hazırlanmış kriterlere göre incelemeleri istendiğinde önermenin açık-anlaşılabilirliği ile ilgili her üç uygulamada Tablo 2 ile uyumlu olarak doğru belirlemeler yapmıştır.

Tablo 4’e göre uygulamalar kendi kriterlerine göre daha önceki incelemelerinde göremedikleri olası bazı eksiklik ve hataları bu kez belirleyebilmiştir. Örneğin, önermede belirsizlik ifadesinin yer almasıyla ilgili kendi oluşturdukları kriterlerle Tablo 2’de verilen incelemede ChatGPT ve Google Gemini bu hatayı belirlemiştir. Writesonic ise bu konuda doğrudan bir inceleme yapmasa da anlam karmaşasına yol açabilecek ifade bulunduğunu belirtmiştir. Araştırmacıların oluşturdukları kriterlere göre yapılan incelemede ise önermedeki bu eksikliği her üç uygulamada belirlemiştir. Ayrıca önermede olumsuz ifadenin kullanımı, birden fazla olumsuz ifadenin yer alması ile ilgili Tablo 2’de bir tek ChatGPT bu duruma değinmiştir. Uygulamalardan kendilerine sunulan kriterlere göre inceleme yapmaları istendiğinde ise Tablo 2’den farklı olarak her üç uygulamada önermenin olumsuz ifade içerdiğini belirleyebilmiştir. Fakat birden fazla olumsuz ifadenin yer almasıyla ilgili ChatGPT doğru belirleme yapsa da Google Gemini ve Writesonic ile “*önerme çifte olumsuzluk içermemektedir.*” açıklamasıyla ilgili hatayı yine tespit edememiştir.

Tablo 4

ChatGPT, Google Gemini ve Writesonic ile Sunulan Kriterlere Göre Önermelerin Gözden Geçirilmesi 1

Doğru-Yanlış Önermesi: It is not unlikely for ice not to melt at room temperature.

ChatGPT	• Önermeler kesinlikle doğru ya da kesinlikle yanlıştır: Belirsiz. İfade, olasılıklar ve olasılıklar hakkında konuştuğu için kesin bir ifade değildir. • Önermeler açık-anlaşılardır: Kısmen doğru. İfade, karmaşık bir yapıya sahiptir ve tam olarak net bir şekilde anlaşılmayabilir. • Önermeler kesinlik ifade etmemektedir. • Her bir önerme tek bir bilginin/fikrin doğru ya da yanlışlığını ölçmektedir: Belirsiz. İfade karmaşık bir yapıya sahiptir ve tek bir fikri ölçmek yerine birden fazla fikri içerebilir. • Önermedeki yanlışlık önemli bir noktadır. • Önerme kişisel bir görüş içermez, ancak belirsizlik ifadeleri kullanılmıştır. • İfade, olumsuz ifadeler içermektedir. • İfade, çifte olumsuzluk içerir. • İfade karmaşık bir yapıya sahiptir ve bazıları için kafa karıştırıcı olabilir. • İfade karmaşık bir yapıya sahip olduğu için öğrenci düzeyine uygun olmayabilir. <i>Önerme, birkaç kriterle uyumlu olsa da genel olarak karmaşık ve belirsiz bir ifade içermektedir. Bu nedenle, öğrencilerin yanıtılmasına veya yanlış bir anlayış geliştirmelerine neden olabilir.</i>
Google Gemini	• Önerme net ve anlaşılır bir şekilde ifade edilmemiştir. "Not unlikely" ifadesi belirsizdir ve tam olarak neyin kastedilip kastedilmediğini anlamak zordur. • Doğru-yanlış önermeleri kesin olmalı. • Önerme daha açık ve net bir şekilde şu şekilde ifade edilebilir: "Buz, oda sıcaklığında erimez." • Kesinlik içeren ifadeler kullanmak, önermeyi daha net ve kesin hale getirebilir. • Önerme tek bir bilginin/fikrin doğru ya da yanlışlığını ölçmektedir. • Önerme cevabı yanlış ise önermedeki yanlışlık önemli bir noktadır. Buz, atmosferik basınç altında 0°C'nin üzerindeki herhangi bir sıcaklıkta erir. Bu nedenle, önermenin yanlış olması önemli bir bilgi hatasıdır. • Önerme, kişisel görüş veya yargı içermemektedir. • Önerme olumsuz bir ifade ("not unlikely") kullanmaktadır. • Olumsuz ifadeler dikkatli kullanılmalı ve mümkünse olumlu ifadelerle değiştirilmelidir. • Önerme çifte olumsuzluk içermemektedir. • Önerme, ortaokul veya lise düzeyindeki öğrenciler için uygun olabilir. <i>Önerme bazı kriterleri karşılarsa da, netlik, kesinlik ve öğrenme çıktılarıyla tam uyumluluk gibi önemli kriterleri karşılamamaktadır. Önerme, daha net, kesin ve öğrenme çıktılarıyla daha uyumlu hale getirilerek daha iyi bir doğru-yanlış önermesi haline getirilebilir. Düzeltilmiş Önerme: "Buz, oda sıcaklığında hızla erir."</i>
Writesonic	• Önerme kesinlik ifade etmemektedir. • Önerme karmaşık bir yapıya sahip olduğu için açık-anlaşılır değildir. • Önerme, kesinlik ifade etmeyen bir ifade içermektedir. • Önerme, tek bir bilgi/fikir üzerine odaklanmaktadır. • Verilen önerme, bilimsel gerçeklere aykırı olduğu için yanlışlığı önemli bir noktadır. • Verilen önerme, kişisel görüşler içermemektedir. • Önerme olumsuz bir ifade içermektedir. • Önerme çifte olumsuzluk içermemektedir. • Verilen önerme, karmaşık bir yapıya sahip olduğu için kafa karıştırıcı olabilir. • Önerme, öğrenci düzeyine uygun bir dil ve karmaşıklık düzeyine sahip değildir. <i>Önerme kriterlere göre incelendiğinde, önermenin birçok kriteri karşılamadığı ve bilimsel gerçeklere aykırı olduğu görülmektedir. Bu nedenle, genel olarak uygun bir doğru-yanlış önermesi değildir.</i>

Tablo 4'ten elde edilen bulgular doğrultusunda her üç yapay zekâ uygulaması kendilerine sunulan kriterlere göre yaptığı incelemede genel olarak önerme için doğru belirlemeler yapmıştır. Her üç yapay zekâ uygulaması da son genel yorum cümlesinde önermeyi kullanmayı uygun bulmamaktadır.

“Dünya haritasında sular kahverengi, ormanlar mavi, dağlar yeşil renkle gösterilir” önermesini araştırmacıların hazırladığı kriterler ile uygulamalar tarafından yapılan inceleme sonuçları Tablo 5’te verilmiştir.

Tablo 5

ChatGPT, Google Gemini Ve Writesonic ile Sunulan Kriterlere Göre Önermelerin Gözden Geçirilmesi 2

Doğru-Yanlış Önermesi: Dünya haritasında sular kahverengi, ormanlar mavi, dağlar yeşil renkle gösterilir.	
ChatGPT	• Önerme kesin doğru veya yanlış ifadesi içerir. • Önerme, dünya haritasının renklerini yanlış tanımlayarak açık ve anlaşılır değildir. • Önerme kesin bir ifade kullanmaktadır. • *Her bir önerme tek bir bilginin/fikrin doğru ya da yanlışlığını ölçmektedir: Doğru. Bu önerme, dünya haritasının belirli unsurlarının renklerini doğru bir şekilde ifade edip etmediğini ölçer. • Önermede dünya haritasının renkleri hakkında yanlış bir bilgi verilmiştir, dolayısıyla yanlışlık önemlidir. • Önerme, coğrafya konusundaki nesnel gerçeklere dayanır ve kişisel görüşleri içermez. • Önermede olumsuz bir ifade bulunmamaktadır. Önerme çifte olumsuzluk içermez. • Kafa karıştırıcı ifadelerden kaçınılmıştır. • Öğrencilerin düzeyine uygun değildir çünkü dünya haritasının renklerini yanlış bir şekilde tanımlar. <i>Önerme, birkaç kriterle uyumlu olsa da genel olarak karmaşık ve belirsiz bir ifade içermektedir. Bu nedenle, öğrencilerin yanıtılmasına veya yanlış bir anlayış geliştirmelerine neden olabilir.</i>
Google Gemini	• Önerme, kesinlikle yanlıştır. Dünya haritalarında sular genellikle mavi, ormanlar yeşil ve dağlar kahverengi renkte gösterilir. • Önerme, açık ve anlaşılır bir şekilde ifade edilmiştir. Kullanılan kelimeler bilindik ve yoruma açık değildir. • Önerme, kesinlik ifade etmektedir. • Önerme, tek bir bilginin doğru veya yanlışlığını ölçmektedir. Bu bilgi, dünya haritalarında kullanılan renklerin coğrafi unsurları temsil etmesidir. • Önerme cevabı yanlış ise, bu önemli bir noktadır. Dünya haritalarında kullanılan renklerin doğru şekilde öğrenilmesi, coğrafi bilgi edinimi açısından önemlidir. • Önerme, kişisel yargı içermemektedir. Coğrafi unsurların renklerle temsili, genel bir kabul görmüş bir bilgidir. • Önerme, olumsuz bir ifade içermemektedir. Önerme, çifte olumsuzluk içermemektedir. • Önerme, kafa karıştırıcı bir ifade içermemektedir. Kullanılan kelimeler açık ve nettir. • Önerme, ortaokul ve lise düzeyindeki öğrenciler için uygundur. <i>Önerme bazı kriterleri karşılarsa da netlik, kesinlik ve öğrenme çıktılarıyla tam uyumluluk gibi önemli kriterleri karşılamamaktadır. Önerme, daha net, kesin ve öğrenme çıktılarıyla daha uyumlu hale getirilerek daha iyi bir doğru-yanlış önermesi haline getirilebilir.</i>
Writesonic	• Dünya haritasındaki gösterim standartları ülkeden ülkeye değişebilir, bu nedenle önerme kesin doğru ya da yanlış ifadesi içermez. • Önerme oldukça açık ve anlaşılırdır. • Önerme kesinlik ifade etmektedir. • Önerme tek bir bilginin doğruluğunu ölçmektedir. Coğrafi unsurların harita üzerindeki gösterimi hakkında *bilgi. • Önerme yanlış ise, haritalardaki standart gösterimlerin coğrafi unsurları temsil etme şekli hakkında yanlış bir anlama neden olabilir. • Önerme kişisel görüşleri içermez. • Önerme olumsuz bir ifade içermez. Önerme çifte olumsuzluk içermez. • Kafa karıştırıcı ifadelerden kaçınılmıştır. • Önerme, coğrafya dersindeki temel kavramları anlama düzeyindeki öğrencilere uygun bir ifade içermektedir. <i>Önerme verilen kriterlere göre incelendiğinde, önermenin birçok kriteri karşılamadığı ve bilimsel gerçeklere aykırı olduğu görülmektedir. Bu nedenle, genel olarak uygun bir doğru-yanlış önermesi olmadığı söylenebilir.</i>

Tablo 5’te verildiği üzere uygulamaların D-Y önermelerini araştırmacılar tarafından hazırlanmış kriterlere göre incelemeleri istendiğinde her üç yapay zekâ uygulaması önermede tek bir fikrin/bilginin ölçüldüğünü belirtmiştir. Önermedeki temel bir problem birden fazla fikrin tek bir önerme içerisinde sunulmuş olmasıdır. Bu durumu üç yapay zekâ uygulaması da kriter olarak verilmesine rağmen belirleyememiştir. Fakat her üç uygulamada son genel geribildirimlerinde sunulan önermeyi kullanmayı uygun bulmamıştır.

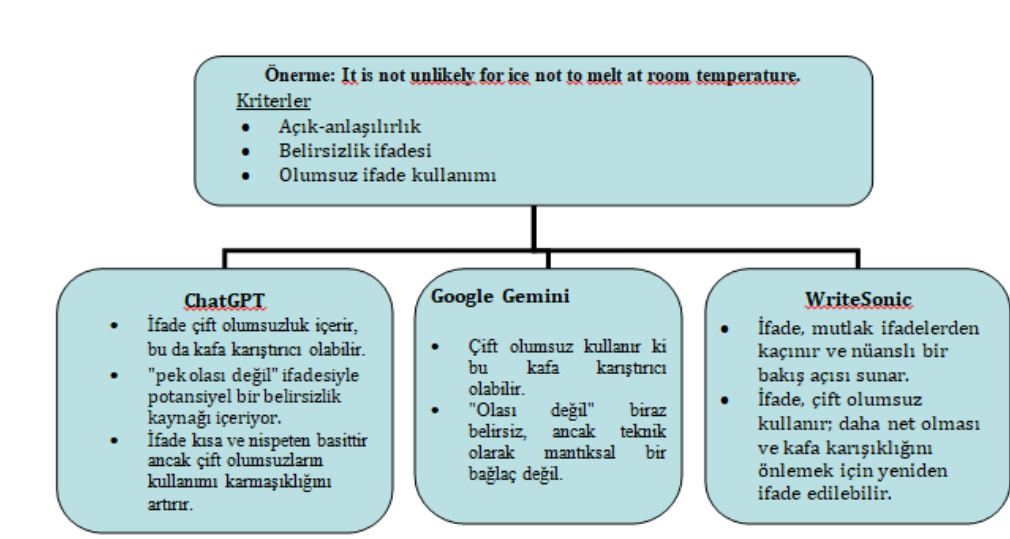
Araştırmacılar tarafından hazırlanan 9 kriter her üç yapay zekâ uygulamasına verilmiş ve önermeyi kritik etmesi istenmiştir. Sunulan kriterlerden en az 6’sı her bir uygulama tarafından önerme için doğru değerlendirilmiştir. Bunların haricinde yapay zekâ uygulamaları önermede sunulan bilginin yanlışlığına odaklanarak bu durumu farklı açılardan uygun bulmamıştır. Örneğin; ChatGPT sunulan yanlış bilginin açık-anlaşılabilirlik ve öğrenci seviyesine uygunluğu etkilediğini belirtmiştir. Google Gemini yanlış bilgiyi öğrenmeyi destekleyeceği için uygun bulmamıştır. Writesonic ise sunulan bilginin ülkeden ülkeye değişebileceği, bu nedenle önermenin kesin doğru ya da yanlış ifade içermediğini belirtmiştir.

Araştırmanın son kısmında yapay zekâ uygulamalarının kullanılan dile yönelik algılamasını test etmek için yapay zekâ uygulamalarının doğru-yanlış önerme hazırlamaya yönelik kriterler oluşturma ve önerme inceleme süreci Tablo 2’de verilen örnek için İngilizce ve Türkçe dilinde ayrı ayrı yürütülmüştür. Yapay zekâ uygulamalarının “It is not unlikely for ice not to melt at room temperature.” önermesine ilişkin her iki dilde verdiği gözden geçirmeler, geri bildirimler incelenmiş ve bulunan en belirgin eksiklikler açısından karşılaştırılmıştır. Şekil 1’de İngilizce gözden geçirme sürecinde üç yapay zekâ uygulamasının önerme için tespit ettiği önemli eksiklikler paylaşılmıştır.

Şekil 1’de görüldüğü gibi, her üç uygulamada açık-anlaşılabilirlik, belirsizlik ifadesi ve olumsuz ifade kullanımı ile ilgili önermedeki eksiklikleri doğru tespit edebilmiştir. Tablo 2’de verilen Türkçe gözden geçirmelerle Şekil 1’de verilen İngilizce gözden geçirmeler arasındaki en temel farklılık ise olumsuz ifade kullanılmasını Türkçe dilinde yapılan incelemede bir tek ChatGPT dikkate alırken, İngilizce incelemede üç uygulamada olumsuz ifade kullanımına ilişkin geri bildirim vermiştir. Yapılan dilsel bu karşılaştırmada üç yapay zekâ uygulaması açık-anlaşılabilirlik ve belirsizlik ifadesi kullanımı açısından ise Türkçe ve İngilizce olarak her iki dilde de gerekli geri bildirimi vermiştir.

Şekil 1

İngilizce Prompt İle Önermeyi Gözden Geçirmede Tespit Edilen Önemli Eksiklikler



TARTIŞMA, SONUÇ VE ÖNERİLER

Ölçme aracı hazırlama sürecinin aşamalarından biri de hazırlanan soru/maddelerin ölçme ve değerlendirme ilkeleri bakımından gözden geçirilmesidir. Sınıf içi ölçme uygulamalarında öğretmenler çoğu zaman hazırladıkları sorular için görüş alacak uzmanlara ulaşamamaktadır. Bu durumda hazırladıkları maddelerin gözden geçirilmesi/incelemesi yapılamamaktadır. Bu açıdan yapay zekâ uygulamaları ön inceleme görevini üstlenebilir mi? Bu doğrultuda araştırmada, doğru-yanlış önermelerini gözden geçirmede belirlenen yapay zekâ uygulamalarının kullanılabilirliğini incelemek amaçlanmıştır.

Yapılan incelemelerde her üç yapay zekâ uygulamasının hazırlanan önermelerin öğrenci düzeyine, ders hedef ve kapsamına uygun olması, önermelerin açık-anlaşılır olması, önermelerde net-kesin bir cevabın olması, önermedeki yanlışlığın önemli bir noktada olması, her bir önermenin tek bir bilgiyi/fikri ölçmesi, doğru-yanlış cevap dengesi gibi kriterleri ele aldığı belirlenmiştir. Bunlar haricinde kriter oluşturmada aralarındaki en belirgin farklılık doğru-yanlış önermelerinde olumsuz ifade kullanımı ve önerme ifadelerini bir kaynaktan olduğu gibi almama kriterleridir. Önermelerde olması gereken bu kriterlere bir tek ChatGPT değinmiştir. Birden fazla olumsuz ifade içeren D-Y önermesi sunulduğunda da bu hatayı bir tek ChatGPT belirlenmiştir.

Önermede olumsuz ifadenin kullanımı, birden fazla olumsuz ifadenin yer alması ile ilgili en iyi performansı ChatGPT göstermektedir. ChatGPT hem kendi oluşturduğu hem de verilen kriterlere göre inceleme yapmada olumsuz ifade kullanımı ile ilgili belirlenmelerini doğru yapmıştır. Google Gemini ve Writesonic ise olumsuzluk kriteri verildiğinde geri bildirimde bulunmuş fakat birden fazla olumsuzluğu belirleyememiştir. Fakat uygulamalara İngilizce prompt verildiğinde olumsuzlukla ilgili her üç uygulamanın da doğru geri bildirimler verdiği sonucuna ulaşılmıştır.

Araştırmada hem kriter oluşturmada hem de verilen önermeleri gözden geçirmede her üç uygulamanın da önermenin açık-anlaşılabilirliği bakımından oldukça iyi geri bildirimler verdiği tespit edilmiştir. Nasution (2023), ChatGPT ile geliştirilen çoktan seçmeli test maddelerine ilişkin öğrencilerin görüşlerini aldıkları çalışmalarında öğrencilerin büyük çoğunluğunun yapay zekâ ile oluşturulan soruları açık ve anlaşılır buldukları sonucuna ulaşmıştır. Ayrıca önermede belirsizlik ifadesinin yer almasıyla ilgili kendi oluşturdukları kriterlere göre incelemede ChatGPT ve Google Gemini bu hatayı belirlemiş, Writesonic ise bu konuda doğrudan bir inceleme yapmamıştır. Araştırmacıların oluşturdukları kriterlere göre yapılan incelemede ise önermedeki bu eksikliğe ilişkin her üç uygulamada doğru geri bildirimde bulunmuştur. Literatürde doğrudan D-Y önermelerinin yapay zekâ uygulamaları ile gözden geçirilmesini içeren çalışma bulunmamakla birlikte; Kıyak vd. (2024), ChatGPT ile oluşturulan soruların kabul edilebilir güvenilirlik düzeyinde olmasını, Taşdelen-Teker vd. (2024) ise ChatGPT4 ile hazırlanan olgu temelli çoktan seçmeli soruların küçük düzeltmelerle birlikte kullanımının madde yazımını kolaylaştırdığı sonuçlarına ulaşılmıştır.

Her üç yapay zekâ uygulaması da hem kendi oluşturduğu kriterlerde hem de kendilerine verilen kriterlerde “her bir önermede tek bir fikrin/bilginin doğruluğunun ölçülmesi” ilkesine ait doğru geri bildirimler verememiştir. Ayrıca Google Gemini’nin düzeltme örnek ifadelerine de yer verdiği, Writesonic’in diğerlerine göre daha kısa ifadelerle, yüzeysel geri bildirimler sunduğu ve daha çok önermedeki bilginin bilimsel olarak yanlışlığı üzerinde durduğu, önermelerdeki bulması beklenen hataları tam olarak tespit edemediği sonuçlarına ulaşılmıştır.

Araştırmada ayrıca uygulamaların ölçme değerlendirme ilkeleri bakımından kendilerine kriterler verildiğinde önermeler için daha uygun geri bildirimler verdiği, Türkçe-İngilizce incelemede dil farklılığının olduğu, uygulamaların İngilizce yazışmalar için daha uygun geri bildirimler verdiği sonuçlarına ulaşılmıştır. Literatürde doğrudan İngilizce ve Türkçe

hazırlanmış olan önermelerin incelenmesine yönelik bir çalışma olmamakla birlikte, Soysal ve Ardiç-Kemerkaya (2024), ChatGPT'nin İngilizce olarak verilen maddeleri cevaplama da daha iyi bir performans gösterdiği sonucuna ulaşmıştır.

Araştırma sonuçlarından hareketle, doğru-yanlış önermelerine ölçme ve değerlendirme ilkeleri bakımından geri bildirim vermede olası hataları görebilme açısından ChatGPT'nin daha uygun olduğu belirlenmiştir. Hazırlana önermelere geri bildirim vermede ChatGPT kullanılması durumunda daha doğru ve açıklayıcı bilgilere ulaşılmaktadır. Konuyla ilgili Kıyak vd. (2024), ChatGPT tarafından oluşturulan soruların kabul edilebilir standartları karşılayan psikometrik özellikler sergilediği, Wang vd. (2024) ise ChatGPT tabanlı otomatik soru üretme yaklaşımının geçerliliği ve güvenilirliği, soru üretimde insan performansına yaklaştığı sonucuna ulaşmıştır. Bu doğrultuda literatürden elde edilen dolaylı bulgular araştırma sonuçlarını destekler niteliktedir. Bu doğrultuda ChatGPT'nin soru kritik etmede, soruların redaksiyonunda, hazırlanan sorulara ön geri bildirim almada ve doğru soru hazırlamada eğitimcilere daha çok katkısı olacağı düşünülmektedir. Bu durum Bozkurt (2023)'ün ifade ettiği üzere, yapay zekâ uygulamalarının eğitimcilere geçerli ve güvenilir değerlendirme araçları sunma potansiyelini destekler nitelikte görülmektedir. Böylece, yapay zekâ destekli araçların eğitimde ölçme ve değerlendirme süreçlerine entegre edilmesinin, soru üretme ve hazırlamada önemli bir potansiyel taşıdığı söylenebilir.

Sonuç olarak, yapay zekânın mantıklı ve doğru soru kritik etmesi için bu talebe ait içeriği, sorunun türüne uygun doğru çerçeveleyerek mümkün olduğunca net bir şekilde ifade etmek gerekir. Bununla birlikte yapay zekâ uygulamalarının geldiği bu noktada Türkçe ve İngilizce dilinde yapılan önerme incelemelerinde farklılıklar tespit edilmiştir. İlgili uygulamalar İngilizce dilinde yapılan sorgulamada önermedeki olası hataları daha iyi belirleyebilmiştir. Ayrıca uygulamalar soruları genel olarak kritik etse de insan doğrulaması gerektirmektedir. Bu açıdan yapay zekâ uygulamalarını sorulara geri bildirim vermede, soruların redaksiyonunda uzmanların yerine geçmesinden ziyade tamamlayıcı olarak kullanmak önerilmektedir. Bu yaklaşımla yapay zekâ ve insanların birlikte çalıştığı "işbirlikçi zekâ" modeli öne çıkmaktadır (Dorsey & Michaels, 2022). Indran vd. (2023), tıp eğitiminde soru üretmeye ChatGPT'yi entegre ettiği çalışmada talimatların iyileştirilmesiyle yapay zekânın standartlı sorular üretebildiğini fakat insan katılımının yerinin doldurulamaz olduğu, içeriğin doğruluğunu, alakalılığını ve kalitesini doğrulamak için konunun uzmanlarını içeren sıkı bir inceleme sürecinin önemli olduğu sonucuna ulaşmıştır. Liu vd. (2024), yapay zekâ ile okuduğunu anlama üzere hazırlanan soruların kalitesinin incelendiği çalışmada ise öğrenci-yapay zekâyı işbirliği ile ortak soru üretmenin okuduğunu anlama için sınıf içi ölçme ve değerlendirmelerde faydalı olduğu, öğrenme sonuçlarını iyileştirdiği, yapıcı geri bildirimleri ile soru kalitesini iyileştirmede faydalı olduğu belirlenmiştir. Ayrıca literatürde, yapay zekâ temelli soru üretimde oluşturulan soruların pedagojik değere sahip olduğu, yapay zekânın soru hazırlama ve sorulara geri bildirim vermede kullanılabileceği (Seldon & Abidoye, 2018), fakat soruların doğruluğunun teyidi için insan denetiminde ihtiyaç duyulduğu, insanlar ve yapay zekâ arasındaki sinerjik işbirliklerinin potansiyeli (Bachiri vd., 2023; Bulut & Yıldırım-Erbaşlı, 2022; Wang vd., 2024; Yeşilyurt, 2024) vurgulanmaktadır. Sonuç olarak literatürden elde edilen bu sonuçlar araştırmadan elde edilen sonuçları destekler niteliktedir.

Bir öğretmen için hazırladıkları soruları gözden geçirmede, doğru soru hazırlamada yapay zekâ uygulamalarını kullanımının zamansal tasarruf sağlayacağı düşünülmektedir. Araştırma sonuçlarından hareketle öğretmenlerin-eğitimcilerin hazırladıkları doğru-yanlış sorularını uzman görüşüne sunmadan önce ön geri bildirim almada yapay zekâ uygulamalarını kullanmaları önerilebilir. Araştırmanın çıkış noktası öğretmen adaylarının hazırladığı doğru-yanlış önermelerini incelemektir. Bu doğrultuda araştırma doğru-yanlış önermeleri ile sınırlıdır. Uygulamaların geliştirilebilmesi için daha fazla araştırmaya ihtiyaç vardır. Araştırma farklı soru türleri için tekrarlanabilir. Öğretmenlere yapay zekâ ile soru yazma eğitimi verilip hazırladıkları

soruların incelemesinin gerçekleştirilmesine yönelik yeni çalışmalar yapılabilir. Ayrıca yapay zekâ uygulamalarının ele alınan davranışa yönelik madde türleriyle birlikte eğitilmesi gerektiği bir gerçektir. Yapay zekadan mantıklı ve sağlam sorular üretmesini beklemek için içerik mümkün olduğunca net çerçevelenmelidir (Bartoli vd., 2024). Öğretmenler tarafından hazırlanan soruları gözden geçirmede çeşitli madde türlerinden soru havuzu oluşturulması ve oluşturulan soruların yapay zekâyla paylaşılması gerekmektedir. Maddelerin yapay zekâyâ verilmesiyle birlikte sorularda bulunan hataların saptanması, soruların yapay zekâ tarafından çözülmesi istenerek uygulamaların ele alınacak konuyu öğrenmesi ve daha güvenilir sonuçlar ortaya sunması sağlanmalıdır (Halaweh, 2023).

Araştırmada yapay zekâ uygulamalarının soru kritik etmede, sorulara geri bildirim vermede nasıl bir performans gösterdiği tartışılmıştır. Farklı bir araştırmada soru hazırlama, puanlama açılarından yapay zekâ uygulamalarının etkisi tartışılabilir. Araştırma kapsamında ChatGPT, Google Gemini ve WriteSonic yapay zekâ sistemleri kullanılmıştır. Gelecek araştırmalarda yeni geliştiren ya da mevcut olarak aktifliğini sürdüren (Co-Pilot, Character.ai vb.) yapay zekâ sistemleri üzerinden de soru maddelerine yönelik gözden geçirme süreçleri incelenebilir. Ayrıca yapay zekâ sistemleri her ne kadar Türkçe'ye uyarlanmış bir yapıda bulunsada veri tabanlarında bulunan bilgilerin yansıtılmasında çeşitli eksiklikler bulunmaktadır. Dil probleminin ortadan kaldırılması amacıyla yerli ve milli yapay zekâ sistemlerinin oluşturulması ve yaygınlaştırılması gerekmektedir. Türkiye'de yapay zekâ sistemlerinin eğitime uyarlanmasına yönelik Millî Eğitim Bakanlığı ile çeşitli yapay zekâ kuruluşları iş birliğine gitmelidir. Bu iş birliği kapsamında öğretmenlere ve öğretmen adaylarına yapay zekâ sistemlerinde eğitime yönelik neler yapabilecekleri açıklanmalıdır.

KAYNAKÇA

- Akçapınar, G., Er, E., Karagül, A. E., & Bakan-Kalaycıoğlu, D. (2024, Ekim). Üretken yapay zekâ modellerinin çoktan seçmeli soru üretme performanslarının karşılaştırması. *Uluslararası Ölçme, Seçme ve Yerleştirme Sempozyumu*. ÖSYM, Ankara.
- Ali, R., Tang, O. Y., Connolly, I. D., Fridley, J. S., Shin, J. H., Zadnik-Sullivan, P. L., Cielo, D., Oyelese, A. A., Doberstein, C. E., Telfeian, A. E., Gokaslan, Z. L., & Asaad, W. F. (2023). Performance of ChatGPT, GPT-4, and google bard on a neurosurgery oral boards prep-aration question bank. *Neurosurgery*, 93(5), 1090–1098. <https://doi.org/10.1227/neu.0000000000002551>
- Bachiri, Y. A., Mouncif, H., & Bouikhalene, B. (2023). Artificial intelligence empowers gamification: Optimizing student engagement and learning outcomes in e-learning and MOOCs. *International Journal of Engineering Pedagogy (iJEP)*, 13(8), 4–19. <https://doi.org/10.3991/ijep.v13i8.40853>
- Baidoo-Anu, D., & Owusu Ansah, L. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Journal of AI*, 7(1), 52–62.
- Bartoli, A., May, A. T., Al-Awadhi, A., & Schaller, K. (2024). Probing artificial intelligence in neurosurgical training ChatGPT takes a neurosurgical residents written exam. *Brain and Spine*, 4, 1–5, 102715. <https://doi.org/10.1016/j.bas.2023.102715>
- Baykul, Y. (2000). *Eğitimde ve psikolojide ölçme: Klasik test teorisi ve uygulaması* (1.Baskı). ÖSYM Yayınları.
- Boulay, B. (2016). Artificial intelligence as an effective classroom assistant. *IEEE Intelligent Systems*, 31(6), 76–81.

- Bozkurt, A. (2023). ChatGPT, üretken yapay zekâ ve algoritmik paradigma değişikliği. *Alanyazın*, 4(1), 63-72. <https://doi.org/10.59320/alanyazin.1283282>
- Boztunç-Öztürk, N., & Şahin, M. G. (2024). Yapay zekâ İngilizce yazma becerilerini puanlamada ne kadar başarılı?. *Uluslararası Ölçme, Seçme ve Yerleştirme Sempozyumu. ÖSYM*, Ankara.
- Bulut, O., & Yildirim-Erbaşlı, Ş. N. (2022). Automatic story and item generation for reading comprehension assessments with transformers. *International Journal of Assessment Tools in Education*, 9, 72-87. <https://doi.org/10.21449/ijate.1124382>
- Ceylan, M., & Gökçe, S. (2024). Test geliştirmede paradigma değişimi: Yapay zekâ destekli çeldiricilerin etkililiği. *Uluslararası Ölçme, Seçme ve Yerleştirme Sempozyumu. ÖSYM*, Ankara.
- Chatterjee, S., & Bhattacharjee, K. K. (2020). Adoption of artificial intelligence in higher education: A quantitative analysis using structural equation modelling. *Education and Information Technologies*, 25, 3443-3463.
- Chen, L., Chen, P., & Lin, Z. (2020). Artificial intelligence in education: A review. *Ieee Access*, 8, 75264-75278.
- Chmiliar, I. (2010). Multiple-case designs. In A. J. Mills, G. Eurepas & E. Wiebe (Eds.), *Encyclopedia of case study research* (pp. 582-583). SAGE Publications.
- Creswell, J. W. (2007). *Qualitative inquiry & research design: Choosing among five approaches* (2. Baskı). SAGE Publications.
- Çam, M. B., Çelik, N. C., Turan-Güntepe, E., & Durukan, Ü. G., (2021). Öğretmen adaylarının yapay zekâ teknolojileri ile ilgili farkındalıklarının belirlenmesi. *Hatay Mustafa Kemal Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, 18(48), 263-285.
- Demir, S. (2023). Investigation of ChatGPT and real raters in scoring open-ended items in terms of inter-rater reliability. *International Journal of Turkish Educational Sciences*, 11(21), 1072-1099. <https://doi.org/10.46778/goputeb.1345752>
- Doğan, N. (2019). İnsan davranışları ve ölçme araçlarını sınıflama. N. Doğan (Ed.), *Eğitimde ölçme ve değerlendirme içinde* (s. 76-111). Pegem Akademi.
- Dorsey, D. W., & Michaels, H. R. (2022). Validity arguments meet artificial intelligence in innovative educational assessment: A discussion and look forward. *Journal of Educational Measurement*, 59(3), 389-394.
- Ercikan, K., & McCaffrey, D. F. (2022). Optimizing implementation of artificial-intelligence-based automated scoring: An evidence centered design approach for designing assessments for AI-based scoring. *Journal of Educational Measurement*, 59(3), 272-287. <https://doi.org/10.1111/jedm.12332>
- Gürbüz, M. Ç., & Gürbüz, T. (2024). Yapay zekâ ile sınav kâğıdının otomatik değerlendirilmesi. *Uluslararası Ölçme, Seçme ve Yerleştirme Sempozyumu. ÖSYM*, Ankara.
- Gürlek, Y., Bozkoyun, E., Ulutürk, M., & Zeyrekgündüz, F. (2023). Yapay zekânın eğitime etkileri ve uygulamaları. *International Journal of Original Educational Research*, 1(1), 25-132.
- Halaweh, M. (2023). ChatGPT in education: Strategies for responsible implementation. *Contemporary Educational Technology*, 15(2). <https://doi.org/10.30935/cedtech/13036>

- Huang, C. H., Hsiao, H. J., Yeh, P. C., Wu, K. C., & Kao, C. H. (2024). Performance of ChatGPT on stage 1 of the Taiwanese medical licensing exam. *Digital Health*, 10. <https://doi.org/10.1177/20552076241233144>
- Huang, S. P. (2018). Effects of using artificial intelligence teaching system for environmental education on environmental knowledge and attitude. *EURASIA Journal of Mathematics, Science and Technology Education*, 14(7), 3277-3284. <https://doi.org/10.29333/ejmste/91248>
- Indran, I. R., Paranthaman, P., Gupta, N., & Mustafa, N. (2023). Twelve tips to leverage AI for efficient and effective medical question generation: A guide for educators using Chat GPT. *Medical Teacher*, 46(8), 1021-1026. <https://doi.org/10.1080/0142159X.2023.2294703>
- Kan, A. (2014). Ölçme aracı geliştirme. S. Tekindal (Eds.), *Eğitimde ölçme ve değerlendirme* içinde (s. 259-296). Pegem Akademi.
- Karakaya, İ. (2022). *Açık uçlu soruların hazırlanması, uygulanması ve değerlendirilmesi*. Pegem Akademi.
- Kıyak, Y. S., Coşkun, Ö., Budakoğlu, I. İ., & Uluoğlu, C. (2024). ChatGPT for generating multiple-choice questions: Evidence on the use of artificial intelligence in automatic item generation for a rational pharmacotherapy exam. *European Journal of Clinical Pharmacology*, 80, 729-735. <https://doi.org/10.1007/s00228-024-03649-x>
- Koçak, D., & Şimşek, A. S. (2024). Yükseköğretimde üretken yapay zekânın öğrenciler tarafından ödev ve performans görevlerinde kullanımı: Avantajlar ve riskler. *Uluslararası Ölçme, Seçme ve Yerleştirme Sempozyumu*. ÖSYM, Ankara.
- Koçyiğit, A., & Darı, A. B. (2023). Yapay zekâ iletişimde ChatGPT: İnsanlaşan dijitalleşmenin geleceği. *Stratejik ve Sosyal Araştırmalar Dergisi*, 7, 427-438.
- Krippendorff, K. H., (2004). *Content analysis: An introduction to its methodology*. Sage Pub.
- Kucam, E., & Malakcıoğlu, İ. (2024). Psikomotor becerilerin ölçülmesinde yapay zekâ tespit araçlarından elde edilen sonuçların doğruluğunun incelenmesi. *Uluslararası Ölçme, Seçme ve Yerleştirme Sempozyumu*. ÖSYM, Ankara.
- Liu, M., Zhang, J. X., Nyagoga, L. M., & Liu, L. (2024). Student-AI question cocreation for enhancing reading comprehension. *IEEE Transactions On Learning Technologies*, 17, 815-826.
- McCarthy, J. (2024, 02 Nisan). What is artificial intelligence?. formal.stanford.edu/jmc/whatisai/.
- Montebello, M. (2018). *AI injected e-learning* (1. Baskı). Springer International Publishing.
- Nasution, N. E. A. (2023). Using artificial intelligence to create biology multiple-choice questions for higher education. *Agricultural and Environmental Education*, 2(1). <https://doi.org/10.29333/agrenvedu/13071>
- Önderöz, F. G., & Karabay, A. (2024). Sınıf öğretmenlerinin ve sınıf öğretmeni adaylarının eğitimde yapay zekâ teknolojilerine ilişkin görüşleri ve yapay zekâ kullanım durumları. *EJER Bildiri Özetleri Kitabı* (pp. 39-40). Anı Yayıncılık.
- Popenici, S. A., & Kerr, S. (2017). Exploring the impact of artificial intelligence on teaching and learning in higher education. *Research and Practice in Technology Enhanced Learning*, 22(12). <https://www.doi.org/10.1186/s41039-017-0062-8>

- Sallam, M., & Al-Salahat, K. (2023). Below average ChatGPT performance in medical microbiology exam compared to university students. *Frontiers in Education*, 8. <https://doi.org/10.3389/feduc.2023.1333415>
- Seldon, A., & Abidoye, O. (2018). *Fourth education revolution*. University of Buckingham Press.
- Shah, P. (2023). *AI and the future of education: Teaching in the age of artificial intelligence*. Jossey-Bass.
- Soysal, S., & Ardiç-Kemer kaya, E. Ö. (2024). PISA 2022 Matematik okuryazarlığı şekil formatındaki maddelerin Türkçe ve İngilizce olarak cevaplanmasında ChatGPT'nin performansı. *Uluslararası Ölçme, Seçme ve Yerleştirme Sempozyumu*. ÖSYM, Ankara.
- Strauss, A., & Corbin, J. M. (1990). *Basic of qualitative research. Grounded theory procedures and techniques*. Sage Publications.
- Şahin, M. Ç., Sözer, A., Kuzucu, P., Türkmen, T., Şahin, M. B., Sözer, E., Tüfek, O. Y., Nernekli, K., Emmez, H., & Çeltikçi, E. (2024). Beyond human in neurosurgical exams: ChatGPT's success in the Turkish neurosurgical society proficiency board exams. *Computers in Biology and Medicine*, 169. <https://doi.org/10.1016/j.compbiomed.2023.107807>
- Taşçı, G., & Çelebi, M. (2020). Eğitimde yeni bir paradigma:Yükseköğretimde yapay zekâ. *OPUS International Journal of Society Researches*, 16(29), 2346-2370.
- Taşdelen-Teker, G., Bakan-Kalaycıoğlu, D., Şahin, M. G., & Esenboğa, S. (2024). Yapay zekâ tarafından oluşturulan olgu temelli çoktan seçmeli maddelerin değerlendirilmesi: Pediatri örneği. *Uluslararası Ölçme, Seçme ve Yerleştirme Sempozyumu*. ÖSYM, Ankara.
- Tezci, i. H., & Bıçak, B. (2024). Yapay zekâ ve nitel veri: Geleceğin analiz yöntemleri. *Uluslararası Ölçme, Seçme ve Yerleştirme Sempozyumu*. ÖSYM, Ankara.
- Turgut, M. F., & Baykul, Y. (2012). *Eğitimde ölçme ve değerlendirme* (4. Baskı). Pegem Akademi.
- Vural, N. E. (2024). Açık uçlu sınavların otomatik değerlendirilmesi: Görüntü işleme ve yapay zekâ destekli iOS tabanlı mobil uygulama geliştirme örneği. *Uluslararası Ölçme, Seçme ve Yerleştirme Sempozyumu*. ÖSYM, Ankara.
- Wang, L., Song, R., Guo, W., & Yang, H. (2024). Exploring prompt pattern for generative artificial intelligence in automatic question generation. *Interactive Learning Environments*, 1–26. <https://doi.org/10.1080/10494820.2024.2412082>
- Yeşilyurt, Ş. (2024). Yapay zekâ ile C1 seviyesi Türkçe okuma sınavı oluşturma ve uygulama. *Uluslararası Ölçme, Seçme ve Yerleştirme Sempozyumu*. ÖSYM, Ankara.
- Yiğiter, M. S., & Boduroğlu, E. (2024). Açık uçlu maddelerin puanlanmasında yapay zekâ performansının incelenmesi. *Uluslararası Ölçme, Seçme ve Yerleştirme Sempozyumu*. ÖSYM, Ankara.
- Zeide, E. (2019). Artificial intelligence in higher education: Applications, promise and perils, and ethical questions. *EDUCAUSE Review*, 1-13.

EXTENDED ABSTRACT

Introduction

When artificial intelligence (AI) is considered in the context of measurement activities in education, the application area of application of AI systems is becoming increasingly widespread. Artificial intelligence applications in classroom assessment and evaluation can bring benefits to educators/teachers. For example, it can reduce the time needed for teachers to give feedback on students' work in classroom assessment, or help the teacher to generate questions and prepare the right questions.

Today, teachers can benefit from artificial intelligence in item creation. One of the stages in the measurement tool preparation process is to review the prepared questions/items in terms of measurement and scoring principles. In classroom measurement practice, teachers often do not have access to experts to get opinions on the questions they prepare. In this context, can artificial intelligence applications take over the task of pre-testing the accuracy and quality of the questions? In this direction, this study aims to investigate the usability of artificial intelligence applications in checking true-false items.

Methods

The research is a case study using qualitative research methods. The purpose of the case study is to provide detailed descriptions and understand the situation as it exists (McMillan, 2004). The research was conducted in the spring semester of the academic year 2023-2024 with the third year students studying in different departments of the Faculty of Education at a state university, who were selected from the incorrectly prepared true-false items prepared within the framework of the course "Measurement and Evaluation in Education". After being trained in the preparation of true-false items in the course, the students were asked to prepare true-false items related to their specialities. Two true-false items were selected for use in the study, taking into account the criteria of subject area, language, and different levels of error. Three different artificial intelligence applications (ChatGPT 3.5, Google Gemini, and Writesonic) were used to critique the selected items. In determining the AI applications used in the research, attention was paid to the criteria of known and widespread use were taken into account. The determination of the AI applications were reviewed and critiqued by an associate professor and a research assistant in the field of "Measurement and Evaluation in Education". The same terms were used in all correspondence with ChatGPT, Google Gemini, and Writesonic. The same true/false statements were analyzed by all three artificial intelligence applications (March-April, 2024). The process of creating criteria with the three AI applications and reviewing the identified items was repeated, discussed, and finalized by the researchers separately and in different time periods. The study used content analysis and descriptive analysis.

Results and Discussion

In the research, it was found that all three AI applications met criteria such as the prepared items being appropriate to the student's level, course objectives, and scope, the items being clear-comprehensible, the items having a clear-certain answer, the inaccuracy in the item being at an important point, each item measuring a single knowledge/idea, and the balance of true/false answers. Apart from these, ChatGPT shows the best performance in terms of the use of negative statements and the presence of more than one negative statement in the item. ChatGPT made the correct determinations regarding the use of negative expressions both in its creation and in the analysis according to the given criteria. Google Gemini and Writesonic, on the other hand, provided feedback when given the negativity criterion, but failed to identify more than one negativity. However, it was concluded that all three applications gave correct feedback on negativity when given an English prompt.

It was concluded that the applications provided more appropriate feedback for items when they were given criteria in terms of evaluation and evaluation principles, that there was a language difference in the Turkish-English analysis, and that the applications provided more appropriate feedback for English correspondence.

Based on the results of the research, it was determined that ChatGPT is more appropriate in terms of being able to see possible errors in giving feedback to true/false items in terms of measurement and evaluation principles. When ChatGPT is used to give feedback on items, more accurate and explanatory information is obtained. Kıyak et. al. (2024) concluded in their study that the questions created by ChatGPT have psychometric properties that meet acceptable standards. In this direction, it is believed that ChatGPT will contribute more to educators in critiquing questions, editing questions, getting preliminary feedback on prepared questions, and preparing correct questions.

In order for AI to criticize a logical and correct question, it is necessary to express the content of this question as clearly as possible by formulating it correctly according to the type of question. However, at this point in the development of AI applications, differences were found in the propositional tests conducted in Turkish and English. The corresponding applications were better able to detect possible errors in the proposition in the English language query. In addition, although the applications criticize the questions in general, they require human verification. In this regard, it is suggested to use artificial intelligence applications as a complement rather than a substitute for experts in providing feedback on questions and proofreading questions. With this approach, the "collaborative intelligence" model, in which artificial intelligence and humans work together, comes to the fore (Dorsey & Michaels, 2022).

Based on the results of the research, it can be suggested that teachers-educators should use artificial intelligence applications to get preliminary feedback before submitting their true-false questions to expert opinion. The starting point of the research is to examine the true-false items prepared by pre-service teachers. Accordingly, the research is limited to true-false items. Further research is needed to improve the applications. The research can be repeated for different types of questions. In future research, review processes for question items can be studied using newly developed or currently active artificial intelligence systems (Co-Pilot, Character.ai, etc.).