Ampirik Araştırma / Original Empirical Research

www.yuksekogretim.org

TÜBA-HER Yükseköğretim Dergisi
TÜBA Higher Education Research/Review

Üniversite Öğrencilerinin Yazılı Üretimlerinin OpenAI GPT ile Değerlendirilmesi

Assessment of University Students' Essays through OpenAI GPT

Ayfer Sayın[1], Deniz Melanlıoğlu[2]

[1] Gazi Üniversitesi, Gazi Eğitim Fakültesi, Eğitim Bilimleri Bölümü, Ankara / Türkiye
[2] İstanbul Üniversitesi, Edebiyat Fakültesi, Dilbilimi Bölümü, İstanbul / Türkiye

Özet

Yazılı üretim ortaya koyma öğrenciler tarafından her zaman zor bir görev olarak nitelenmektedir. Yazmaya ilişkin tutum, motivasyon, değerlendirme ölçütleri, konuya hâkimiyet, dili kullanma yetisi gibi değişkenler yazmaya önyargıyla yaklaşmaya neden olmaktadır. Ancak yazma sadece bir akademik başarı göstergesi değil aynı zamanda iş dünyasında da ihtiyaç duyulan bir beceridir. Bu nedenle üniversite eğitimi sırasında öğrencilerin etkili yazma becerisi kazanmaları önemli görülmektedir. Öğrencilerin yazmada yetkinlik kazanması, yazma uygulaması yapmaya ve ürünlere geribildirimde bulunmayla yakından ilişkilidir. Anlamlı dönüt verilmesinde yazılı üretimlerin objektif değerlendirilmesi gerekmektedir. Yazmaya yönelik objektif değerlendirmede bulunma bazen puanlayıcıların yaklaşımından kaynaklı olarak geçerli ve güvenilir sonuçlar vermeyebilir. Böyle bir durumda teknolojinin sunduğu imkânlardan faydalanılabileceğine inanılmaktadır. Bu bağlamda araştırmada, üniversite öğrencilerinin yazılı üretimlerinin insan puanlayıcılar ve yapay zekâ tarafından puanlanmıştır. Bu iki puanlamanın karşılaştırılarak incelenmesi araştırmanın amacını oluşturmuştur. Böylece yazılı üretimleri değerlendirmede OpenAI tarafından geliştirilen GPT yapay zekâ sistemlerinin kullanılabilirliği sınanacaktır. Söz konusu amaç doğrultusunda araştırma, ilişkisel tarama modelinde yürütülmüştür. Katılımcılar, gönüllülük esasına bağlı olarak bir devlet üniversitesinde öğrenim gören 60 birinci sınıf öğrencisidir. Araştırma kapsamında katılımcılardan yazılı üretim görevi doğrultusunda bir metin oluşturmaları istenmiş ve bunlar bütüncül puanlama anahtarı ile puanlanmıştır. Elde edilen verilerin analizi sonrasında GPT ile uzmanların puanları arasında pozitif yönde ve orta düzeyde bir ilişki olduğu saptanmıştır.

Anahtar Kelimeler: OpenAI GPT, Türkçe Eğitimi, Yazma Becerisi, Bütüncül Puanlama, Üniversite Öğrencisi.

Abstract

Students always characterize written essays as a difficult task. Variables such as attitude towards writing, motivation, evaluation criteria, mastery of the subject, and ability to use language cause prejudice towards writing. However, writing is an indicator of academic success and a skill needed in the business world. For this reason, it is considered essential for students to gain effective patching skills during university education. Students gaining competence in writing is closely related to writing practice and giving feedback to products. In providing meaningful feedback, written essays should be evaluated objectively. Objective evaluation of writing may sometimes not give valid and reliable results due to the approach of the scorers. In such a case, the opportunities offered by technology can be utilized. In this context, the study aims to examine the scoring of university students' written essays by human raters and artificial intelligence, compare and analyze the two scorings, and test the usability of GPT artificial intelligence systems developed by OpenAI in evaluating written essays. The research was conducted in the relational survey model in line with this purpose. The participants were 60 first-year students studying at a state university voluntarily. Within the scope of the research, the participants were asked to create a text in line with the written essay task, and these were scored with a holistic scoring key. After the analysis of the data obtained, it was found that there was a positive and moderate relationship between GPT and experts' scores.

Keywords: OpenAI GPT, Turkish Education, Writing Skill, Holistic Scoring, Higher Education Attendee.

İletişim / Correspondence:
Prof. Dr. Ayfer Sayın
Gazi Üniversitesi, Gazi Eğitim Fakültesi, Eğitim Bilimleri Bölümü, Ankara
e-posta: ayfersayin@gazi.edu.tr

Yükseköğretim Dergisi / TÜBA Higher Education Research/Review (TÜBA-HER), 14(3), 121-134. © 2024 TÜBA
Geliş tarihi / Received: Ocak / January 12, 2024; **Kabul tarihi / Accepted:** Nisan / April 4, 2024
Bu makalenin atıf künyesi / How to cite this article: Sayın, A. & Melanlıoğlu, D. (2024). Üniversite Öğrencilerinin Yazılı Üretimlerinin OpenAI GPT ile Değerlendirilmesi. *Yükseköğretim Dergisi, 14*(3), 121-134. https://doi.org/10.53478/yuksekogretim.1418870

Bu çalışma 11-14 Nisan 2024 tarihinde düzenlenen National Council on Measurement in Education (NCME), Philadelphia, ABD'de sözlü bildiri olarak sunulmuştur.

ORCID: A. Sayın: 0000-0003-1357-5674; D. Melanlıoğlu: 0000-0002-3663-0894.

Ana dili edinimi, anne karnında başlayıp eğitim öğretimle desteklenen ve yaşam boyu gelişim gösteren bir süreçtir. Bu nedenle okul öncesinden üniversiteye değin ana dili eğitimine yönelik içeriklerle öğrencinin karşılaşması sağlanmaktadır. Eğitim sürecinde; dinleme, konuşma, okuma ve yazma becerisini geliştirme ile bunların etkin kullanımını alışkanlık hâline dönüştürmeye yönelik bir içerik sunulmaktadır. Bu bağlamda etkili iletişim kurabilen, iletişim yoksunluğu yaşamayan bireylerin topluma kazandırılması noktasında ana dili eğitimi, temel gereksinim olarak düşünülmektedir. Barker ve Hall'a (1995) göre iletişim becerileri, akademik başarının yanı sıra kariyer gelişimi için de kritik öneme sahiptir. Araştırmalar, akademide ve iş dünyasında sözlü ve yazılı iletişim becerilerinin başarıda oynadığı rol üzerinde hemfikirdir (Quible ve Griffin, 2007). Son yıllarda e-posta, kısa mesaj ve anlık mesajlaşmanın yaygınlık kazandığı teknolojik araçların kullanımı, kurum ve kuruluşlarda bilgi paylaşımının daha hızlı olması gibi nedenler dolayısıyla yazılı iletişim becerilerine verilen önem gün geçtikçe artmaktadır (Brandt, 2005). Bu anlamda etkili yazma, yükseköğrenimde önemli bir beceri olarak belirtilmektedir (Kellogg ve Raulerson, 2007). Ancak birçok üniversite mezununun etkili yazma becerisinden yoksun olduğu yapılan araştırmalarda vurgulanmaktadır (Ashton, 2007; Henricks, 2007). Belirtilen durum öğrencilerin, yazma becerisindeki eksikliklerle üniversite eğitimine başlamasıyla ilişkilendirilmekte (NAEP, 2002) ve eksikliklerin giderilmesinde dört yıllık üniversite eğitiminin yeterli olmadığı (Bacon ve Anderson, 2004), birçok öğrencinin etkili yazma becerisinden yoksun kaldığı ifade edilmektedir (May ve ark., 2012). Oysa etkili yazma; işe girme (Stevens, 2005), kariyer sahibi olma (Rowh, 2006) ve finansal refaha ulaşmayla (Fisher, 1999) ilişki bir beceri şeklinde yorumlanmaktadır. İşverenlerin birçoğunun yeni işe alınan üniversite mezunlarının yazma eksikliklerine karşı kaygı duyduğu (Henricks, 2007) ve çalışanlarını yazma becerisinde geliştirmek amacıyla yüksek bütçeler ayırmak zorunda kalmaktan yakındıkları bilinmektedir (Quible ve Griffin, 2007; Smerd, 2007). Bu nedenle üniversite derslerinin yazma uygulamaları içermesi gerektiğine inanılmaktadır (Stevens, 2005).

Üniversite eğitiminin ilk yıllarından itibaren öğrencilere metin oluşturma, farklı metin türlerinde yazma imkânı verilmesi ve öğrencilerin gelişimlerinin izlenmesi tavsiye edilmektedir (Benjamin ve Chun, 2003). Enos (2010) üniversite öğrencilerinin yazma becerisini geliştirmek için birinci sınıfta yazılı üretimler üzerinden düzenleme ve düzeltme uygulamalarına öncelik verilmesi gerektiğini savunmaktadır. Bu uygulamaların ilk sınıftan başlamasının bir gereksinim olduğuna işaret eden araştırma sayısı da dikkat çekmektedir (May ve ark., 2012; Riordan ve ark., 2000; Rowe, 2006). Kahraman ve Yalvaç (2015) yazma başarısının diğer dil becerilerinin gelişiminden doğrudan etkilenmesi ve yazma kurallarının keskin olması gibi değişkenler nedeniyle öğrencilerin yazma görevlerinde zorlandığını söylemektedir. Üretim becerisi olarak yazma, duygu ve düşünceleri ifade etme ihtiyacını karşılamada etkili araçlardan biridir. Ancak öğrencinin yazma performansını etkileyen metnin konusu, amacı, ana fikri, kullanılan yazma stratejisi gibi bazı etmenler vardır. Ayrıca öğrencinin tutumu, motivasyonu gibi duyuşsal değişkenler de yazılı üretimin niteliğini belirlemede rol oynamaktadır (Monis ve Rodriques, 2012). Öğrencilerin yazmaya ilişkin inançları genellikle yazma pratiklerini ve alışkanlıklarını da etkilemektedir (Greenberg, 1988). Çünkü öğrenciler yazma becerisine ön yargılı yaklaşmakta, yazmayı en zor dil becerisi olarak nitelemekte ve kendilerini yazma görevini yerine getirecek donanımda hissetmemektedir. Bu açıdan bakıldığında yazma eğitiminde öğrencilerin yazılı üretimlerine dönüt verilmesi ile ölçme değerlendirme çalışmalarının yapılması becerinin gelişimi açısından oldukça önemli olduğu ifade edilebilir. Bir yazılı üretimi değerlendirmede, öğrencinin yazmaya ilişkin yeteneği dışında birçok değişken, yazma puanında farklılaşmaya yol açmaktadır. Olası varyans kaynakları; konu (belirlenmiş veya öğrencinin kendi seçtiği), söylem tarzı, metin türü (açıklama, tartışma vb.), yazma süresi, yazma modu (kâğıt ve kalem veya metin işlemci), değerlendirici tutarsızlığı, puanlama prosedürü (bütünsel veya analitik) ve puanlanacak özellikler (içerik, dil kullanımı veya yazım) şeklinde sıralanabilir (Huot, 1990). Değişken sayısının çok olduğu böyle bir durumda neyin hata kabul edilip edilmeyeceği konusu bir tartışmaya dönüşmektedir. Marcoulides (1998) değerlendirmenin geçerli ve güvenilir olması için farklı yazma görevlerinin aynı kabul edilerek değerlendirilmede bulunulmaması ve her görevin kendi nitelikleriyle değerlendirilmesi gerektiğini belirtmektedir. Yazma görevinin değerlendirilmesinde geçerli ve güvenilir sonuçlar elde etme, derecelendirme güvenilirliği ve görev güvenilirliği açısından oldukça sorunlu bir alan olarak görülmektedir (Kroll, 1998). Örneğin değerlendiriciler/puanlayıcılar aynı metinlere yönelik yaptıkları değerlendirmede farklılık gösterebilmekte hatta farklı zaman aralıklarında aynı metne bakan bir değerlendirici çoğu zaman kendiyle de çelişebilmektedir (Charney, 1984). Yazma konusu, öğrencinin dünya bilgisi, retorik özellikler yazma performansını etkilediği için söz konusu değişkenlerin verilen görevi puanlamayı da etkilediği düşünülmektedir (Schoonen, 2005). Schoonen ve diğerlerine (1997) göre görevin ve değerlendiricinin etkisine, kullanılan puanlama yöntemini de eklemek gerekmektedir. Puanlama yöntemi, değerlendiricilerin görevleri derecelendirmek için aldıkları bilgi ve yönergeleri içermektedir. Görüldüğü gibi yazma performansını değerlendirmek oldukça karmaşık ve zor bir süreci içermektedir. Üniversite öğrencilerini yazma becerisinde geliştirme ve yazmayı alışkanlığa dönüştürme noktasında yapılacak değerlendirmenin dönüt verme noktasında hayati öneme sahip olduğu söylenebilir. Değerlendirici güvenirliğini sağlama açısından teknolojinin sunduğu imkânlardan yararlanılabileceği düşünülmektedir. Bu bağlamda yapay zekânın sunduğu imkânlardan faydalanılabilceğine inanılmaktadır.

Yazılım mühendisliği ve yapay zekâ alanlarının bir bileşeni olan Doğal Dil İşleme (Natural Language Processing - NLP); doğal dil metinlerini otomatik olarak anlama, yorumlama ve üretme yeteneğini geliştirmeyi amaçlayan bilgisayar destekli bir analitik tekniktir (Kang ve ark., 2020; Saravia, 2018). İnsanların konuştuğu dili analiz etmek ve anlamak, iletişim, bilgi alma, otomatik dil işleme, doğal dil arayüzleri, duygu analizi, bilgi çıkarma, çeviri ve üretme gibi birçok alanda önemli faydalar sağlamaktadır (Hirschberg ve Manning, 2015; Kyparisis, 1987; Tan ve ark., 2015; Zech ve ark., 2022). Bu nedenle NLP, insanlar ve makineler arasında etkili iletişimi sağlamak, büyük veri setlerini anlamlandırmak ve metin tabanlı uygulamaların geliştirilmesine yardımcı olmak için kullanılmaktadır (Hirschberg ve Manning, 2015; Xu ve Lv, 2022). Adams ve Chuah (2022) başta yazma becerisi olmak üzere teknolojinin dil öğretiminde etkin kullanıldığını söylemektedir. Düzeltme için e-geri bildirim kullanma (Tuzi, 2004), wiki tabanlı iş birliği uygulama (Hsu, 2019), yazma kalitesini iyileştirmek için derlem destekli hata çözümleme (Crosthwaite ve ark., 2020), sosyal medya ile yazma (Lee, 2020) teknolojinin yazma becerisine entegre edilmesine örnek olarak sıralanabilir. Teknolojinin dil eğitiminde kullanımı yeni olmamakla birlikte GPT ile yaygınlık kazandığı dile getirilebilir (Okonkwo ve Ade-Ibijola, 2021). Akıllı, hızlı ve çok dilli yanıt verme yeteneğiyle kendinden önceki uygulamalardan ayrılan GPT, OpenAI tarafından 2022 yılının Kasım ayında GPT-3.5 büyük ölçekli dil modelleri ailesinin üzerine inşa edilen yeni nesil bir sohbet robotudur (OpenAI, 2022). Dönüştürücü mimarisine dayalı büyük dil modeli olan GPT modelleri, ilk olarak 2018'de OpenAI tarafından tanıtılmıştır. Bunu 2019'da GPT-2, 2020'de GPT-3 ve 2023'te GPT-4 izlemiştir. 2023'te bu modeller büyük dil modellerinin çoğu benzer özellikleri paylaştığından genellikle GPT'ler olarak adlandırılmıştır. "Ön eğitimli (pre-trained)" ve "dönüştürücü (transformer)" kelimelerinin birleşiminden oluşan GPT, metin okuma ve yazma yeteneğine sahip bir yapay zekâ kodudur (Pavlik, 2023). Kelime dizilerini istatistiksel olarak tahmin etmek için makine çevirisinde yaygın olarak kullanılır. GPT'nin eğitim için dil sözlüğü İngilizcedir ancak diğer dillerde de kullanımı mevcuttur (Brown ve ark., 2020). OpenAI, GPT modellerini araştırmacılar ve uygulama geliştiriciler için erişilebilir hâle getirdiğinden geniş bir kullanıcı grubu tarafından farklı alanlarda yaygın olarak kullanılmaktadır. GPT modellerinin bir türü olan ve bir sohbet robotu olarak tasarlanan ChatGPT, kullanıcı erişiminin kolaylığı neticesinde bazı çalışmalara konu edinmiştir. Örneğin Michel-Villarreal vd. (2023) ChatGPT'yi yüksek öğrenime sorumlu bir şekilde entegre etmek için açık politikalara, yönergelere ve çerçevelere ihtiyaç vurgulamakta ayrıca kullanıcı deneyimlerini ve algılarını anlamak için deneysel araştırmaların yapılması gerektiğinin altını çizmiştir. Talan ve Kalınkara (2023) ChatGPTnin bir sınavdaki soruları cevaplama yeteneğini, üniversite öğrencilerinin performansıyla karşılaştırdıkları çalışmada ChatGPTnin öğrencilere kıyasla daha yüksek doğru cevap oranına sahip olduğunu saptadı. Imran ve Almusharraf (2023) de ChatGPT'nin yazma asistanı olarak kullanıldıkları araştırmalarında akademik süreci kolaylaştırmak ve desteklemek açısından nispeten faydalı cihazlar olduğundan öğrenciler ve eğitmenler için kolaylaştırıcı bir rol üstlendiğini ifade etmektedir. Bununla birlikte her ne kadar çeşitli konuşma verilerinden yararlanarak sohbet ortamlarında insan benzeri yanıtlar üretmede başarılı olsa da ChatGPT'nin (Ivanov & Soliman, 2023) dil çevirisi, metin özetleme ve duygu analizi gibi konuşma yapay zekâsının ötesindeki işlemlerde iyi performans gösteremediği; ürettiği bilgilerin doğruluğu ve güvenirliği konusunda ve üretimlerde standartlaştırılmasının sağlanmasında endişelerin olduğu, sohbet robotu dışındaki işlemlerde GPT modellerinin kullanılması gerektiği belirtilmiştir (Athaluri ve ark., 2023; Floridi & Chiriatti, 2020). Başka bir anlatımla hem ChatGPT hem de GPT aynı temel dönüştürücü mimarisine dayansa ve metin oluşturma yeteneklerinde benzerlikler olsa da ChatGPT özellikle diyalogsal yapay zekâ uygulamaları için uyarlanmışken GPT modelleri doğal dil işlemede geniş bir potansiyel kullanım alanı sunmaktadır. Bu doğrultuda da araştırma kapsamında otomatik puanlama süreci GPT 3.5 modeli ile gerçekleştirilmiştir. GPT dil yardımı, çeviri, düzenleme ve redaksiyon dahil olmak üzere birçok fayda sağlama potansiyeline sahiptir (Jiao ve ark., 2023). Örneğin GPT dil bilgisi, sözdizimi, imla ve söz varlığı hakkında gerçek zamanlı geri bildirim sağlayarak öğrencinin dil yeterliliğini ve yazma becerisini geliştirebilir böylece yazılı üretimlerin genel kalitesini artırabilir (Bishop, 2023; Nagata ve ark., 2020). GPT'nin metin oluşturma becerisi, yazma görevlerini tamamlamadaki büyük potansiyelini göstermektedir (Stokel-Walker, 2022). GPT tarafından oluşturulan metinler, genel kalite, içerikte özgünlük (Yeadon ve ark., 2022) ve literatür tarama (Aydın ve Karaarslan, 2022) bakımından nitelikli bulunmaktadır. Ancak GPT ile oluşturulan metinlerin intihal olma durumu tartışılagelmektedir (Jabotinsky ve Sarel, 2022; Susnjak, 2022). Çünkü GPT, karmaşık veya teknik konularda anında metin oluşturma yeteneğine sahiptir (Lund ve Wang, 2023). Üstelik GPT, açıklayıcı cevaplar oluşturmada insana eşdeğer görülmektedir (Wenzlaff ve Spaeth, 2022). Genelde dil öğretiminde özelde ise yazma becerisinde GPT'nin işe koşulduğu çalışmaların sonuçlarının süreci planlama açısından yol gösterici olacağına inanılmaktadır. Bu çalışmada yapay zekâ bir değerlendirme aracı olarak kullanılmaktadır, yazma becerisini değerlendirmede puanlamanın ne kadar güvenilir olduğu konusunda bir bakış açısı sunmak araştırmanın özgün tarafını oluşturmaktadır.

Bu bağlamdan hareketle araştırmada üniversite öğrencilerinin yazılı üretimlerini değerlendirmede OpenAI tarafından geliştirilen GPT yapay zekâ sistemlerinin kullanılabilirliğini sınamak amaçlanmaktadır. Söz konusu amaç doğrultusunda cevap aranacak araştırma soruları şöyledir: (i) Uzmanların puanlarına ilişkin betimsel istatistikler nelerdir ve puanlayıcılar arası güvenirlik ne düzeydir? (ii) GPT puanlarına ilişkin betimsel istatistikler nelerdir ve GPT puanları arası güvenirlik ne düzeydir? (iii) GPT puanları ile uzmanların puanları arasında nasıl bir ilişki bulunmaktadır? (iv) GPT puanları, yazılı üretimlerin özellikleri bakımından nasıl bir değişiklik göstermektedir?

Yöntem

Araştırmanın Modeli

Bu araştırma ilişkisel tarama modelinde yürütülmüştür. İlişkisel tarama araştırmaları; nedensellik ilişkisi kurulmadan araştırmaya konu olan değişkenler arasında beklenen ilişkilerin test edilmesine ve tahminlerin yapılmasına izin vermektedir (Stangor ve Walinga, 2019). Bu araştırmada üniversite öğrencilerinin yazılı üretimlerini değerlendirmede OpenAI tarafından geliştirilen GPT yapay zekâ sistemlerinin kullanılabilirliğini sınamak amaçlandığı için katılımcıların yazılı üretimlerine verilen uzman puanları ile GPT tarafından üretilen puanlar karşılaştırılarak incelenmiştir.

Araştırma Grubu

Araştırma, 2023-2024 eğitim öğretim yılında İstanbul Üniversitesi Dilbilimi Bölümüne devam eden 60 öğrenci ile yürütülmüştür. Katılımcıların seçiminde Türkçeye yönelik farkındalığın yüksek olduğu bir bölümde okumaları, Yazılı Anlatım adlı bir ders almaları ve araştırmaya katılmaya gönüllü olmaları etkin olmuştur. Araştırma yapay zekâ destekli otomatik puanlamaya ilişkin bir uygulama örneği sunduğu için evren ve örneklem seçimine gidilmemiş, çalışma bir araştırma grubu ile yürütülmüştür. Araştırmaya katılan öğrencilerin tamamı birinci sınıf öğrencisi olup öğrencilerin 32'si (%53) kadın, 28'i (%47) erkektir. Öğrencilerin yaşı 18 ila 25 arasında değişiklik göstermektedir.

Veri Toplama Araçları

Bu araştırma kapsamında üniversite öğrencilerinden "Yazılı Üretim Görevi" doğrultusunda bir metin oluşturmaları istenmiştir. Oluşturulan metinler "Bütüncül Puanlama Anahtarı" ile puanlanmıştır.

Yazılı Üretim Görevi

Üniversite öğrencilerinin yazılı üretim ortaya koymaları için hazırlanan görev, Kaggle platformunda yer alan otomatik kompozisyon puanlama eğitim setinden seçilmiştir. Kaggle, veri setleri ve yarışmalar düzenleyerek veri bilimcilerin ve makine öğrenimi uzmanlarının birbirleriyle etkileşimde bulunmasını ve becerilerini geliştirmesini sağlayan, otomatik puanlama çalışmaları yürüten veri bilimi ve makine öğrenim platformudur (Banachewicz ve ark., 2022). Görevleri ve verileri açık erişimde olan platformun veri setleri ve görevleri birçok araştırmacı tarafından bilimsel çalışmalarda kullanılmaktadır (Ramachandran ve ark., 2015; Ramalingam ve ark., 2018; Ramesh ve Sanampudi, 2022; Yang ve ark., 2020). Bu araştırmada yapay zekâ tarafından gerçekleştirilen otomatik puanlamaların değerlendirilmesi amaçlandığından yazma görevi, Kaggle platformundan seçilmiştir. Seçilen görev, Türkçeye ve üniversite öğrencilerinin seviyesine uyarlanarak düzenlenmiştir. Yazılı üretim görevine ilişkin bilgiler Tablo 1'de belirtilmiştir.

Tablo 1'de görüldüğü gibi üniversite öğrencilerinden bilgilendirici bir metin üretmesi istenmiştir. İkna etme ve açıklama amaçlarıyla ele alınacak göreve yönelik 60 üniversite öğrencisi yazılı üretim ortaya koymuştur. Öğrencilerin üretimlerinin ortalama kelime sayısı 255'tir. Elde edilen yazılı üretimler, üç puanlayıcı tarafından değerlendirilmiştir. Puanlama, araştırmacıların yanı sıra bir Türkçe alan uzmanı tarafından gerçekleştirilmiştir. Sonrasında veriler GPT tarafından da puanlanmıştır. Her iki değerlendirmede de 1-6 aralığında bütüncül puanlama anahtarı kullanılmıştır. Katılımcılara verilen yazma görevinin içeriği Tablo 2'de gösterilmiştir.

Tablo 1.
Yazılı Üretim Görevinin Özellikleri.

Metin türü	Bilgilendirici
Metnin amacı	İkna etme ve açıklama
Öğrenci düzeyi	Üniversite öğrencisi
Uygulanan öğrenci sayısı	60
Ortalama kelime uzunluğu	255
Puanlayıcılar	Uzman1, Uzman2, Uzman3, GPT
Puanlama şekli	Bütüncül puanlama anahtarı
Puanlama aralığı	1-6

Tablo 2.
Yazılı Üretim Görevi.

Konu: Teknolojinin eğitime yansımasıyla birlikte tüm sınıfların bilgisayar laboratuvarı şeklinde düzenlenmesi; her öğrencinin sırasında bir bilgisayar olması yönünde bir görüş tartışılıyor. Ancak bu durumun öğrenciler için faydalı olup olmayacağı konusunda uzmanlar hemfikir değil. Sizden istenen aşağıdaki görüşlerden birini seçip muhatabınızı ikna eden yazılı bir metin hazırlamanız. Görüşler şunlardır:
* Bazı uzmanlar bilgisayarların öğrenciler ve öğrenme üzerinde olumlu bir etkisi olduğuna inanıyor. Onlara göre bilgisayarlar; öğrencilere uzak yerler ve insanlar hakkında bilgi edinme yeteneği verir, bilgiye ulaşmalarını kolaylaştırarak öğrenmeyi olumlu etkiler. Öğrencilerin okula devam etmelerini ve isteklerini artırır.
* Bazı uzmanlar ise öğrencilerin bilgisayarlarında çok fazla zaman harcayarak spor yapmaya, doğanın tadını çıkarmaya, aile ve arkadaşlarıyla, öğretmenleriyle etkileşimlerine daha az zaman ayırmalarından endişe duyuyor. Ayrıca bilgisayarların güvenilir olmayan birçok bilgiyi barındırdığını ve öğrencilerin derslere odaklanmada sorun yaşayabileceklerini belirtiyor.
Sen de şimdi tarafını seç ve okurlarını ikna edecek bir metin oluştur.

Araştırmaya katılan öğrencilerden 35'i (%58) birinci görüş olan sınıfların bilgisayar laboratuvarı şeklinde düzenlenmesini seçip yazılarını üretmiştir. 25 öğrenci de (%42) yazılarında ikinci görüşü seçip sınıfların bilgisayar laboratuvarı şeklinde düzenlenmesine karşı çıkan bir üretimde bulunmuştur.

Bütüncül Puanlama Anahtarı: Üniversite öğrencilerinin yazılı üretimlerini kullanmak için Kaggle platformunda yer alan puanlama anahtarı kullanılmıştır. Puanlama anahtarı öncelikle Türkçeye çevrilmiş, ardından ifadeler öğretim programında yer alan hususlar doğrultusunda düzenlenmiştir. Türkçe çeviri ve düzenlemelerden sonra puanlama anahtarına yönelik Türkçe eğitimi alanında öğretim üyesi olarak görev yapan dört konu alan uzmanının görüşü alınmıştır. Konu alan uzmanlarının görüşleri doğrultusunda puanlama anahtarındaki ifadeler yeniden gözden geçirilmiştir. Örneğin "Is awkward and fragmented." açıklaması bire bir çevirisi olan "Garip ve parçalıdır." yerine "Giriş, gelişme ve sonuç bölümleri yoktur ve yazının organizasyonu iyi değildir." şeklinde düzenlenmiştir. Ardından rastgele seçilen 6 öğrencinin yazılı üretimi, üç uzman tarafından uzman-paneli şeklinde değerlendirilmiştir. Puanlama sürecinde farklı anlaşılan ifadelerde gerekli müdahaleler yapılarak puanlama anahtarına son şekli verilmiştir. Puanlama anahtarı Tablo 3'tedir.

Tablo 3 incelendiğinde üniversite öğrencilerinin yazılı üretimleri değerlendirilirken öğrencilerden "konulardan birini tam olarak seçmesi, onu ayrıntı, örnek ve diğer destekleyici ifadelerle ele alması; giriş, gelişme ve sonuç bölümlerini organize etmesi, cümle ve paragraf bazında akıcılık sağlaması ve okurun dikkatini çekecek unsurlara yer vermesi" ölçütlerinin dikkate alındığı görülmektedir.

Verilerin Toplanması

Araştırma kapsamında veriler, 2023-2024 eğitim-öğretim yılının bahar döneminde üniversite öğrencilerinden çevrim içi olarak toplanmıştır. Araştırmaya katılım gönüllülük esasında olup https://iusinav.istanbul.edu.tr/#!/login platformunda öğrencilerin yazılı üretimlerini ortaya koyması istenmiştir. Katılımcıların hepsi aynı anda yazılı üretimlerine başlamış ve 60 dakika içerisinde tamamlamıştır. Araştırmanın verileri, Gazi Üniversitesi Etik Komisyonunun 07.11.2023 tarihli ve 19 sayılı kararı doğrultusunda yürütülmüştür.

Yazılı Üretimlerin Puanlanması

Uzman Puanlaması: Üniversite öğrencilerinin yazılı üretimleri öncelikle üç puanlayıcı tarafından değerlendirilmiştir. Puanlama dört aşamada gerçekleştirilmiştir. Birinci aşamada yazılı üretimler içinden rastgele seçilen 6 ürün; puanlayıcılar tarafından uzman-paneli şeklinde puanlanmıştır. Bu süreçte üç uzmanın da katıldığı panelde seçilen ürünler okunarak her bir uzman bağımsız bir şekilde puanlama yapmış ve puanının nedenini açıklamıştır. Bu aşamada farklılık gösteren puanlara yönelik puanlama anahtarında düzenlemelere gidilmiştir. İkinci aşamada rastgele 5 ürün daha seçilerek düzenlenen puanlama anahtarıyla ürünler yine bağımsız olarak puanlanmıştır.

Tablo 3.
Bütüncül Puanlama Anahtarı.

1 Puan: Hangi konuyu seçtiği tam olarak anlaşılamamış ya da seçtiği konuda yönergede verilen ifadeleri tekrar eden bir yazı üretmiştir. Bu yazıda; • Ayrıntı, örnek ve destekleyici ifadeler az sayıda yer almaktadır ya da tutarsızdır. • Giriş, gelişme ve sonuç bölümleri yoktur ve yazının organizasyonu iyi değildir. • Cümle ve paragraf bağlantıları kopuk olduğu için yazının okuması ve anlaşılması zordur. • Okurun dikkatini çekecek unsurlar yoktur.
2 Puan: Hangi konuyu seçtiği tam olarak anlaşılamamış ya da seçtiği konudaki yönergeden uzaklaşamayan yazı üretmiştir. Bu yazıda; • Ayrıntı, örnek ya da destekleyici ifadeler yer almamakta, görüşler genel ifadelerle sunulmaktadır. • Giriş, gelişme ve sonuç bölümleri arasındaki geçişler- metin organizasyonu- zayıftır. • Birçok cümle ve paragraf bağlantısında kopukluk vardır. • Okurun dikkatini çok az çekecek unsur vardır.
3 Puan: Konulardan birini tam olarak seçmiş ancak destekleyici ifade ve ayrıntılara yetersiz düzeyde yer veren bir yazı üretmiştir. Bu yazıda; • Ayrıntı, örnek ya da destekleyici ifadelere yeteri kadar yer verilmeden genel nedenler sunulmaktadır. • Giriş, gelişme ve sonuç bölümleri arasındaki geçişler- metin organizasyonu- basit düzeydedir. • Bazı cümle ve paragrafların bağlantısında kopukluk vardır. • Okurun dikkatini orta düzeyde çekecek unsurlar vardır.
4 Puan: Konulardan birini tam olarak seçmiş ancak destekleyici ifade ve ayrıntılara biraz yer veren bir yazı üretmiştir. Bu yazıda; • Ayrıntı, örnek ya da destekleyici ifadelere biraz yer almakla birlikte genel nedenler sunulmaktadır. • Giriş, gelişme ve sonuç bölümleri arasındaki geçişler- metin organizasyonu- orta düzeydedir. • Bazı cümle ve paragrafların bağlantısı akıcıdır. • Okurun dikkatini yeterli düzeyde çekecek unsurlar vardır.
5 Puan: Konulardan birini tam olarak seçmiş, destekleyici ifade ve ayrıntılara makul düzeyde yer veren bir yazı üretmiştir. Bu yazıda; • Ayrıntı, örnek ya da destekleyici ifadelere yeterince yer almakla birlikte bunlar herkesçe bilinen ifadelerdir. • Giriş, gelişme ve sonuç bölümleri arasındaki geçişler- metin organizasyonu- iyi düzeydedir. • Cümle ve paragraflar arasındaki geçiş, metin boyunca orta derecede akıcıdır. • Okurun dikkatini tüm metin boyunca çekecek unsurlar vardır.
6 Puan: Konulardan birini tam olarak seçmiş, destekleyici ifade ve ayrıntılara yeterli düzeyde yer veren özgün bir yazı üretmiştir. • Özgün ayrıntı, örnek ya da destekleyici ifadelere yeterli düzeyde yer almaktadır. • Giriş, gelişme ve sonuç bölümleri arasındaki geçişler- metin organizasyonu- çok iyi düzeydedir. • Cümle ve paragraflar arasındaki geçiş, metin boyunca çok iyi derecede akıcıdır. • Okurun dikkatini tüm metin boyunca artıracak unsurlar vardır.

Uzmanların puanları bire bir aynı olduğu için üçüncü aşamada 60 ürün uzmanlarca bağımsız olarak puanlanmıştır. Uzman paneli şeklinde gerçekleştirilen dördüncü aşamada da 60 öğrencinin her biri için nihai puana oy birliği ile karar verilerek puanlama süreci tamamlanmıştır.

Yapay zekâ puanlaması: Araştırmada uzman puanlamasının yanı sıra puanlama OpenAI GPT 3.5 tarafından da gerçekleştirilmiştir. GPT-3.5, eğitim ve dönüştürme adımlarını içeren "bilgi istemi" adı verilen belirli bir girdiye dayalı olarak sözcükler, kodlar veya veri dizileri üreten bir bilgi işlem sistemidir (Chen ve ark., 2021). GPT'ler gibi ön eğitim ve ince ayar içeren iki aşamalı bir süreçle eğitilir. Bu metodoloji, büyük ölçekli dil modelleri için yaygın olarak kullanılır ve belirli görevlerde veya alanlarda uzmanlaşmadan önce modelin genel dil anlayışı edinmelerine olanak tanır (Brown ve ark., 2020; Kamnis, 2023; Liu ve Gibson, 2023; Zhu ve ark., 2022). Ön eğitim aşamasında, GPT modelleri için büyük bir metin veri seti toplanır ve ön işleme adımları ile hazırlanır. Bu veri seti, kitaplar, makaleler, web sayfaları veya diğer metin belgeleri gibi çeşitli kaynaklardan elde edilebilir. Bu aşama, modelin dil bilgisi, sözdizimi, semantiği kavramasını ve önemli miktarda genel bilgi biriktirmesini sağlar (Kamnis, 2023). Ön eğitimden sonra GPT modelleri, daha küçük, göreve özgü veri kümelerinde ince ayar yaparak belirli görevlere veya etki alanlarına uyarlanır. İnce ayar, modelin istenen görevde uzmanlaşmasını ve ön eğitim sırasında öğrenilen genel dil anlayışını korurken alana özgü veriler üzerindeki performansını geliştirmesini sağlar. Örneğin, duygusal olarak etiketlenmiş veriler, bir duyarlılık analizi görevi için kullanılabilir (Kamnis, 2023). Görüldüğü gibi ilgili çıktıları sağlamak için modellerini eğitmek, büyük miktarda veri gerektirir ve GPT-3 modelleri, 175 milyar parametre ile eğitilmiştir. Bu nedenle GPT-3'ün özetleme, çeviri, dil düzeltme, soru yanıtlama, sohbet robotları, e-posta oluşturma, metin tamamlama, sınıflandırma, oluşturma, puanlama soru-cevap eşleştirme ve diğer dil tabanlı yapay zekâ uygulamalarında uygulamaları bulunmaktadır (Floridi ve Chiriatti, 2020; Zhang ve ark., 2021). Bu araştırmada ön eğitimi hâlihazırda 175 milyar parametre ile tamamlanan GPT 3.5'nin ikinci aşaması olan ince ayar ve istem özellikleri düzenlenmiştir. Çalışmada GPT-3.5 serisinin bir parçası olarak 30 Kasım 2022ıde yayınlanan OpenAI'nin text-davinci-003 modeli kullanılmıştır (Mizumoto ve Eguchi, 2023). Bunun için öncelikle kompozisyonlar CSV uzantılı bir dosyaya kaydedilmiştir. Ardından Python 3.11 yazılımı kullanılarak text-davinci-003 modeline emreden komut özelliği tanımlanmıştır. Komuttan sonra konu, puanlama anahtarı ve yazılı üretimler girdi olarak verilmiştir. Komut Python 3.11'de OpenAI API'si kullanılarak çalıştırılmıştır. Ayrıca GPT, token (karakter) tabanlı bir dil modelidir ve her bir giriş ve çıkış metni için belirli bir token sınırlaması bulunmaktadır. GPT-3.5ıe verilen bir komutun ardından, 2048ıden fazla token işlendiğinde modelin başlangıçtaki metni unutma olasılığı artmaktadır (Liu ve ark., 2023). Bu nedenle bir döngü komutu da tanımlanarak her 2000 tokenden sonra komut yeniden çalıştırılmıştır. Bu süreçte "eğer yazılı üretim token sayısı>maksimum token sayısı" özelliği de tanımlanarak işlem sıfırlanarak yazılı üretim yeniden değerlendirilmiştir. OpenAI'nin API'si aracılığıyla text-davinci-003'ü kullanmanın maliyeti 1.000 token başına 0,02 USD maliyeti olmuştur. 60 öğrenci tarafından ortaya konulan yazılı ürünler 5,26 dakika sonra puanlanarak sonuçlar CSV uzantılı bir dosyaya yazdırılmıştır.

Verilerin Analizi

Araştırmada öncelikle puanlayıcılardan elde edilen sonuçlara göre betimsel istatistikler (frekans, yüze, ortalama ve standart sapma) yapılmıştır. Ardından uzman puanlayıcılar arasında puanlayıcı güvenirliğinin hesaplanması için Pearson korelasyon ve Fleiss Kappa katsayıları hesaplanmıştır. Puanlayıcı güvenirliğinin yüksek bulunmasının ardından puanlamanın son aşamasında uzman-paneli gerçekleştirilerek her bir yazılı ürüne nihai bir puan verilerek bu puanların dağılımı için frekans ve yüzde değerleri hesaplanmıştır. Sonra GPT puanlarına ilişkin betimsel istatistikler hesaplanmıştır. GPT puanlarının güvenirliğini belirlemek için GPT tarafından üç farklı puanlama yapılmıştır. Puanlar arasındaki ilişkileri belirlemek için Pearson korelasyon katsayısı, puan ortalamaları arasında anlamlı farklılık olup olmadığını belirlemek için de bağımlı ölçümlerde t testi ile Wilcoxon işaretler testine başvurulmuştur. GPT tarafından gerçekleştirilen puanlar ile uzmanların puanları karşılaştırmak için çapraz tablodan ve histogram grafiğinden yararlanılmıştır. Puanlar arasındaki ilişkilerin belirlenmesi amacıyla Pearson korelasyon katsayısı hesaplanmıştır. Puan ortalamaları arasında anlamlı bir fark olup olmadığının belirlenmesi için bağımlı ölçümlerde t testi kullanılmıştır. Daha sonra GPT puanları ile yazılı üretimlerin metinsel özellikleri (karakter, harf, kelime ve cümle sayısı ile okunabilirlik değerleri) arasındaki ilişkilerin incelenmesi amacıyla Pearson korelasyon katsayısı hesaplanmıştır. Yazılı üretimlerin savunduğu görüşe göre GPT puanlarının anlamlı bir değişiklik gösterip göstermediğinin belirlenmesi için Mann Whitney U testi kullanılmıştır. Ayrıca GPT puanları ile uzmanların puan farkının 2 olduğu ürünler uzmanlarca betimsel olarak incelenmiştir. Metinlerin okunabilirlikleri Ateşman tarafından Türkçeye uyarlanan okunabilirlik değeri ile hesaplanmıştır.

Araştırma kapsamında gerçekleştirilen puanlara yönelik çarpıklık ve basıklık değerleri hesaplanmıştır. GPT 3.5 okuma dışındaki puanlarda veri sayısı>30 olduğundan ve çarpıklık basıklık değerleri ±1 arasında değer aldığı için (Büyüköztürk, 2018) söz konusu puanların karşılaştırılmasında parametrik istatistikler kullanılmıştır. Bununla birlikte GPT'nin üçüncü puanlamasına yönelik karşılaştırmalarda parametrik olmayan istatistiklerden yararlanılmıştır. Ayrıca birinci görüşü ele alan yazılı üretimlerde GPT puanları sola çarpık bir dağılım gösterdiği (birinci görüş çarpıklık: -1.465; basıklık: 2.329) için karşılaştırmada parametrik olmayan istatistikten faydalanılmıştır. Araştırmada GPT tarafından puanlar Python 3.11'de OpenAI API'si kullanılarak gerçekleştirilmiştir. Ayrıca yazılı üretimlerin karakter, harf, kelime ve cümle sayısı ile okunabilirlik değerleri Python 3.11 kullanılarak gerçekleştirilmiştir. Betimsel istatistik, ilişki ve

fark testleri için SPSS 26.0 programından yararlanılmıştır.

Bulgular

Uzmanların Puanlarına İlişkin Betimsel İstatistikler ve Puanlayıcılar Arası Güvenirlik

Araştırmada kapsamında öncelikle üç uzman tarafından gerçekleştirilen puanların dağılımı belirlenmiş ve sonuçlar

Tablo 4.
Uzmanların Puanlarına İlişkin Frekans Değerleri.

Puanlar	Uzman1	Uzman2	Uzman3
1	1	1	1
2	4	3	3
3	13	12	8
4	14	14	18
5	20	19	17
6	8	11	13
Ortalama (SS)	4.2 (1.2)	4.3 (1.2)	4.4 (1.2)

Tablo 4'te gösterilmiştir.

Tablo 4'te görüldüğü gibi Puanlayıcı 1'in 60 üniversite öğrencisinin yazılı anlatım ürünlerini değerlendirdiği puanların ortalaması 4.2 (1.2); Puanlayıcı 2'nin 4.3 (1.2) ve Puanlayıcı 3'ün 4.4 (1.2)'tür. Puanlayıcı güvenirliğinin belirlenmesi amacıyla öncelikle Pearson korelasyon katsayısı

Tablo 5.
Uzmanların Puanları Arası İlişkiler.

Puanlayıcı	Uzman1	Uzman2	Uzman3
Uzman1	1	0.927**	0.893**
Uzman2		1	0.913**
Uzman3			1

hesaplanmış ve sonuçlar Tablo 5'te sunulmuştur.

Tablo 5 incelendiğinde puanlayıcıların üniversite öğrencilerinin yazılı anlatım ürünlerine vermiş olduğu puanlar arasındaki ilişkiler 0.893-0.927 arasında değişiklik göstermektedir (p=0.00). Hesaplanan korelasyon katsayısı puanlar arasındaki ilişkilerin oldukça yüksek olduğunu göstermektedir (Taylor, 1990). Bununla birlikte hem korelasyona dayalı puanlayıcı güvenirliği hesaplamalarında tesadüfe dayalı uyumluluklar hesaplanmadığı için (Bilgen ve Doğan, 2017) hem de üç puanlayıcı tarafından elde edilen sonuçların aynı anda değerlendirilmesi de amaçlanmıştır. Bu nedenle araştırmada puanlayıcı güvenirliğini belirlemek için ayrıca Fleiss Kappa katsayısı hesaplanmış ve elde edilen sonuçlar Tablo 6'da gösterilmiştir.

Tablo 6'ya göre üç uzmanın puanları arasında önemli düzeyde (>0.60) bir uyum olduğu görülmektedir (Fleiss ve ark., 1979). Uzmanlardan elde edilen puanlar arasında önemli bir uyum olduğunun belirlenmesinin ardından uzman paneli düzenlenmiş ve uzmanlar her bir yazılı ürün için nihai bir puan belirlemişlerdir.

GPT Puanlarına İlişkin Betimsel İstatistikler ve Güvenirlik

Araştırmada GPT 3.5 serisine ait text-davinci-003 modeli ile öncelikle 60 öğrencinin yazılı üretimleri puanlanmıştır.

Tablo 7.
GPT Puanlarına İlişkin Frekans Değerleri.

Puanlar	GPT
1	1
2	5
3	8
4	12
5	30
6	4
Ortalama (SS)	4.3 (1.2)

Ulaşılan sonuçlar Tablo 7'de verilmiştir.

Tablo 7 incelendiğinde GPT tarafından 60 yazılı ürünün puan ortalaması 4.3 (1.2)'tür. Yazılı üretimlerin 30'u (%50) GPT tarafından 5 olarak puanlanmıştır. Bir öğrencinin yazılı üretimi 1; dört öğrencinin yazılı üretimi de GPT tarafından 6 olarak puanlanmıştır. GPT tarafından gerçekleştirilen puanların güvenirliği iki aşamada incelenmiştir. Birinci aşamada 60 öğrencinin yazılı üretimlerinin csv formatındaki veri dosyasındaki sırası değiştirilmiştir. GPT tarafından birinci ve ikinci değerlendirmeler arasında anlamlı bir farklılık olup olmadığının tespit edilmesi için bağımlı t testi hesaplanmış ve ortalamaların anlamlı bir farklılık göstermediği belirlenmiştir (r= 0.892; $t_{(59)}$=0.241; p=0.811). GPT puanlarının güvenirliğini belirlemek için ikinci aşamada 60 öğrenci içinden rastgele 10 öğrencinin yazılı üretimi seçilmiştir. Bu yazılı üretimlerin sırası da değiştirilerek komutlar yeniden çalıştırılarak GPT tarafından ikinci puanlama yapılmıştır. 10 öğrencinin yazılı üretimi için GPT tarafından gerçekleştirilen puanlar

Tablo 6.
Uzmanların Puanları Arası Uyum- Fleiss Kappa Katsayısı.

	Katsayı	SE	Z	p	En düşük 95% güven sınırı	En yüksek 95% güven sınırı
Fleiss Kappa	.646	.040	16.351	.000	.569	.724

arasında da Wilcoxon işaretler testine göre anlamlı bir farklılık olmadığı saptanmıştır (r=0.747; z= -0.447; p=0.655). Her iki aşamada da GPT ile yapılan puanlamaların ilk puanlarla yüksek korelasyon göstermesi ve ortalamalar arası farklılık olmaması, puanların güvenirliğinin yüksek olduğunu göstermektedir.

GPT Puanları ile Uzman Puanlarının Karşılaştırılması

GPT 3.5 puanları ile uzman puanlarının karşılaştırılmasını

Tablo 8.
GPT 3.5 Puanları ile Uzmanların Puanlarının Dağılımı.

Puanlar		GPT Puanlaması (4.3±1.2)					
		1	2	3	4	5	6
Uzman puanlaması (4.3±1.2)	1	1	0	0	0	0	0
	2	0	0	2	1	0	0
	3	0	2	4	1	3	0
	4	0	2	2	4	8	0
	5	0	1	0	5	11	2
	6	0	0	0	1	8	2

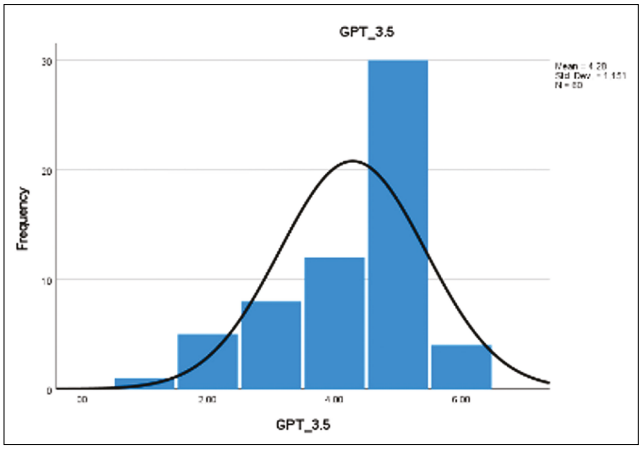
içeren frekans dağılımı Tablo 8'de ifade edilmiştir.

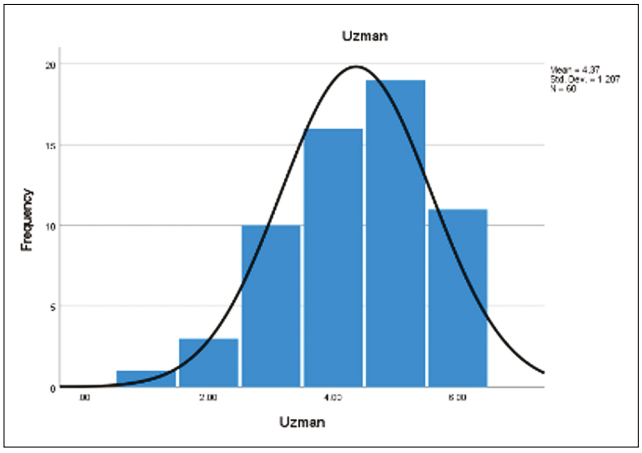
Tablo 8'de görüldüğü gibi hem uzmanlarca hem de GPT tarafından bir öğrencinin yazılı üretimi 1 olarak puanlanmıştır. Bununla birlikte uzmanlarca üç ürün 2 olarak puanlanmıştır. GPT tarafından bu ürünlerden ikisi 3; biri de 4 olarak puanlanmıştır. Benzer şekilde on bir ürün 6 olarak puanlanırken GPT tarafından bu ürünlerden 2 tanesi olmak üzere toplam 4 yazılı ürün tam puan almıştır. Puanlar grup bazında incelendiğinde hem uzmanların hem de GPT puanlarının ortalamasının 4.3 olarak hesaplandığı görülmektedir. GPT ve uzmanların puanları arasındaki ilişkilerin belirlenmesi amacıyla Pearson korelasyon katsayısı hesaplanmıştır. Ayrıca puan ortalamaları arasında anlamlı bir fark olup olmadığının tespiti için de bağımlı ölçümlerde t testi hesaplanmış ve ulaşılan sonuçlar Tablo 9'da gösterilmiştir.

Tablo 9'da görüldüğü gibi uzmanların puanları ile GPT'nin puanları arasında pozitif yönde ve orta düzeyde bir ilişki bulunmaktadır (r=0.638; p=0.00). GPT puanları ile uzmanların puan ortalamaları arasında da anlamlı bir farklılık olmadığı belirlenmiştir ($t_{(599=0.265;\ p=0.792)}$). GPT puanları ile uzman puanları grup bazında incelendikten sonra birey bazında inceleme gerçekleştirilmiştir. Puan dağılımlarına ilişkin histogram grafiği oluşturulmuş ve Şekil 1'de gösterilmiştir.

Şekil 1' de görüldüğü gibi GPT ve uzman puanlamasının modu 5'tir. Ancak bu değer uzman puanlamasında 19 ürün iken GPT değerlendirmesinde 30 üründür. Hem uzman hem de GPT tarafından sadece bir tane yazılı ürüne 1 puan verilmiştir. GPT ile uzman puanları arasındaki farklara ilişkin frekans ve yüzde değerleri Tablo 10'da gösterilmiştir.

Şekil 1.
GPT 3.5 Puanları İle Uzmanların Puanlarının Dağılımı





Tablo 9.
GPT ile Uzman Puanları Arasındaki İlişki ile Ortalamaları Arasındaki Fark Testi Sonuçları.

Puanlama	r	Ortalama fark	SS	SE Ort.	95% güven aralığı		t	sd	p
					En Düşük	En Yüksek			
GPT-Uzman	0.667**	-.033333	.97366	.12570	-.28486	.21819	-.265	59	.792

**p<0.01

Tablo 10.
GPT 3.5 Puanları ile Uzmanların Puan Farklarının Dağılımı.

Puanlar Arası Fark (Mutlak)	f	%
0	22	37
1	32	53
2	6	10
3	0	0
4	0	0
5	0	0

Tablo 10'a göre 60 üniversite öğrencisinin katıldığı araştırmada GPT ile uzmanlar, yazılı ürünlerin 22'sine (%37) aynı puanı vermiştir. 32 yazılı üründe (%53) GPT ile uzmanlar arasında 1 puan; 6 üründe (%10) 2 puan fark olduğu belirlenmiştir. GPT ile uzman puanı arasında 3, 4 veya 5 puan fark olan bir yazılı ürün bulunmamaktadır.

GPT Puanlarının Yazılı Üretim Özellikleri Açısından İncelenmesi

GPT puanlarının yazılı ürünlerin metinsel özelliklerine göre incelenmesi: Üniversite öğrencilerinin yazılı üretimlerinde GPT puanları, yazılı üretimlerin metinsel özellikleri bakımından incelenmiştir. Bu doğrultuda metinlerin karakter, harf, kelime ve cümle sayısı ile üretimlerin okunabilirlik değerleri hesaplanmıştır. GPT puanları ile metinsel özellikler arasındaki ilişkilerin belirlenmesi amacıyla Pearson korelasyon katsayısı hesaplanmıştır. Hesaplama sonucunda GPT puanları ile yazılı üretimlerin karakter sayısı (r=0.262; p=0.043), hece sayısı (r=299; p=0.020), cümle sayısı (r=0.353; p=0.006) arasında pozitif yönde ancak düşük düzeyde ilişkiler olduğu belirlenmiştir. Bununla birlikte GPT puanları ile metinlerdeki kelime sayısı (r=0.222; p=0.089) ve metinlerin okunabilirlik değerleri (r=0.221; p=0.090) arasında anlamlı ilişkiler olmadığı saptanmıştır.

GPT puanlarının yazılı üretimlerin ele aldığı görüşe göre incelenmesi: Araştırmaya katılan öğrencilerden 35'i (%58) birinci görüş olan sınıfların bilgisayar laboratuvarı şeklinde düzenlenmesini seçerek yazılarını üretmiştir. 25 öğrenci de (%42) yazılarında ikinci görüşü seçerek sınıfların bilgisayar laboratuvarı şeklinde düzenlenmesine karşı çıkan bir görüş ele almıştır. GPT puanlarının yazılı üretimin savunduğu görüşe göre anlamlı bir farklılık gösterip göstermediğinin belirlenmesi amacıyla Mann Whitney U testi yapılmıştır. Hesaplama sonucunda birinci görüşü savunan yazılı üretimlerin (sıra ort: 34.19), ikinci görüşü savunan üretimlerden (sıra ort: 25.34) anlamlı bir şekilde daha yüksek puanlara sahip olduğu tespit edilmiştir (U=308.500; z=2.081; p=0.037). Uzmanların puanlarında ise savunulan görüşe göre puanlar anlamlı bir farklılık göstermemektedir (U=386.500; z=0.787; p=0.431).

GPT puanlarının uzman görüşüne göre incelenmesi: Araştırmaya katılan 54 öğrencinin yazılı ürünlerinin değerlendirilmesinde GPT tarafından gerçekleştirilen puanlama ile uzmanlar tarafından yapılan puanlama arasında hiç fark olmadığı ya da 1 puan fark olduğu belirlenmiştir.

Tablo 11.
GPT Puanlarına İlişkin Uzman Görüşleri.

GPT Puanı	Uzman Puanı	Fark	Görüş	Açıklama
2	4	2	2	43 kodlu ürün: Bu yazılı üretim tez-antitez şeklinde sınıfların bilgisayar laboratuvarları şeklinde düzenlenmemesini ele alınmıştır. Bu nedenle GPT "konunun tam olarak seçilmediğini" belirttiği için bu puanı vermiştir. Ancak üretim yeniden değerlendirildiğinde yazının kendi içinde tutarlılığı olduğu belirlenmiştir.
2	4	2	2	51 kodlu ürün: Bu yazılı üretim tez-antitez şeklinde sınıfların bilgisayar laboratuvarları şeklinde düzenlenmemesini ele alınmıştır. Bu nedenle GPT "konunun tam olarak seçilmediğini" belirttiği için bu puanı vermiştir. Ancak üretim yeniden değerlendirildiğinde yazının kendi içinde tutarlılığı olduğu belirlenmiştir.
5	3	2	1	11 kodlu ürün: Bu yazılı üretimde giriş ve gelişme bölümleri bulunmakla birlikte yazının sonuç bölümü iyi organize edilmemiştir. Destekleyici ifadelere yer verilmekle birlikte cümle ve paragraf geçişlerinde kopukluklar bulunmaktadır. GPT, metin organizasyonundan ziyade ana fikri destekleyen unsurlara dikkat ederek bu puanı vermiş olabilir.
5	3	2	1	60 kodlu ürün: Bu yazılı üretimde giriş ve gelişme bölümleri bulunmakla birlikte yazının sonuç bölümü iyi organize edilmemiştir (bulunmamaktadır). GPT, metin organizasyonundan ziyade ana fikri destekleyen unsurlara dikkat ederek bu puanı vermiş olabilir.
5	3	2	1	36 kodlu ürün: Bu ürün konuyu belirlemiş olmasına karşın cümlelerde anlatım bozuklukları, cümleler arası kopukluklar ve tutarsızlıklar olduğu için metin akıcılıktan uzaktır. Ancak GPT ürünün giriş, gelişme ve sonuç bölümlerinin iyi organize edildiğini ve yeterince destekleyici ifadeye yer verildiğini belirttiği için 5 puan vermiştir.
4	2	2	1	23 kodlu ürün: Bu ürün, kendi içinde giriş, gelişme ve sonuç bölümleri barındırarak genel olarak bir akıcılığa sahip olsa da verilen yönergedeki ana fikre yer vermemektedir. Bilgisayarın yararlarına odaklanan ürün sınıfların bilgisayar laboratuvarı olmasına yönelik bilgiler içermemektedir. Bununla birlikte GPT; anahtar kelime tarayarak metnin akıcı olduğunu ve organizasyonunun bulunduğu için 4 puan verdiği belirtmiştir.

Bununla birlikte 6 öğrencinin yazılı ürününde GPT tarafından gerçekleştirilen puanlama ile uzmanlar tarafından yapılan puanlama arasında 2 fark olduğu tespit edilmiştir. Söz konusu 6 öğrenciye ilişkin yazılı ürünler ve puanlar, puanlama yapan üç uzman tarafından incelenmiş, sonuçlar Tablo 11'de sunulmuştur.

Tablo 11 incelendiğinde GPT'nin uzmanlardan 2 puan daha düşük puan verdiği öğrencilerin ikinci görüşü savunduğu belirlenmiştir. GPT puanlama gerekçesinde "konunun tam olarak seçilmediğini" ifade eden bir açıklama yazmıştır. Metinler yeniden incelendiğinde öğrencilerin görüşlerini tez-antitez şeklinde ele aldığını, karşılaştırmalar yaparak görüşlerini açıkladıkları görülmüştür. GPT'nin uzmanlardan 2 puan daha fazla verdiği yazılı üretimlerde birinci görüşün savunulduğu belirlenmiştir. Bu metinlerde metin organizasyonu ve bağdaşıklıktan kaynaklı sorunlar varken GPT puanlama gerekçesinde bu metinlerin "akıcı" olduğunu ifade etmiştir. Örneğin 60 kodlu üründe öğrenci, muhtemelen süresinin yetmemesi nedeniyle, üretimini cümlenin yarısında bırakmış ve metninin sonuç bölümünü yazmamıştır. Giriş ve gelişme bölümleri kendi içinde destekleyici ifadeler barındırsa ve akıcı da olsa metin bir bütün olarak ele alındığında eksiktir. Bununla birlikte GPT söz konusu ürünü 5 olarak puanlamıştır.

Tartışma

Bu araştırmada üniversite öğrencilerinin yazılı üretimlerini değerlendirmede OpenAI tarafından geliştirilen GPT yapay zekâ sistemlerinin kullanılabilirliği araştırılmıştır. Bu doğrultuda 60 üniversite öğrencisinden yazılı bir ürün ortaya koyması istenmiştir. Yazılı üretimler üç uzman ve GPT tarafından puanlanmıştır. Araştırmada puanlar karşılaştırılarak incelenmiş ayrıca GPT puanları da farklı açılardan değerlendirilmiştir. Araştırmada ilk olarak üç uzmanın puanlarının güvenirliği incelenmiş ve uzman puanlarının önemli düzeyde uyum gösterdiği belirlenmiştir. Ardından uzmanlarca her bir yazılı üretime nihai bir puan verilmiştir. Araştırmanın ikinci aşamasında katılımcıların yazılı üretimleri, GPT tarafından puanlanmıştır. GPT puanlarının güvenirliğini belirlemek için GPT'ye üç farklı puanlama yaptırılmış ve puan ortalamaları arasında anlamlı bir farklılık olmadığı belirlenmiştir. Bu durumda; karşılaştırmada uzmanlarca belirlenen nihai puanlar ile GPT tarafından gerçekleştirilen ilk puanlama sonuçları kullanılmıştır. Araştırma sonucunda GPT ile uzmanların puanları arasında pozitif yönde ve orta düzeyde bir ilişki olduğu saptanmıştır ($r=0.667$; $p<0.01$). Aynı zamanda 1-6 aralığında yapılan puanlamada GPT puanları ile uzman puanları arasındaki en yüksek farkın 2 olduğu ve 2 puanlık farkın 6 yazılı üründe (%10) olduğu belirlenmiştir. Böyle bir tabloda, yazma performansını değerlendirmede GPT puanlaması umut verici bir sonuç içerdiği söylebilir. Araştırma kapsamında elde edilen sonucun öğrenme çıktılarını iyileştirmek, öğrenci katılımını artırmak ve öğrenci görevlerini otomatikleştirerek eğitim süreçlerini geliştirme bağlamında GPT'nin kullanılabilirliğine yönelik yapılan araştırma sonuçlarıyla örtüştüğü anlaşılmaktadır (Baidoo-Anu ve Owusu Ansah, 2023; Jalil ve ark., 2023; Mhlanga, 2023). Bununla beraber GPT'nin yazılı ürünlerin otomatik puanlanmasına yönelik alan yazında henüz sınırlı sayıda çalışma bulunmaktadır. Bu alanda ilk çalışma Mizumoto ve Eguchi (2023) tarafından gerçekleştirilmiş, TOEFL kapsamında yazılmış 12.100 yazılı üretimi GPT-3 text-davinci-003 modeli ile puanlamışlardır. Bireylerin ikinci dilde ortaya koyduğu ürünlerin değerlendirilmesi için GPT'nin otomatik puanlama için umut verici bir araç potansiyeline sahip olduğu tespit edilmiştir.

Araştırma kapsamında ayrıca GPT tarafından verilen puanlar incelenmiştir. GPT puanlarının yazılı üretimlerin karakter, harf, kelime, cümle sayısı ve okunabilirlik değerleri arasında anlamlı ilişkiler olmadığı ya da ilişkilerin düşük düzeyde olduğu belirlenmiştir. Bu durumun üniversite öğrencilerinin yönergeye uyarak 200 kelimeden oluşan metinler oluşturmasından kaynaklı olabileceği düşünülmektedir. Çünkü yazılı üretimlerin bu tür yüzeysel yönleri ile metnin niteliğinin değerlendirilmesi arasındaki ilişkinin incelendiği araştırmalar hem eğitimsel ölçüm hem de dil teknolojisi alanında metin uzunluğunun, metinleri otomatik olarak puanlamak için kullanılan bilgisayar algoritmalarının yanı sıra insan değerlendiriciler tarafından yapılan metin derecelendirmelerini güçlü bir şekilde etkilediğini göstermektedir (Chodorow ve Burstein, 2004; Guo ve ark., 2013; Powers, 2005). Crossley (2020), metin uzunluğunu dilsel bir özellik olarak dikkate almadığını ancak metin uzunluğunun yazma gelişimi ve niteliğinin en güçlü yordayıcısı olduğunu kabul ettiğini söylemektedir. Bu da alanyazındaki araştırmaların metin uzunluğu ile insanların metin kalitesine ilişkin derecelendirmeleri arasında pozitif bir ilişki bulunmasının gerekçesini açıklamaktadır (Chenoweth ve Hayes, 2001; McNamara ve ark., 2015). Metin uzunluğu ile insan puanları arasındaki ilişkinin, metin uzunluğu ile metnin niteliği arasındaki ilişkiyi yansıtma durumu veya bunun puanlayıcı yanlılığından mı kaynaklandığı belirsizliğini korumaktadır. İlki, metin uzunluğunun yapıyla ilgili bir unsur olduğunu ve makale isteminde sunulan konu hakkında etkili bir bakış açısı geliştirmek için belirli bir uzunluğa ihtiyaç duyulduğunu ve bunun puanlamada dikkate alınan yönlerden biri olduğunu ileri sürerken (Quinlan ve ark., 2009) ikincisi metin uzunluğunun yazma yeterliliğinin yapısıyla tamamen veya kısmen alakasız olduğunu ve insan muhakemesi üzerindeki güçlü etkisinin bir önyargı olarak değerlendirilebileceğini iddia etmektedir (Powers, 2005). Deane (2013) akıcı yazmanın yapıyla ilgili bir unsur olarak görülmesi gerektiğini ifade etmektedir. Metin niteliğinin analitik derecelendirmesinde Attali (2016) daha uzun metinlerin genellikle daha uyumlu araçlar içerdiğini ve bunun da metin kalitesi derecelendirmeleri üzerinde olumlu bir etkiye sahip olduğunu belirlemiştir. Özetle uzun metinlerin bağdaşıklık ve tutarlılık açısından daha nitelikli metinler ortaya koyduğu dile getirilebilir.

GPT'nin, uzmanlardan 2 puan daha fazla verdiği yazılı

üretimlerde bağdaşıklık problemi olduğu belirlenmiştir. Bağdaşıklık, bir yapıya metin olma hüviyeti kazandıran, metin içi ilişkileri kuran dille ilgili özelliklerin bütünü şeklinde tanımlanmaktadır (Günay, 2017). Bağdaşıklık; gönderim, değiştirim, eksiltme, bağlama ögeleri ve kelime bağdaşıklığı olmak üzere beş ögenin çatı kavramı olarak düşünülmektedir. Metnin tutarlılığı için bağdaşıklık ön koşuldur. Araştırmalar, üniversite öğrencilerinin bağdaşıklık unsurlarını kullanma noktasında bazen eksikleri olduğunu ortaya koymaktadır (Göçer, 2010; Kalı, 2016; Seçkin ve ark., 2014). Crossley (2020) daha yüksek puan alan yazıların genellikle daha karmaşık sözcüksel öğeler ve sözdizimsel özellikler gösterdiğini ayrıca daha fazla uyum içerdiğini söylemekte böylece metinlerdeki dilsel özelliklerin yazma kalitesi ve gelişimi hakkında önemli bilgiler verebileceğine dair güçlü göstergeler sunmaktadır. Bu durum, araştırma kapsamında ulaşılan sonuç ile benzerlik göstermektedir. GPT' ile insan puanlayıcılar arasında değerlendirmede fark olmasının nedeni, Türkçenin dil yapısına hâkimiyet noktasında yapay zekânın eksik kalması şeklinde yorumlanabilir. Araştırmada öğrencilere bir konu ile ortaya konulan iki farklı görüş sunulmuş ve öğrencilerden bu görüşlerden birini ele alan bir yazılı metin üretmeleri istenmiştir. Birinci görüş olan "tüm sınıfların bilgisayar laboratuvarı şeklinde düzenlenmesi gerektiği"ni belirten öğrenci puanlarının; "sınıfların bilgisayar laboratuvarı şeklinde düzenlenmemesi gerektiği"ni savunan öğrencilerden daha yüksek olduğu saptanmıştır. Bu durumun nedeni uzman görüşüne göre incelendiğinde ikinci görüşü savunan öğrencilerin ana fikri ortaya koymak için tez- antitez yöntemiyle yazılarını oluşturmalarıyla ilişkili olduğu tespit edilmiştir. Sıralarda bilgisayar olmasının yanı sıra bilgisayarın neden daha fazla zarar vereceğini açıklayan bu görüşlerde GPT, konuya tam olarak karar verilemediğini ifade etmiş ve öğrencilere daha düşük puan verme eğiliminde olmuştur. Bu durumun katılımcıların tartışmacı metin türüne yönelmesiyle ilgili olduğu düşünülmektedir. Tartışmacı metin, düşünceyi geliştirme yollarını kullanarak yazarın bir konu hakkında öne sürdüğü iddiasını veri ve gerekçelerle destekleyip karşı iddiaları çürütmeye çalıştığı böylece konuyu sonuca bağladığı bir metin türüdür (Coşkun ve Tiryaki, 2011). Bir anlamda tartışmacı metin, muhatabı ikna etme, bunun için gerekçeler oluşturmayı kapsamaktadır. GPT'nin puanlamasında metin türünün özelliklerinden kaynaklı değişkenleri sürece dahil edememesinin etkili olduğuna inanılmaktadır.

Yazma değerlendirmesi, yazılı üretimin yazma kalitesini gösteren özelliklerinin belirlenmesi ve değerlendirilmesi ile ilgilidir. Bu araştırmada yazılı üretimi değerlendirmede uzman puanlaması ile yapay zekâ için GPT puanlaması kullanılmıştır. Öğrencilerin yazılı üretimlerine puan vermek zor bir görevdir çünkü metinden gelen farklı ögeler değerlendiricilerin veya öğretmenlerin kararlarının doğruluğunu etkileyebilmektedir (örneğin, el yazısı, imla: Graham ve ark., 2011; uzunluk, sözcük çeşitliliği: Wolfe ve ark., 2016). Değerlendirilecek yapıya bağlı olarak bu yönlerin etkisi, yargılama yanlılığı olarak ifade edilebilir. Ayrıca yorgunluk ve puanlama anahtarındaki ifadelerin farklı yorumlanması, değerlendirme sonuçlarının güvenirliğini olumsuz etkilemektedir (Hussein ve ark., 2019). Bununla birlikte yazılı üretimlerin uzmanlar tarafından değerlendirilmesi zaman alıcıdır ve oldukça yoğun bir emek gerektirir. Araştırmada elde edilen sonuçlar, OpenAI tarafından geliştirilen GPT 3.5 modelinin öğrencilerin yazılı üretimlerinin değerlendirilmede kullanım potansiyeline olduğunu göstermektedir. Ayrıca puanlama gerekçelerinin de üretilmesi, öğrencilere geri bildirim verme sürecinde de kolaylıklar sağlamaktadır. Bu doğrultuda GPT'nin yazılı üretimlerin değerlendirilmesinde ön kanıt sağlamak için kullanılması önerilmektedir. Bununla birlikte puanların uzmanlarca hâlen kontrol edilmesine ihtiyaç olduğu da bir gerçektir. GPT puanlarını farklı yönlerde detaylı bir şekilde inceleyebilmek amacıyla bu araştırmada 60 öğrenci tarafından üretilen yazılı üretimler değerlendirilmiştir. Araştırma sonuçlarından hareketle daha geniş örneklem büyüklüklerinde çalışma yinelenerek genellenebilir sonuçlara ulaşılabilir. Uzman ve GPT sonuçlarının yanı sıra makine öğrenmesine dayalı otomatik puanlamalar da yapılarak sonuçlar karşılaştırılabilir. Bu araştırmada GPT puanları yazılı metinlerin metinsel özellikleri, savunduğu görüş ve uzman görüşlerine göre incelenmiştir. Metinlerin dilsel özellikleri gibi farklı özellikler ve ileri düzey istatistiklerle değerlendirilebilir.

Yazarların Katkısı

Birinci yazar; araştırmacının tasarımı, veri toplama aracının düzenlenmesi, yazılı üretimlerin insan puanlanması, yapay zekâ destekli otomatik puanlama, yöntem, analiz ve bulgular sürecinde rol oynamıştır.

İkinci yazar; literatür taraması, veri toplama aracının düzenlenmesi, verilerin toplanması, yazılı üretimlerin insan puanlaması, sonuçların tartışılması ve yorumlanması sürecinde rol oynamıştır.

Teşekkür

Bu araştırmada üniversite öğrencilerinin yazılı üretimlerinin puanlanma sürecine katkı sunan Sena Kızılcakaya'ya teşekkür ederiz.

Kaynakça

Adams, D. ve Chuah, K.M. (2022). Artificial intelligence-based tools in research writing: current trends and future potentials. *Artificial Intelligence in Higher Education*, 169-184.

Ashton, R. (2007). The write skills: Rob Ashton looks at the challenges of improving graduates' business writing skills. *Training Journal*, 33-38.

Athaluri, S., Manthena, S., Kesapragada, V., Yarlagadda, V., Dave, T., & Duddumpudi, R. (2023). Exploring the boundaries of reality: investigating the phenomenon of artificial intelligence hallucination in scientific writing through chatgpt references. *Cureus*. https://doi.org/10.7759/cureus.37432

Attali, Y. (2016). A comparison of newly-trained and experienced raters on a standardized writing assessment. *Language Testing*, *33*(1), 99–115. https://doi.org/10.1177/0265532215582282

Aydın, Ö. ve Karaarslan, E. (2022). Penaı chatgpt generated literature review: digital twin in healthcare. In Ö. Aydın (Ed.), *Emerging Computer Technologies 2* (pp. 22-31). İzmir Akademi Derneği.

Bacon, D. R. ve Anderson, E. S. (2004). Assessing and enhancing the basic writing skills of marketing students. *Business Communication Quarterly*, *67*(4), 443-455.

Baidoo-Anu, D. ve Owusu Ansah, L. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Available at SSRN 4337484*.

Banachewicz, K., Massaron, L. ve Goldbloom, A. (2022). *The Kaggle Book: Data Analysis And Machine Learning For Competitive Data Science*. Birmingham: Packt Publishing Ltd.

Barker, R. T. ve Hall, B. S. (1995). Using the business briefing to develop oral communication skills. *Journal of Management Education*, *19*(4), 513-518.

Benjamin, R. ve Chun, M. (2003). A new field of dreams: the collegiate learning assessment project. *Peer Review*, *5*(4), 26-29.

Bilgen, Ö. B. ve Doğan, N. (2017). Puanlayıcılar arası güvenirlik belirleme tekniklerinin karşılaştırılması. *Journal of Measurement and Evaluation in Education and Psychology*, *8*(1), 63-78. https://doi.org/10.21031/epod.294847

Bishop, L. (2023). A computer wrote this paper: What chatgpt means for education, research, and writing. *Research, and Writing*, 26.

Brandt, D. (2005). Writing for a living: Literacy and the knowledge economy. *Written Communication*, *22*(2), 166-197. https://doi.org/10.1177/0741088305275218

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G. ve Askell, A. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877-1901.

Büyüköztürk, Ş. (2018). *Sosyal bilimler için veri analizi el kitabı*. Pegem Atıf İndeksi, 001-214.

Charney, D. 1984: The validity of using holistic scoring to evaluating writing: a critical overview. *Research in the Teaching of English*, 18, 65-81.

Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. d. O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N. ve Brockman, G. (2021). Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Chenoweth, N. A. ve Hayes, J. R. (2001). Fluency in writing: generating text in L1 and L2. *Written Commun. 18*(1), 80–98. https://doi.org/10.1177/07410883010180010

Chodorow, M. ve Burstein, J. (2004). Beyond essay length: evaluating e-rater 's performance on toefl essays. *ETS Research Reports-73*, i–38. https://doi.org/10.1002/j.2333-8504.2004.tb01931.x

Coşkun, E. ve Tiryaki E. N. (2011). Tartışmacı metin yapısı ve öğretimi. *Mustafa Kemal Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, *8*(16), 63-73.

Crossley, S. (2020). Linguistic features in writing quality and development: an overview. *Journal of Writing Reseearch*, *11*(3), 415–443. https://doi.org/10.17239/jowr-2020.11.03.01

Crosthwaite, P., Storch, N. ve Schweinberger, M. (2020). Less is more? The impact of written corrective feedback on corpus-assisted L2 error resolution. *Journal of Second Language Writing*, *49*. https://doi.org/10.1016/j.jslw.2020.100729

Deane, P. (2013) On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing*, 18(1), 7–24. https://doi.org/10.1016/j.asw.2012.10.002

Enos, M. F. (2010). Instructional interventions for improving proofreading and editing skills of college students. *Business Communication Quarterly*, *73*(3), 265-281. https://doi.org/10.1177/1080569910376535.

Fisher, A. (1999). Ask Annie. *Fortune*, *145*(5), 223-225.

Fleiss, J. L., Nee, J. C. ve Landis, J. R. (1979). Large sample variance of kappa in the case of different sets of raters. *Psychological Bulletin*, *86*(5), 974. https://doi.org/10.1037/0033-2909.86.5.974

Floridi, L. ve Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, *30*, 681-694.

Göçer, A. (2010). Eğitim fakültesi öğrencilerinin yazılı anlatım becerilerinin süreç yaklaşımı ve metinsellik ölçütleri ekseninde değerlendirilmesi Niğde üniversitesi örneği. *Kastamonu Eğitim Dergisi*, *18*(1), 271-290.

Graham, S., Harris, K. R. ve Hebert, M. (2011). It is more than just the message: presentation effects in scoring writing. *Focus Exceptional Children*, *44*(4), 1–12.

Greenberg, K. L. (1988). *Effective writing choices and conventions*. St. Martin's Press.

Guo, L., Crossley, S. A. ve McNamara, D. S. (2013). Predicting human judgments of essay quality in both integrated and independent second language writing samples: a comparison study. *Assessing Writing*,*18*(3), 218–238. https://doi.org/10.1016/j.asw.2013.05.002

Günay, D. (2007). *Metin bilgisi*. Multilingual Yayınları.

Henricks, M. (2007). Writing skills are vital for today's employees, but few have them. *Entrepreneur*, *35*(7), 85-86.

Hirschberg, J. ve Manning, C. D. (2015). Advances in natural language processing. *Science*, *349*(6245), 261-266. https://doi.org/10.1126/science.aaa8685

Hsu, H. C. (2019). Wiki-mediated collaboration and its association with L2 writing development: An exploratory study. *Computer Assisted Language Learning*, *32*(8), 945–967. https://doi.org/10.1080/09588221.2018.1542407

Huot, B. (1990). The literature of direct writing assessment: major concerns and prevailing trends. *Review of Educational Research*, 6082), 237-263. https://doi.org/10.3102/00346543060002237

Imran, M. ve Almusharraf, N. (2023). Analyzing the role of ChatGPT as a writing assistant at higher education level: A systematic review of the literature. *Contemporary Educational Technology*, 15(4).

Ivanov, S. and Soliman, M. (2023). Game of algorithms: chatgpt

implications for the future of tourism education and research. *Journal of Tourism Futures*, *9*(2), 214-221. https://doi.org/10.1108/jtf-02-2023-0038

Jabotinsky, H. Y. ve Sarel, R. (2022). *Co-authoring with an AI? Ethical Dilemmas and Artificial Intelligence* (July 31, 2023). Arizona State Law Journal, Forthcoming, Available at SSRN: https://ssrn.com/abstract=4303959 or http://dx.doi.org/10.2139/ssrn.4303959

Jalil, S., Rafi, S., LaToza, T. D., Moran, K. ve Lam, W. (2023, April). Chatgpt and software testing education: Promises & perils. In *2023 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW)* (pp. 4130-4137). IEEE.

Jiao, W., Wang, W., Huang, J. T., Wang, X. ve Tu, Z. (2023). Is ChatGPT a good translator? A preliminary study. https://arxiv.org/abs/2301.08745

Kahraman, A. ve Yalvaç, F. (2015). EFL Turkish university students' preferences about teacher feedback and its importance. *Procedia - Social and Behavioral Sciences*, 199, 73 – 80. https://doi.org/10.1016/j.sbspro.2015.07.489

Kalı, G. (2016). *Türkçe öğretmeni adaylarının öyküleyici anlatımlarının bağdaşıklık ve tutarlılık açısından incelenmesi*. [Yayımlanmamış yüksek lisans tezi, Muğla Sıtkı Kocaman Üniversitesi Eğitim Bilimleri Enstitüsü].

Kamnis, S. (2023). Generative pre-trained transformers (GPT) for surface engineering. *Surface and Coatings Technology*, 129680. https://doi.org/10.1016/j.surfcoat.2023.129680

Kang, Y., Cai, Z., Tan, C.-W., Huang, Q. ve Liu, H. (2020). Natural language processing (NLP) in management research: A literature review. Journal of Management Analytics, 7(2), 139-172. https://doi.org/10.1080/23270012.2020.1756939

Kellogg, R. T. ve Raulerson, B. A. (2007). Improving the writing skills of college students. *Psychonomic Bulletin & Review*, *14*(2), 237-242.

Kroll, B. (1998). Assessing writing abilities. *Annual review of applied linguistics*, 18, 219-240.

Kyparisis, J. (1987). Sensitivity analysis framework for variational inequalities. *Mathematical programming*, *38*, 203-213.

Lee, L. (2020). An Exploratory study of using personal blogs for l2 writing in fully online language courses. In B. Zou & M. Thomas (Eds.), *Recent Developments in Technology-Enhanced and Com- puter-Assisted Language Learning* (pp. 145–163). Information Science Reference.

Liu, L. ve Gibson, D. (2023). *Exploring the Use of ChatGPT for Learning and Research: Content Data Analysis and Concerns*. Society for Information Technology & Teacher Education International Conference, Mar 13, 2023 in New Orleans, LA, United States.

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F. ve Liang, P. (2023). Lost in the Middle: How Language Models Use Long Contexts. *arXiv preprint arXiv:2307.03172*.

Lund, B. D. ve Wang, T. (2023). Chatting about ChatGPT: how may AI and GPT impact academia and libraries? *Library Hi Tech News*. https://www.researchgate.net/profile/Brady-Lund/publication/367161545_Chatting_about_ChatGPT_How_may_AI_and_GPT_impact_academia_and_libraries/links/6412235c315dfb4cce80f0e4/Chatting-about-ChatGPT-How-may-AI-and-GPT-impact-academia-and-libraries.pdf

Marcoulides, G. A. (1998). Applied Generalizability Theory Models. In Marcoulides, G.A., Editor. *Modern methods of business research*. NJ: Lawrence Erlbaum Associates.

May, G. L., Thompson, M. A. ve Hebblethwaite, J. (2012). A process for assessing and improving business writing at the MBA level. *Business Communication Quarterly*, *75*(3) 252-270. https://doi.org/10.1177/1080569912441822

McNamara, D. S., Crossley, S. A., Roscoe, R. D., Allen, L. K. ve Dai, J. (2015). A hierarchical classification approach to automated essay scoring. *Assessing Writing*, 23, 35–59. https://doi.org/10.1016/j.asw.2014.09.002

Mhlanga, D. (2023). The Value of Open AI and Chat GPT for the Current Learning Environments and the Potential Future Uses. *Available at SSRN 4439267*. http://dx.doi.org/10.2139/ssrn.4439267

Michel-Villarreal, R., Vilalta-Perdomo, E., Salinas-Navarro, D. E., Thierry-Aguilera, R., & Gerardou, F. S. (2023). Challenges and opportunities of generative AI for higher education as explained by ChatGPT. *Education Sciences*, 13(9). https://doi.org/10.3390/educsci13090856

Mizumoto, A. ve Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*, *2*(2), 100050. https://doi.org/10.1016/j.rmal.2023.100050

Monis, M. ve Rodriques, M. V. (2012). Teaching creative writing in English language classroom. *Indian Streams Research Journal*, *2(10)*,1-7.

NAEP (National Assessment of Educational Progress). (2002). *The nation's report card. Writing 2002 major results.* nces.ed.gov/nationsreportcard/writing/results2002/

Nagata, R., Hashiguchi, T. ve Sadoun, D. (2020). Is the simplest chatbot effective in english writing learning assistance? In L.-M. Nguyen, X.-H. Phan, K. Hasida & S. Tojo (Eds.), *Computational Linguistics* (ss. 245–256). Springer.

Okonkwo, C. W. ve Ade-Ibijola, A. (2021). Chatbots applications in education: A systematic review. *Computers and Education: Artificial Intelligence*, *2*. https://doi.org/10.1016/j.caeai.2021.100033

OpenAI. (2022, November 30). *ChatGPT: Optimizing Language Models for Dialogue*. OpenAI. https://openai.com/blog/chatgpt

Pavlik, J. V. (2023). Collaborating with chatgpt: considering the implications of generative artificial intelligence for journalism and media education. *Journalism & Mass Communication Educator*, 78 (1), 84-93.10776958221149577. https://doi.org/10.1177/107769582211495

Powers, D. E. (2005). Wordiness: a selective review of its influence, and suggestions for investigating its relevance in tests requiring extended written responses. *ETS Research Reports*, i–14.

Quible, Z. K. ve Griffin, F. (2007). Are writing deficiencies creating a lost generation of business writers?. *Journal of Education for Business*, *83*(1), 32-36.

Quinlan, T., Higgins, D. ve Wolff, S. (2009). Evaluating the construct-coverage of the e-rater scoring engine. *ETS Research Reports*, i–35.

Ramachandran, L., Cheng, J. ve Foltz, P. (2015). *Identifying patterns for short answer scoring using graph-based lexico-semantic text matching*. Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications, Colorado.

Ramalingam, V. Pandian, A., Chetry, P. ve Nigam, H. (2018, January). Automated essay grading using machine learning algorithm. Journal of Physics: Conference Series.

Ramesh, D. ve Sanampudi, S. K. (2022). An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review*, *55*(3), 2495-2527. https://doi.org/10.1007/s10462-021-10068-2

Riordan, D. A., Riordan, M. P. ve Sullivan, M. C. (2000). Writing across the accounting curriculum: An experiment. *Business Communication Quarterly*, *63*(3), 49-59. https://doi.org/10.1177/108056990006300

Rowh, M. (2006). Write well, go far, it's the skill every employer demands. Here's how to build it. *Career World*, *A Weekly Reader Publication*, *34*(4), 18- 23.

Saravia, E. (2018). Deep learning for NLP: An overview of recent trends. *Retrieved November*, *27*, 2018.

Schoonen, R. (2005). Generalizability of writing scores: an application of structural equation modeling. *Language Testing*, *22* (1) 1–30. https://doi.org/10.1191/0265532205lt295oa

Schoonen, R., Vergeer, M. & Eiting, M. (1997). The assessment of writing ability: expert readers versus lay readers. *Language Testing*, 14 (2), 157-184. https://doi.org/10.1177/026553229701400203

Seçkin, P., Arslan, N. ve Ergenç, S. (2014). Bağdaşıklık ve tutarlılık bakımından lise ve üniversite öğrencilerinin yazılı anlatım becerileri. *Uluslararası Türkçe Edebiyat Kültür eğitim (TEKE) Dergisi*, 3(1), 340-353. https://doi.org/10.7884/teke.269

Smerd, J. (2007). New workers solely lacking literacy skills. *Workforce Management*, *86*(21), 6.

Stangor, C. ve Walinga, J. (2019). 3.5 psychologists use descriptive, correlational, and experimental research designs to understand behaviour. *Introduction to Psychology*.

Stevens, B. (2005). What communication skills do employers want? Silicon Valley recruiters respond. *Journal of Employment Counseling*, *42*(1), 2-9. https://doi.org/10.1002/j.2161-1920.2005.tb00893.x

Stokel-Walker, C. (2022). AI bot ChatGPT writes smart essays—Should professors worry? *Nature*, *613*, 620–621 https://doi.org/10.1038/d41586-022-04397-7

Susnjak, T. (2022). ChatGPT: The End of Online Exam Integrity? *https://arxiv.org/abs/2212.09292*

Tan, R. G. R., Aviso, K. B. ve Uy, O. M. (2015). Comprehensive sensitivity analysis in NLP models in PSE applications using space-filling DOE strategy. *Chemical Engineering Transactions*, *45*, 523-528. https://doi.org/10.3303/CET1545088

Talan, T. ve Kalınkara Y. (2023). The role of artificial ıntelligence in higher education: ChatGPT assessment for anatomy Course. *Uluslararası Yönetim Bilişim Sistemleri ve Bilgisayar Bilimleri Dergisi* 7 (1).

Taylor, R. (1990). Interpretation of the correlation coefficient: a basic review. *Journal of diagnostic medical sonography*, *6*(1), 35-39. https://doi.org/10.1177/875647939000600106

Tuzi, F. (2004). The impact of e-feedback on the revisions of L2 writers in an academic writing course. *Computers and Composition*, *21*(2), 217–235. https://doi.org/10.1016/j.compcom.2004.02.003

Wenzlaff, K. ve Spaeth, S. (2022). Smarter than humans? Validating how OpenAI's ChatGPT model explains crowdfunding, alternative finance and community finance. *SSRN Electronic Journal* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4302443

Wolfe, E. W., Song, T. ve Jiao, H. (2016). Features of difficult to score essays. *Assessing Writing.* 27, 1–10. https://doi.org/10.1016/j.asw.2015.06.002

Xu, H. ve Lv, Y. (2022). Mining and application of tourism online review text based on natural language processing and text classification technology. *Wireless Communications and Mobile Computing*, 2022. https://doi.org/10.1155/2022/9905114.

Yang, R., Cao, J., Wen, Z., Wu, Y. ve He, X. (2020). Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking. *Findings of the Association for Computational Linguistics:* EMNLP 2020, 1560-1569.

Yeadon, W., Inyang, O.-O., Mizouri, A., Peach, A. ve Testrow, C. (2022). The death of the short-form physics essay in the coming AI revolution. *Physics Education*, 58, 1-13. https://doi.org/10.1088/1361-6552/acc5cf

Zech, J. M., Steele, R., Foley, V. K. ve Hull, T. D. (2022). Automatic rating of therapist facilitative interpersonal skills in text: A natural language processing application. *Frontiers in Digital Health*, *4*, 917918. https://doi.org/10.3389/fdgth.2022.917918

Zhang, Z., Han, X., Zhou, H., Ke, P., Gu, Y., Ye, D., Qin, Y., Su, Y., Ji, H. ve Guan, J. (2021). CPM: A large-scale generative Chinese pre-trained language model. *AI Open*, *2*, 93-99. https://doi.org/10.1016/j.aiopen.2021.07.001

Zhu, X., Zhu, J., Li, H., Wu, X., Li, H., Wang, X. ve Dai, J. (2022). Uni-perceiver: Pre-training unified architecture for generic perception for zero-shot and few-shot tasks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition