THE JOURNAL OF TURKISH EDUCATIONAL SCIENCES
TÜRK EĞİTİM BİLİMLERİ DERGİSİ

Kimya Sorularının Cevaplanmasında Yapay Zekâ Tabanlı Sohbet Robotlarının Performansının İncelenmesi*

Investigating the Performance of AI-Based Chatbots in Answering Chemistry Questions

Ayşe Yalçın-Çelik, Özgür K.Çoban

Yazar Bilgileri

Ayşe Yalçın-Çelik
Doç. Dr., Gazi Üniversitesi, Matematik ve Fen Bilimleri Eğitimi,
ayseyalcin@gazi.edu.tr

Özgür K.Çoban
Doktora Öğrencisi, Milli Savunma Üniversitesi, Kara Harp Okulu,
okcoban@msu.kho.edu.tr

ÖZ

Yapay zekâ son yıllarda sağlık, bankacılık ve finans, teknoloji, endüstri, psikoloji ve eğitim gibi birçok alanda kullanılmaktadır. Özellikle doğal dili anlayan ve dil modellerini etkili bir şekilde kullanarak cevaplar verebilen yapay zekâ tabanlı sohbet robotlarının (chatbot) ortaya çıkmasıyla beraber sohbet robotlarının sorulara verdikleri cevapların doğruluk düzeyi tartışma konusu olmuştur. Bu araştırmanın amacı, iki sohbet robotunun üniversite seviyesinde, Bloom'un bilişsel alan taksonomisi dikkate alınarak, yüzey gerilimi konusu ile ilgili hazırlanmış sorulara verdikleri cevapların doğruluk düzeylerini belirlemektir. Araştırmanın deseni durum çalışması olarak belirlenmiştir. Veri toplama aracı olarak yüzey gerilimi ile ilgili Bloom'un bilişsel alan taksonomisi dikkate alınarak hazırlanmış altı adet açık uçlu sorudan oluşan ölçek kullanılmıştır. Sohbet robotlarının yüzey gerilimi ile ilgili sorulara verdiği cevaplar üç uzman tarafından değerlendirilmiştir. Araştırmanın bulgularına göre sohbet robotlarının 60 puan üzerinden 35 ve 38 puan aldıkları, aynı sorularda aynı puan ortalamalarına sahip oldukları, çözümleme düzeyindeki soruyu yanlış cevapladıkları, yaratma düzeyindeki sorudan en yüksek puanı aldıkları ve cevaplarında yanlışlıklar/eksiklikler olduğu ancak açıklamalarının %66,7 oranında net olduğu belirlenmiştir. Bu sonuçlardan yola çıkarak; sohbet robotlarının performansının zorluk seviyesi kolaydan zora doğru olan farklı konularda belirlendiği, istem (prompt) girişinin birden fazla yapılarak bu uygulamanın daha doğru cevapların üretilmesine etki edip etmediği ve sohbet robotların cevaplarında yanlış kavramaların olup olmadığının belirlendiği çalışmaların yapılması önerilmektedir.

Makale Bilgileri

Anahtar Kelimeler
Yapay zekâ
Sohbet robotu
Kimya eğitimi
Bloom'un bilişsel alan taksonomisi

Keywords
Artificial intelligence
Chatbot
Chemistry education
Bloom's cognitive domain taxonomy

Makale Geçmişi
Geliş: 16.09.2023
Düzeltme: 20.10.23
Kabul: 05.11.2023

ABSTRACT

Artificial intelligence has developed rapidly in recent years and is used in many fields, such as health, banking and finance, technology, industry, psychology and education. Especially with the emergence of artificial intelligence-based chatbots that understand natural language and can answer using language models effectively, the accuracy level of the answers given by chatbots to questions has been a subject of discussion. This study aims to determine the accuracy levels of the answers provided by two chatbots to the questions prepared about surface tension at university level, taking into account Bloom's cognitive domain taxonomy. The research design was determined as a case study. A scale of six open-ended questions about surface tension prepared using Bloom's cognitive domain taxonomy was used as a data collection tool. Three experts evaluated the answers of chatbots to the questions about surface tension. According to the results of the study, the chatbots scored 35 and 38 out of 60 points, they had the same average scores on the same questions, they answered the question at the analysis level incorrectly, they got the highest score on the question at the creation level, and there was misinformation/insufficient information in their answers, but 66.7% of their explanations were clear. Based on these results, it is recommended to carry out studies in which the performance of chatbots is determined in different subjects with difficulty levels from easy to difficult, whether this application affects the production of more accurate answers by making more than one prompt input, and whether there are misconceptions in the responses of chatbots.

* Bu çalışma birinci yazarın danışmanlığında ikinci yazarın devam etmekte olan doktora tezinden üretilmiştir.

Makale Türü Araştırma

Önerilen Atıf Yalçın-Çelik, A. & K.Çoban, Ö. (2023). Kimya sorularının cevaplanmasında yapay zekâ tabanlı sohbet robotlarının performansının incelenmesi. *TEBD*, *21*(3), 1540-1561. https://doi.org/10.37217/tebd.1361401

Giriş

21. yüzyılda gelişen teknoloji ile birlikte yapay zekâ çalışmaları değer kazanmıştır. Yapay zekâ (artificial intelligence - AI), bir bilgisayar veya bilgisayar tarafından kontrol edilen bir robotun, insan zekâsı ve ayırt etme yeteneği gerektirdiği için genellikle insanlar tarafından yapılan görevleri yerine getirme yeteneği olarak tanımlanmaktadır (Meço ve Coştu, 2022). Diğer bir ifadeyle "insanı taklit etme yeteneğine sahip, etkileşim, öğrenme, uyum sağlama ve tecrübelerini genişleterek uygulama imkânı olan dijital teknoloji ve/veya uygulamalar" olarak tanımlanmaktadır (Tamer ve Övgün, 2020).

Yapay zekâ; günümüzde sağlık, bankacılık ve finans, teknoloji, endüstri, psikoloji, güvenlik, eğitim gibi birçok alanda kullanılmaktadır (AlAfnan, Dishari, Jovic ve Lomidze, 2023; Das vd., 2023; Fergus, Botha ve Ostovar, 2023; Geerling, Mateer, Wooten ve Damodaran, 2023; Kung vd., 2023; Zhu, Jiang, Yang ve Ren, 2023). Kullanımı sıklıkla artan ve her geçen gün yeni uygulamaları çıkan yapay zekâ terimi tarihte ilk kez 1956'da Hannover'da yapılan bir konferansta ortaya atılmıştır (Haenlein ve Kaplan, 2019). Alanyazındaki birçok çalışma, yapay zekânın birçok farklı kullanım alanlarını örneklendirmekte ve AI kullanımının avantajlarını ve dezavantajlarını ortaya koymaktadır (AlAfnan vd., 2023; Das vd., 2023; Lo, 2023; Motlagh, Khajavi, Sharifi ve Ahmadi, 2023; Shawar ve Atwell, 2007). Özellikle eğitim alanında, yapay zekânın kullanım amaçlarının çeşitlilik gösterdiği ve bu kullanımın eğitimciler için yararlı olabildiği gibi yeni sorunlar yarattığı da belirlenmiştir (Gregorcic ve Pendrill, 2023). Yapay zekânın eğitimdeki etkili uygulama alanlarından birisi de öğrencilerin bireysel takibini yapabilen ve "Intelligent Tutoring Systems (ITS)" olarak adlandırılan sistemlerin kullanımıdır (Steenbergen-Hu ve Cooper, 2014).

Lo'ya (2023) göre yapay zekâ, öğretmenler tarafından öğretim faaliyetlerinde ders materyalleri oluşturmak, dil çevirisi yapmak, öğretmenlere önerilerde bulunmak, değerlendirme görevi oluşturmak ve öğrencilerin performansını değerlendirmek için kullanılmaktadır. Benzer şekilde, yapay zekâ uygulamaları öğrencilerin sorularını yanıtlayarak, metinleri özetleyerek veya ödevlerini hazırlayarak öğrenmelerini desteklemektedir. Ancak yapay zekânın hazırladığı cevapların her zaman doğru olmadığı, yanlış ve/veya eksik verilere dayanarak yorum yaptığı veya yanlış ve/veya gerçek olmayan bilgileri üretmesi gibi eleştirilerde mevcuttur (Han, Battaglia, Udaiyar, Fooks ve Terlecky, 2023, s. 32). Yapay zekâ araçlarının verdiği yanlış veya eksik cevaplar, öğrenciler için anlamlı öğrenmeyi engelleyen önemli bir zorluk olarak karşımıza çıkabilir (Clark, 2023). Ayrıca yapay zekâ araçlarının öğrenci yerine ödevleri yapması veya sınavları cevaplaması etik açıdan da uygun değildir (Mhlanga, 2023; Susnjak, 2022). Bu gibi olumsuzlukları ortadan kaldırmak veya en aza indirmek için önerilen yollardan biri çevrim içi sınavları veya verilen ödevleri çoktan seçmeli veya kısa cevaplı sorular yerine üst düzey düşünme becerisi gerektiren sorular ve ödevler şeklinde verilmesidir (Geerling vd., 2023; Susnjak, 2022). Ancak bu gerçekten işe yarar mı?

2022 yılından itibaren eğitim alanında kullanılan bir yapay zekâ aracı da sohbet robotlarıdır (chatbots). İnsan diyaloğunu taklit etmek için doğal dil işleme (NLP) teknolojisini kullanan yapay zekâ tabanlı robotlar, sohbet robotu olarak bilinir. 1966'dan günümüze kadar farklı sürümleri olan sohbet robotları şu an insanlardan komut almakta, komutları yorumlayabilmekte ve hatta sesli cevap bile verebilmektedir (Suta, Lan, Wu, Mongkolman ve Chan, 2020). Sohbet robotları, bir sohbet aracı olarak, kullanıcılar tarafından sorulan soruları yanıtlayabildiği gibi ödevler hazırlayabilmekte veya sınav sorularını yanıtlayabilmektedir (Fergus vd., 2023). Sohbet robotlarının eğitimde kullanımı OpenAI tarafından geliştirilen ChatGPT ile hızlı bir şekilde artış göstermiştir. Benzer şekilde sıklıkla kullanılan diğer yapay zekâ tabanlı sohbet robotları Chat, Bard, Bing ve Ernie'dir (Rudolph, Tan ve Tan, 2023). Sohbet robotlarının eğitimde kullanılmasıyla öğrencilerin bu araçları farklı amaçlar için kullanmaya başladığı da tespit edilmiştir. Özellikle uzaktan eğitim sürecinde sınavlarda veya ödev hazırlama süreçlerinde sohbet robotu kullanımı artmıştır. Sohbet robotlarının ödev sürecinde ve sınavlarda kullanımları eğitimcileri fazlasıyla tedirgin etmektedir (Motlagh vd., 2023).

Sohbet robotlarının öğrencilerin sorduğu soruları cevapladığı, üst düzey sorulara bile cevap verdiğini belirten çalışmalar mevcuttur. Örneğin Susnjak (2022), bir yapay zekâ tabanlı sohbet robotu olan ChatGPT'den üniversite düzeyinde üst düzey düşünme gerektiren sorular hazırlamasını, bu soruları cevaplamasını ve cevaplarını analiz etmesini istemiştir. ChatGPT'nin verdikleri cevapları analiz ederek yapay zekânın kapasitesi hakkında yorumda bulunmuştur. Araştırma sonucunda sohbet robotu teknolojilerinin olağanüstü seviyelere ulaştığını ve sadece bilgi veren bir araç olmasının yanı sıra eleştirel düşünme yeteneğine sahip olduklarını deneysel olarak ortaya koymuştur.

Sohbet robotu teknolojisinin kapasitesinin araştırıldığı bir diğer çalışma da Jalil, Rafi, LaToza, Moran ve Lam (2023) tarafından yürütülmüştür. Jalil vd. (2023), lisans düzeyinde "software testing" dersi ile ilgili bir kitapta bulunan soruları ChatGPT'ye sormuş ve soruları cevaplama kapasitesini belirlemeye çalışmıştır. ChatGPT'nin soruların %77,5'ini cevapladığını, cevapların %55,6'sının ise doğru olduğunu belirlemiştir. Das vd. (2023) de ChatGPT'ye mikrobiyoloji konusu ile üst düzey düşünme ve yorumlama becerisi gerektiren sorular sormuş ve %80 doğruluk oranında cevaplar elde etmiştir.

Korsakova vd. (2022) öğrencileri sınava hazırlayan bir yapay zekâ robotu kullanarak bir araştırma yapmışlardır. Araştırmada kimya öğrencilerine konu anlatan, sınav ödevleri veren veya eğitsel oyunlar oynatabilen bir sohbet robotu tanıtılmıştır. 465 öğrencinin katıldığı çalışmada öğrenciler 3 ay sohbet robotu ile etkileşime geçmiştir. Araştırma sonucunda ülke genelinde uygulanan bir sınava katılan öğrencilerin çok yüksek başarılar elde ettikleri belirlenmiştir. İlgili araştırmacılar bu başarıya yol açan etmenlerden birisinin sohbet robotu olabileceği yönünde görüş bildirmişlerdir.

Clark (2023), kimya biliminin soyut doğasına vurgu yaparak sohbet robotlarının kimya ile ilgili soru ve problemleri cevaplama, açıklama ve problemleri çözme kapasitelerini araştırmıştır. Bu amaçla sohbet robotuna açık uçlu ve kapalı uçlu sorular sormuş, aldığı cevapları değerlendirmiş ve öğrenci cevapları ile karşılaştırmıştır. Kapalı uçlu sorular, sayısal soru ve sayısal olmayan soru olarak ikiye ayrılarak analiz gerçekleştirilmiştir. Sayısal sorularda sohbet robotunun ortalaması ile öğrencilerin ortalaması, sırasıyla, %40 ve %41 iken sayısal olmayan sorularda ise sohbet robotunun ortalaması ile öğrencilerin ortalaması, sırasıyla, %47 ve %65'tir. Açık uçlu sorular ise Bloom'un bilişsel alan taksonomisine göre sınıflandırılmıştır. Cevaplar, sohbet robotunun verdiği cevapların soru ile ilişkisi, cevabın doğruluğu ve yanlışlığı dikkate alınarak analiz edilmiştir. Araştırmacı analiz sonucunda sohbet robotunun öğrenciler gibi uzun metinler yazabildiğini ama öğrencilerin sohbet robotundan farklı olarak cevaplarını şekiller ve grafikler ile desteklerini ifade etmiştir. Ayrıca sohbet robotunun, temel kimya düzeyindeki kavramlara cevap içinde başarılı bir şekilde yer verdiğini ancak anlamlarını yanlış yorumladığı veya olguyu yanlış modellediğini bu yüzden bir sınava girecek olsa cevapların puanlanması kısmında ise en düşük öğrenciden bile daha düşük bir not alabileceğini ifade etmiştir. Bunun sebebini de cevapların bir kısmında çok yaygın olan yanlış kavramaların bulunmasıyla ilişkilendirmiştir.

Sohbet robotlarının performansı ile ilgili alan yazın gözden geçirildiğinde, yukarıdaki örneklerde olduğu gibi, performansının düşük veya yüksek olduğu durumların daha fazla araştırılması ve performansını etkileyen durumların konuya/alana bağlı olarak değerlendirilmesi gereklidir. Bu nedenle bu araştırma, yapay zekâ tabanlı sohbet robotlarının kimya alanında performansını belirlemeye odaklanmaktadır. Bu araştırmanın amacı, yapay zekâ tabanlı iki farklı sohbet robotunun üniversite seviyesinde Bloom'un bilişsel alan taksonomisi dikkate alınarak yüzey gerilimi konusu ile ilgili hazırlanmış sorulardaki performanslarını belirlemektir. Yüzey gerilimi konusundaki performanslarını belirlemek için sadece iki sohbet robotunun kullanılmasını içeren bu araştırmanın hedefi, sohbet robotlarını birbirleriyle karşılaştırarak hangisinin kimya ile ilgili sorulara doğru yanıt verdiğini tespit etmek değildir. Aksine, bu araştırmanın iki farklı sohbet robotu kullanılarak yapılmasının nedeni, sohbet robotlarının kimya ile ilgili alanlardaki performans düzeyini ortaya koyarak kimyada ne kadar başarılı olabilecekleri hakkında öğretmenlere, öğrencilere ve eğitimcilere bilgi sağlamaktır. Araştırmada, güncel olan, sıklıkla tercih edilen iki adet sohbet robotunun (ChatGPT ve Bard) kimya sorularına verdikleri cevaplar analiz edilerek performansları belirlenmiştir.

Sohbet robotlarının yüzey gerilimi konusundaki performanslarını belirlemeyi amaçlayan bu çalışmanın alt problemleri aşağıdaki gibidir:

1. Sohbet robotları yüzey gerilimi konusu ile ilgili hazırlanmış ölçekten kaç puan almıştır?

2. Sohbet robotlarının Bloom'un bilişsel alan taksonomisine göre hazırlanmış sorulardaki başarı düzeyi nedir?
3. Sohbet robotlarının ilgili sorulara verdikleri cevaplar nasıldır?
4. Sohbet robotlarının yanlış cevaplarının kaynakları neler olabilir?
5. Sohbet robotlarının cevaplarının netliği (anlaşılırlığı) nasıldır?

Yöntem

Araştırmanın Deseni

Bu araştırma, iki farklı yapay zekâ tabanlı sohbet robotunun performanslarını, Bloom'un bilişsel alan taksonomisi dikkate alınarak hazırlanmış yüzey gerilimi konusu ile ilgili sorulara verdikleri cevaplar analiz ederek belirlemeyi amaçladığı için nitel araştırma yöntemi dikkate alınarak gerçekleştirilmiştir. Veriler, sohbet robotlarının sorulara verdikleri cevapların içerik analizi ile sağlanmıştır. Araştırmada sohbet robotlarının kimya sorularına nasıl cevap verdikleri ayrıntılı olarak ele alınacağı için çalışmanın deseni durum çalışmasıdır. Durum çalışması, bir kişiyi, olayı veya durumu doğal bir ortamda derinlemesine tanımlamak, incelemek ve bütünsel olarak yorumlamak için kullanılan bir tekniktir (Yıldırım ve Şimşek, 2018).

Araştırmada Kullanılan Sohbet Robotları

Araştırmada iki farklı yapay zekâ tabanlı sohbet robotu kullanılmıştır: OpenAI tarafından piyasaya sunulan ChatGPT (https://chat.openai.com/) ve Google tarafından piyasaya sunulan Bard (https://bard.google.com/). Araştırmada bu iki sohbet robotunun tercih edilmesinin nedeni, her iki sohbet robotunun rahatlıkla ulaşılabilir ve Türkçe dil desteğinin olmasıdır. Ayrıca bu modeller, kullanıcı girdisine yanıt olarak insan konuşmasına benzeyen metinler üretmek için doğal dil işleme (Natural Language Processing -NLP) teknolojilerini kullanmaktadır. ChatGPT (Chat Generative Pre-trained Transformer) Kasım 2022'de Open IA tarafından piyasaya sürülmüş hızlı bir şekilde dünya genelinde çok fazla kullanıcıya sahip olmuştur (Lo, 2023; Rahaman, Ahsan, Anjum, Rahman ve Rahman, 2023). ChatGPT'nin bu kadar hızlı yaygınlaşmasında kullanımının kolay olması ve ücretsiz sürümlerinin olması önemli bir faktör olmuştur. ChatGPT'nin yaygın olarak kullanılan "3,5" sürümü yanında ChatGPT'nin yeni bir sürümü, "ChatGPT 4.0", Mart 2023'te yayınlanmıştır. Ancak bu sürümü ücretsiz değildir. Bu araştırmada herkes tarafından ulaşılabilirliği daha kolay olmasından dolayı ChatGPT 3,5 sürümü tercih edilmiştir.

Araştırmada kullanılan diğer yapay zekâ tabanlı sohbet robotu ise Bard'dır. Bu sohbet robotu, Mart 2023'te Google firması tarafından kullanıma sunulmuştur. İlk kullanımında Türkçe desteği bulunmamakla birlikte Temmuz 2023'te Türkçe dil desteği de vermeye başlamıştır. Bu araştırmada

Bard'ın tercih edilmesindeki en önemli neden Türkçe dil desteğine yeni başlaması ve uygulamaya ücretsiz olarak herkesin erişebilmesidir.

Veri Toplama Aracı

Araştırmada, yüzey gerilimi ile ilgili altı adet açık uçlu soru içeren bir ölçek kullanılmıştır. Ölçekteki her bir soru Bloom'un yenilenmiş bilişsel alan taksonomisi dikkate alınarak hazırlanmıştır. Ölçekte her bir düzey için birer adet soru bulunmaktadır. Bahsi geçen ölçek araştırmacılar tarafından hazırlanmıştır. Ölçeğin hazırlanma aşamasında ilk önce Bloom'un bilişsel alan taksonomisindeki her bir düzey için iki adet soru olacak şekilde on iki adet sorudan oluşan bir soru havuzu oluşturulmuştur. Sorular, Bloom'un güncellenmiş bilişsel alan taksonomisindeki ana ve alt grupların açıklamaları ile bilişsel süreç gruplarındaki fiiller dikkate alınarak oluşturulmuştur (Anderson vd., 2001, s. 28-30).

Ayrıca hazırlanan sorularda, sohbet robotlarının okuyamamasından kaynaklı şekillerin, görsellerin ve bilimsel sembollerin olmamasına dikkat edilmiştir. Sadece bir tane soruda bilimsel bir gösterim olarak "^{o}C" kullanılmıştır. Sohbet robotlarının verdikleri cevaplardan okunduğu anlaşılan "^{o}C" sembolik gösterimini içeren soru havuzda tutulmuştur. Soruların, Bloom'un bilişsel alan taksonomisine göre sınıflandırılmasının uygunluğu için fen alanında uzman iki kişiden görüş alınmıştır. Uzmanlara, araştırmacılar tarafından hazırlanmış belirtke tablosu verilmiş ve soruların Bloom'un taksonomisine göre dağılımının uygun olma durumunu belirtmeleri istenmiştir. Uzmanlar arasındaki tutarlığı belirlemek için Miles ve Huberman (1994) uyum katsayısı kullanılmış ve uyum %91,6 olarak hesaplanmıştır.

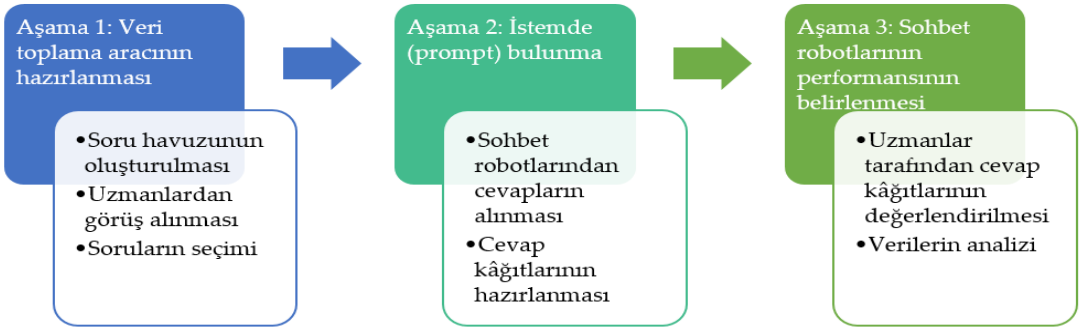
Soru havuzundan aynı düzeyden birer tane olacak şekilde altı adet soru seçilmiştir (Ek 1). Soruların Bloom taksonomisine göre sınıflandırılması Tablo 1'de verilmiştir. Bu sorular sohbet robotlarının performansını belirlemek amacıyla kullanılmıştır.

Tablo 1. Bloom'un Bilişsel Alan Taksonomisine Göre Soruların Dağılımı

Bilgi Birikimi	***Bilişsel Alan***					
Boyutu	***Hatırlama***	***Anlama***	***Uygulama***	***Çözümleme***	***Değerlendirme***	***Yaratma***
Olgusal	S1					
Kavramsal		S2		S4	S5	
İşlemsel			S3			S6
Üstbilişsel						

Araştırma Süreci

Araştırma üç aşamada gerçekleştirilmiştir (Şekil 1). İlk aşama, sohbet robotlarına sorulacak olan soruların hazırlanmasıdır. Bu aşama veri toplama aracının tanıtıldığı alt başlıkta ayrıntılı olarak açıklanmıştır.



Şekil 1. Araştırma süreci

İkinci aşama sohbet robotlarından cevapların alınması aşamasıdır. Araştırmacılardan birine her iki sohbet robotundan kullanıcı hesabı açılmıştır. Aynı sohbet (chat) sekmesinde sorulan sorulara verilen cevapların doğruluğu daha yüksek olduğu için (Jalil vd., 2023) açılan hesaptaki sohbet kısmına her bir istem (prompt) teker teker yazılmıştır. Bu araştırmadaki istemler sohbet robotu tarafından cevap oluşturulmasını sağlamak için kullanılan veri toplama aracındaki sorulardır. Her ne kadar yapay zekâ tabanlı sohbet robotlarının İngilizce istemlere daha iyi cevap verdiklerini ifade eden çalışmalar olsa da (Wood vd., 2023), bu araştırmada istemler Türkçe olarak girilmiş ve cevaplar Türkçe olarak alınmıştır. Bunun nedeni öğrencilerin İngilizceyi yeterince doğru kullanamaması ve sohbet robotlarının Türkçe dil desteğinin olmasıdır. İstemlerin sohbet robotları tarafından anlaşılırlığını artırabilmek için "bir kimya öğrencisi gibi düşün" veya "bir laboratuvar deneyi tasarla" gibi sorulara ilave ifadeler kullanılmamıştır. Çünkü sohbet robotları tarafından soruların anlaşılırlığını etkileyebilecek ifadeler soru havuzu oluşturulurken önceden düşünülerek sorulara dâhil edilmiştir. Beşinci sorudaki "…insan olsan…" gibi. Bütün ölçek maddeleri sohbet robotlarına aynı soru sorma sırasında, aynı sohbet (chat) bölmesine ve aynı gün sorulmuştur. Sohbet robotlarına sorular sadece bir kez sorulmuş, robotların soruları tekrar cevaplaması veya cevapların farklı taslaklarını üretmesi istenmemiştir. Sorulara verilen cevaplar kopyalanarak her bir sohbet robotu için bir cevap kâğıdı oluşturulmuştur.

Üçüncü aşama, sohbet robotlarının oluşturduğu cevapların uzmanlar tarafından incelenerek sohbet robotlarının performanslarının belirlendiği aşamadır. Bu aşamada fizikokimya ve/veya fizikokimya laboratuvarı dersini yürüten ve bu konuda çalışmaları olan üç uzmandan görüş alınmıştır. Uzmanlara, sohbet robotları tarafından oluşturulmuş cevap kâğıtları verilmiş ve değerlendirmeleri istenmiştir. Kâğıtlarda, sohbet robotları hakkında tanılayıcı bir bilgi bulunmamaktadır. Uzmanlardan cevapları ilk sorudan başlayarak ve her iki cevap kâğıdında da aynı soruyu okuyarak 0-10 arasında puan vererek değerlendirmeleri istenmiştir. Uzmanlar arasındaki objektifliği sağlamak amacıyla uzmanların birbirlerinin puanlarından haberdar olmamaları sağlanmıştır. Her bir sohbet robotunun 6 sorudan alabileceği en yüksek puan 60'tır.

Verilerin Analizi

Araştırma sorularına cevap bulmak amacıyla veriler nitel analiz gerçekleştirilerek elde edilmiştir. Sohbet robotlarının sorulara verdikleri cevaplar üç uzman tarafından incelenmiştir. Uzmanların cevaplara verdikleri puanlardan her bir sohbet robotunun toplam puanı ve ortalama puanı hesaplanmıştır. Uzmanların verdikleri puanlardan puanlayıcılar arası güvenirliği belirleyebilmek için sınıf içi korelasyon katsayısı sohbet robotları için sırasıyla, 0,67 ve 0,78 olarak hesaplanmıştır. Sınıf içi korelasyon katsayısının 0,5 den küçük değer olması zayıf güvenirliği; 0,5-0,75 arasında değer alması orta güvenirliği ve 0,75-0,9 arasında değer alması büyük güvenirliği ifade etmektedir (Koo ve Li, 2016). Bu sonuçlara göre ilk sohbet robotu için puanlayıcılar arası güvenirlik orta düzeyde iken ikinci sohbet robotu için puanlayıcılar arası güvenirlik iyi düzeyi ifade etmektedir.

Araştırmada aynı zamanda, sohbet robotlarının performansını belirlemeyebilmek için, verilen cevaplar "doğru", "kısmi doğru" ve "yanlış" olarak sınıflandırılmıştır. Bu aşamada Jalil vd.'nin (2023) sınıflandırılması dikkate alınmıştır. Jalil vd. (2023), araştırmalarında sohbet robotlarının cevabını "cevap kısmı" ve "açıklama kısmı" olarak ikiye ayırmış ve her bir kısmı ayrı olarak değerlendirmiştir. Bu araştırmada ise Jalil vd.'nin (2023) çalışmasından farklı olarak verilen cevap bir bütün olarak ele alınmış, uzmanların cevaplara verdikleri puanlar dikkate alınarak, 0-4 puan aralığı "yanlış", 5-8 puan aralığı "kısmi doğru" ve 9-10 puan aralığı ise "doğru" olarak kategorilendirilmiştir. Araştırmada aynı zamanda cevaplar bilgi eksikliği, yanlış bilgi içermesi ve bilgilerin yanlış yorumlanması açısından da incelenmiştir (Jalil vd., 2023). Böylelikle, yanlış veya kısmi yanlış olan cevapların nedeni belirlenmeye çalışılmıştır. Son olarak cevaplar açıklamaların netliği (anlaşılırlığı) açısından da analiz edilmiştir. Bu amaçla betimsel analiz ve içerik analizi birlikte yapılmıştır. Analizler araştırmacılar tarafından gerçekleştirilmiştir.

Bulgular

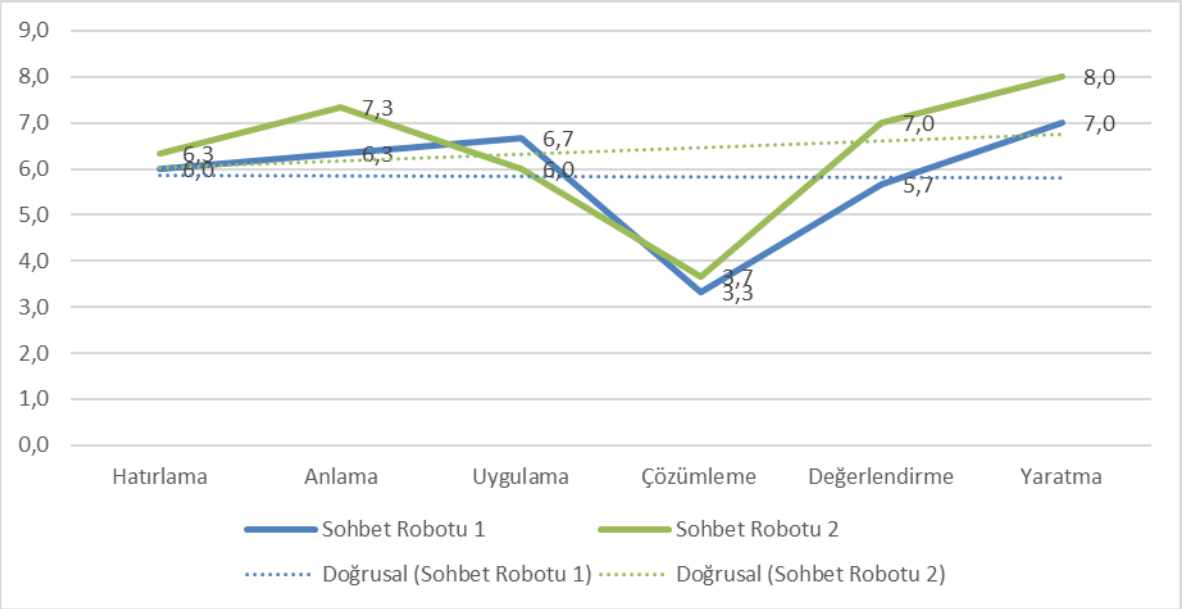
Araştırma bulguları, sohbet robotlarının sorulara verdikleri cevapların uzmanlar tarafından değerlendirilmesi ile hazırlanmıştır. Uzmanların sohbet robotlarının cevaplarına verdikleri puanlara ait betimsel istatistik değerleri Tablo 2'de verilmiştir.

Tablo 2. Sohbet Robotlarının Puanlarının Betimsel İstatistik Sonuçları

		Toplam Puan	$\bar{X}$	*SS*	*Ortanca*	*Min.*	*Mak.*
Uzman 1	Sohbet Robotu 1	36,0	6,0	1,09	6,0	4,0	7,0
	Sohbet Robotu 2	39,0	6,5	1,64	7,0	4,0	8,0
Uzman 2	Sohbet Robotu 1	34,0	5,7	1,51	6,0	4,0	7,0
	Sohbet Robotu 2	37,0	6,2	1,83	7,0	3,0	8,0
Uzman 3	Sohbet Robotu 1	35,0	5,8	2,04	6,0	2,0	8,0
	Sohbet Robotu 2	39,0	6,5	2,42	7,0	2,0	9,0
Sohbet Robotu 1	Toplam	35,0	5,8	1,50	6,0	2,0	8,0
Sohbet Robotu 2	Toplam	38,0	6,4	1,88	7,0	2,0	9,0

Tablo 2'ye göre yapay zekâ tabanlı sohbet robotlarının cevapları en düşük 2; en yüksek 9 puan olarak değerlendirilmiştir. Sohbet robotu 1'in toplam puanlarının ortalaması 5,8 iken, sohbet robotu 2'nin puan ortalaması 6,4'tür. Uzmanların puanlamaları birlikte değerlendirildiğinde sohbet robotu 1, 60 puan üzerinden 35 puan (100 üzerinden 58) ve sohbet robotu 2, 60 puan üzerinden 38 puan (100 üzerinden 64) almıştır.

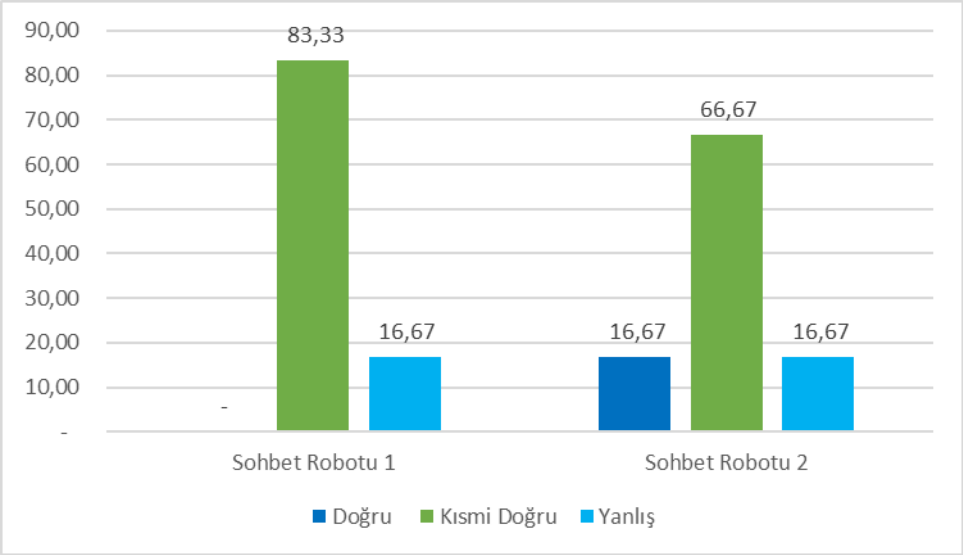
Araştırmanın bir diğer amacı da sohbet robotlarının Bloom'un bilişsel alan taksonomisindeki düzeylerdeki performanslarını belirlemektir. Bu amaçla her düzeydeki soru için uzmanların verdikleri puanların ortalamaları belirlenmiş ve Şekil 2'de verilmiştir.



Şekil 2. Bloom'un bilişsel alan taksonomisine göre sohbet robotlarının puan ortalamalarının dağılımı

Şekil 2'e göre her iki sohbet robotunun aynı taksonomi düzeyinde benzer ortalamalarda puan aldıkları görülmektedir. Bloom taksonomisinin ilk basamağı olan "hatırlama" düzeyinden "yaratma" düzeyine doğru sohbet robotu 2'nin puan ortalamalarında bir artış mevcuttur. Sohbet robotlarının en yüksek puan aldıkları soru "yaratma" düzeyine ait soru iken en düşük puan aldıkları soru ise "analiz etme" düzeyine ait soru olmuştur. Ayrıca bu düzeydeki soruya her iki sohbet robotu da yanlış cevap vermiştir.

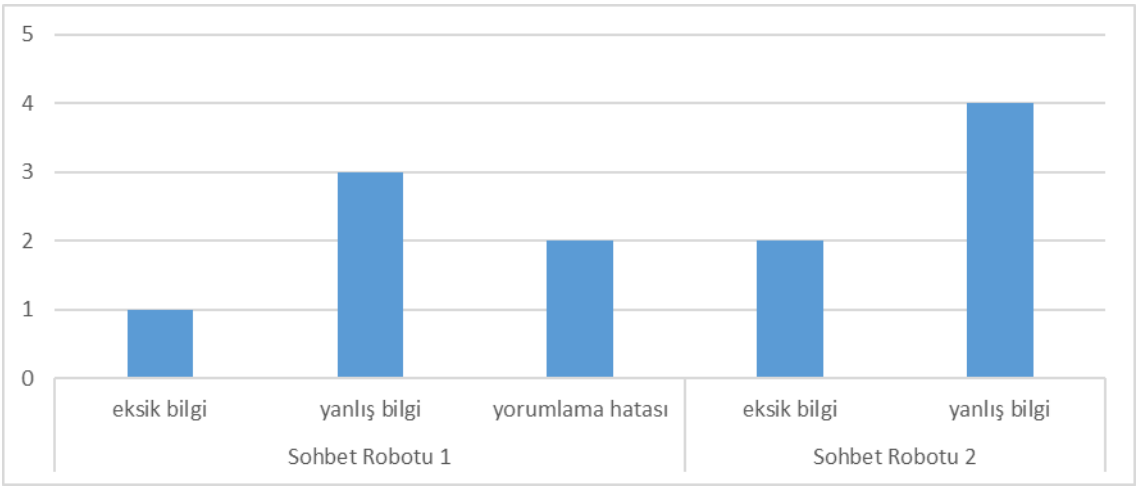
Araştırmanın bir başka amacı sohbet robotlarının performansını cevapların içeriğinin doğruluğu açısından sınıflandırmaktır. Sohbet robotlarının verdikleri cevaplar "doğru", "kısmi doğru" ve "yanlış" olarak değerlendirilmiştir ve Şekil 3'te verilmiştir.



Şekil 3. Sohbet robotlarının cevaplarının doğru/kısmı doğru/yanlış dağılımı

Şekil 3'e göre her iki sohbet robotunun cevapları, çoğunlukla, kısmi olarak (%83,33 ve %66,67, sırasıyla) doğrudur. Ayrıca sohbet robotu 1 hiçbir soruya tam olarak doğru cevap veremezken (9-10 puan aralığında alamamış), sohbet robotu 2 sadece bir soruya (%16,67) doğru cevap verebilmiştir (9-10 puan aralığında almış). Her iki sohbet robotu da birer soruyu yanlış olarak cevaplamıştır (0-4 puan aralığında almış). Her iki sohbet robotunun cevaplarının yanlış olarak sınıflandırıldığı bu soru "çözümleme" düzeyinde bir sorudur (Şekil 2). Çözümleme düzeyindeki soruda sohbet robotlarından aynı derişimdeki 2 faklı alkol çözeltisinin yüzey geriliminin kıyaslanması istenmektedir. Bu soruda sohbet robotu 1, yüzey gerilimine etki eden faktörleri doğru şekilde sıralayıp "*her iki alkolün çözeltileri 0,2 M (mol/litre) konsantrasyonda olduğu için, konsantrasyon etkisi göz ardı edilebilir… (çözelti durumunu dikkate almıyor)... Sonuç olarak 0,2 M izobütil alkol çözeltisi, 0,2 M amil alkol çözeltisine göre daha düşük bir yüzey gerilimine sahip olma eğilimindedir*" cevabını vererek yanlış kıyaslama yapmıştır. Sohbet robotu 2 ise, yüzey gerilimlerini sadece molekül ağırlığına bağlı olarak kıyaslama yapmıştır. Sohbet robotu 2 "*izobütil alkolün moleküler ağırlığı amil alkolden daha düşüktür … sonuç olarak, aynı ortamda bulunan 0,2 M izobütil alkol çözeltisinin yüzey gerilimi, 0,2 M amil alkol çözeltisinin yüzey geriliminden daha düşüktür*" cevabını vermiştir. Bu soruda her iki sohbet robotu da karbon sayısı ve moleküler yapıyla ilişkili bir açıklama yapmamıştır. Bu nedenle her iki robotun bu soru için puan ortalamaları 4,0'ten düşük kalarak soruyu yanlış cevaplandırdıkları kabul edilmiştir.

Araştırmada sohbet robotlarının performansını belirlerken dikkate alınan bir diğer veri, sohbet robotlarının yanlış veya kısmi yanlış cevaplarının nedenleridir. Bu amaçla cevaplar analiz edilmiş ve sonuçlar Şekil 4'te verilmiştir.



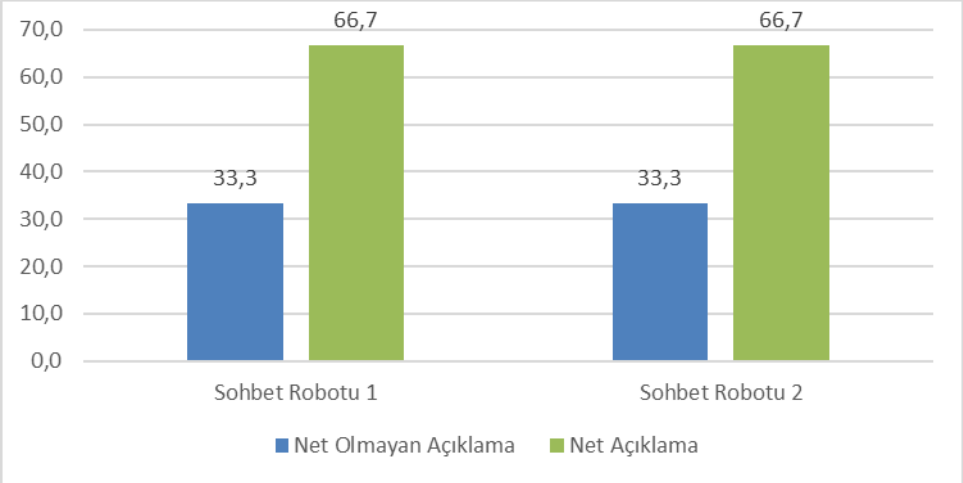
Şekil 4. Sohbet robotların yanlış veya kısmi yanlış cevaplarının nedenlerinin dağılımı

Şekil 4'e göre sohbet robotların yanlış veya kısmi yanlış cevap vermesinde (tam not alamamasında) en etkin olan faktörün yanlış bilgi olduğu görülmektedir. Her iki sohbet robotu da farklı sorularda yanlış bilgi içeren cevaplar vermişlerdir. Örneğin, sohbet robotu 2, benzen maddesinin yüzey geriliminden yararlanarak bilinmeyen bir sıvının yüzey geriliminin hesaplanması işlemi sırasında yüzey gerilimi ile bir damlanın kütlesini ilişkilendiren matematiksel formülü yanlış yazmış (yanlış bilgi) ve problemi yanlış cevaplamıştır. Bu soruda sohbet robotu 2'nin kullandığı yanlış formül "*Yüzey gerilimi (bilinmeyen sıvı)=Yüzey gerilimi (benzen)*(Kütle (benzen)/Kütle (bilinmeyen sıvı)*" şeklindedir. Bu formülde, damlaların kütleleri yanlış pay ve paydada yazılmıştır. Her ne kadar formülün bir kısmı yanlış olsa da cevaplardaki açıklamalardan kaynaklı uzmanlar bu soruya puan vermişlerdir. Bir başka soruda ise sohbet robotu 1, iki farklı çözeltinin yüzey gerilimini kıyaslarken "*daha büyük moleküllerin yüzey gerilimini artırma eğilimi vardır çünkü daha fazla molekül, yüzeyde daha fazla çekim kuvveti yaratma potansiyeline sahiptir*" cevabını vermiştir. Bu cevaba göre büyük moleküllerin yüzey gerilimini arttırma eğilimi vardır ve molekül sayısı (daha fazla molekül) ile yüzey gerilimi arasında bir ilişki mevcuttur. Ancak bu iki ifade de bilimsel olarak yanlıştır.

Sohbet robotlarının yanlış ve kısmi yanlış cevap vermesindeki ikinci faktör ise cevaplarda eksik bilgi bulunmasıdır. Örneğin, suyun yüzey geriliminden yararlanarak farklı alkol çözeltilerinin yüzey geriliminin kıyaslanabilmesi için bir deney tasarlamasının istendiği soruda sohbet robotu 1, stalagmometre kullanarak bir deney önermiştir. Önerisinde *"…İlk olarak, stalagmometre ile saf suyu kullanarak yüzey gerilimini ölçün ve kaydedin. Sonra, aynı stalagmometre ile ilgili alkolü kullanarak yüzey gerilimini ölçün ve kaydedin…"* cevabını vermiştir. Ancak deneyde stalagmometre ile yüzey geriliminin nasıl kıyaslanabileceği (damla sayısı ile ilişkilendirme) hakkında bilgi vermemiştir. Ayrıca aynı derişime sahip iki farklı alkol çözeltisinin yüzey gerilimlerinin kıyaslanmasının istenildiği soruda her iki sohbet robotunun karbon sayısı ve moleküler yapıyla ilişkili bir açıklama yapmamaları cevaplarda eksik bilgi olduğuna bir başka örnektir.

Ayrıca sohbet robotu 1'in yanlış ve kısmi yanlış cevap vermesinde yorumlama hatası da bulunmaktadır (Şekil 4). Örneğin sohbet robotu 1, yüzey gerilimi kavramını "*yüzey gerilimi, sıvının yüzeyinin bir tür gerilmiş bir zar gibi davrandığı ve bu nedenle yüzeydeki moleküllerin iç kısımdaki moleküllerden daha fazla enerjiye ihtiyaç duyduğu bir olgu*" olarak tanımlanmıştır. Bu açıklamaya göre yüzeydeki taneciklerin yüzeyde kalabilmesi için enerjiye ihtiyaçları vardır. Bir başka örnek olarak, "Gökten alkol yağacak olsa, aynı derişimde amil alkol, izobütil alkol ve bütil alkolden hangisinin yağmasını istersin" sorusuna sohbet robotu 2, "*Yüzey gerilimi, bir sıvının atmosferle etkileşimini ve damla oluştururken çekim gücünü etkileyen önemli bir faktördür. Daha düşük yüzey gerilimine sahip bir alkol, damlanın daha hızlı büyümesine ve daha büyük damlaların oluşmasına yol açabilir… Bu verilere dayanarak, izobütil alkolün en düşük yüzey gerilimine sahip olduğunu görüyoruz. Bu nedenle, izobütil alkolün yağması durumunda, daha büyük damlaların oluşma olasılığı daha yüksektir*" cevabını vermiştir. Bu cevapta düşük yüzey gerilimi ile damla büyüklüğü ve damla sayısı arasındaki ilişki yanlış yorumlanmıştır.

Son olarak araştırmada sohbet robotlarının verdikleri cevapların net olup olmadığı da araştırılmıştır ve analiz sonuçları Şekil 5'te verilmiştir. Araştırma sonucuna göre sohbet robotlarının çoğunlukla (%66,7) "net açıklamalar" verdikleri bazı cevaplarının ise net olmadığı (%33,3) görülmektedir. Örneğin, sıcaklıkla yüzey geriliminin nasıl değiştiğinin açıklanması istenilen soruda sohbet robotu 1, "*sıvıların yüzey gerilimi sıcaklıkla ilgili karmaşık bir şekilde değişebilir ve A sıvısının 25°C'deki yüzey geriliminin daha yüksek olması, özel kimyasal ve fiziksel özelliklerine bağlı olabilir. Bu tür bir durumu tam olarak anlamak için A sıvısının özel özelliklerini incelemek gerekebilir*" şeklinde bir cevap vermiştir. Bu cevaptaki "özel kimyasal özellikler" ve "özel özellikler" ifadelerin ne anlama geldiği anlaşılmamaktadır.



Şekil 5. Sohbet robotlarının açıklamalarının netliği

Sonuç olarak; araştırmadan elde edilen tüm bulgular özetlenecek olursa;

1. Sohbet robotları ölçekten yaklaşık olarak aynı ortalamaları almışlardır.

2. Sohbet robotlarının ölçekten aldıkları puanların birbirine oldukça yakın olduğu belirlenmiştir (60 puan üzerinden sırasıyla 35 puan ve 38 puan).
3. Her iki sohbet robotu da aynı soruyu yanlış cevaplamıştır ve bu soru "çözümleme" düzeyindedir.
4. En yüksek düzeyde cevapladıkları soru her ikisinde de aynı düzeyde ve "yaratma" düzeyidir.
5. Sohbet robotların cevaplarında yanlışlıklar, eksiklikler veya yorumlama hataları olabilmektedir.
6. Sohbet robotları genellikle net açıklamalar vermektedir.

Tartışma

Bu çalışmanın amacı, yapay zekâ tabanlı sohbet robotlarının yüzey gerilimi ile ilgili Bloom'un bilişsel alan taksonomisine uygun sorulara verdikleri cevapların analizinden sohbet robotlarının performanslarını belirlemektedir. Bu amaçla iki farklı sohbet robotuna altı adet açık uçlu soru aynı sırayla sorulmuştur. Sohbet robotlarının cevapları üç uzman tarafından analiz edilmiştir. Sohbet robotu 1'in soruları %58 oranında doğru cevapladığı, sohbet robotu 2'nin %64 oranında doğru cevapladığı belirlenmiştir. Bu araştırma için sohbet robotların performansının yaklaşık %60 olduğunu varsayarsak bu sonuç alan yazındaki çalışmalar ile çelişmektedir. Çünkü sohbet robotlarının %70-80 üzerinde doğrulukla cevap verdiklerini belirten çalışmalar mevcuttur (Das vd., 2023; Geerling vd., 2023; Jalil vd., 2023; Wood vd., 2023). Örneğin Das vd. (2023), ChatGPT'ye mikrobiyoloji ile ilgili 96 tane soru sormuşlardır. Bu soruların yarısı karmaşık ve üst düzey düşünme becerisi gerektiren sorudur. ChatGPT düzey fark etmeksizin bu soruların %80'ini doğru cevaplamıştır. Geerling vd. (2023) de sohbet robotlarının sorulan soruları yüksek başarı ile cevapladığını belirleyen bir çalışma gerçekleştirmiştir. Çalışmada, ChatGPT'ye ekonomi ile ilgili sorular sorulmuş ve sohbet robotunun mikroekonomi testindeki başarısı %63,3; makroekonomideki başarısı ise %86,7 olarak belirlenmiştir. Bu iki örnek, bu çalışmanın sonuçlarıyla örtüşmezken, farklılığın nedenini daha çok merak etmemize neden olmuştur.

Bu araştırmada belirlenen bir diğer sonuç da uzmanlar tarafından sohbet robotlarının cevaplarının, çoğunluğunun, kimya bilimi açısında "tam doğru" olarak değil, "kısmi doğru" olarak kabul edilmesidir. Araştırmanın bu iki bulgusunun nedeni kimya disiplini ile ilgili olabilir. Kimya disiplini, fizik ve biyoloji gibi, soyut kavramlar içeren, bilimsel modellere ve teorilere dayalı bir bilim dalıdır (Nakhleh, 1992; Taber, 2002, 2009). Kimya gibi fen alanındaki birçok çalışma öğrencilerin temel kavramları modellemeleri, olgu ve olayları anlamaları üzerine odaklanmıştır. Bu çalışmalar öğrencilerin birçok yanlış kavramaya ve yanlış modellemelere sahip olduğunu da ortaya koymuştur (Gilbert, 2004; Gilbert ve Watts, 1983; Greca ve Moreira, 2000). Düşünebilen, muhakeme edebilen ve yorum yapabilen öğrenciler bile kimya gibi derslerde zorlanıyorlarsa yapay zekânın hata yapması,

şimdilik, çok normaldir. Yapay zekâ tabanlı sohbet robotlarının kimya ile ilgili kavram/olay/olguları modelleyememesi sorulara yanlış cevap vermesine neden olmuş olabilir. Gregorcic ve Pendrill'e (2023) göre, sohbet robotları her ne kadar "insan gibi düşünüyor" gibi görünse de gerçek durum böyle değildir; sadece "kavramları" çok iyi kullanıp "düşünüyor" gibi görünmektedir. Kavramları kullanmakta, bir muhakeme sürecini izlemekte ama düşünüp anlayamadığı için yanlış ve/veya eksik sonuçlara ulaşmaktadır. Gregorcic ve Pendrill (2023) fizik konusu ile ilgili benzer bir çalışma yaparak bu yorumu ortaya atmıştır. Çalışmada fizik konusu ile ilgili ChatGPT'ye bir dizi soru sormuşlar ve ChatGPT'nin çok basit fizik sorularına dahi cevap veremediğini ve hatta öğrencilerle benzer hatalar yaptığını belirtmişlerdir. Humphry ve Fuller (2023) de sohbet robotlarının kimya ile ilgili konularda çok iyi olmadığını belirtmiştir. Bu sonuçlar, mevcut araştırmanın bulguları ile örtüşmektedir.

Araştırmada, sohbet robotlarının Bloom'un bilişsel alan taksonomisine göre hazırlanmış sorulardaki performansı da incelenmiştir. Analiz sonuçlarına göre, her iki yapay zekâ tabanlı sohbet robotunun "çözümleme" düzeyindeki soruyu yanlış ve "yaratma" düzeyindeki soruyu ise en yüksek doğrulukta cevapladıkları belirlenmiştir. Açıkçası bu beklenmedik bir durumdur. Fergus vd.'nin (2023) çalışmasında ChatGPT'ye kimya ile ilgili sorular sorulmuş ve Bloom'un bilişsel alan taksonomisindeki "hatırlama ve anlama" düzeyindeki sorulara cevap verebildiğini ancak "uygulama" düzeyindeki birçok soruya cevap veremediği belirlenmiştir. Yazarlar, ayrıca sohbet robotunun yorumlama becerisi gerektiren sorularda başarısız olduklarını da belirtmişlerdir. Mevcut araştırmada, "çözümleme" düzeyindeki soru yorumlama gerektiren bir soru olduğu için sohbet robotları başarısız olmuş olabilir.

Bir diğer araştırma sorusu da sohbet robotlarının cevaplarının neden yanlış olduğunun belirlenmesidir. Sohbet robotlarının her ikisinin cevapları eksik bilgi ve yanlış bilgi (hatta yanlış kavramaya neden olabilecek ifadeler) içermektedir. Ayrıca, sohbet robotu 1 yorumlama hatası da yapmıştır. Sohbet robotları sorulara cevap verirken internet tabanlı araştırmalar yapmaktadır. İnternette bulunan her bilginin doğru olduğunu kabul etmek çok yanlıştır (Acar-Sesen ve Ince, 2010). Bu nedenle sohbet robotları yanlış bilgi içeren cevaplar vermiş olabilir. Gungordu, Yalcin-Celik ve Kilic (2017) öğrencilerin çevre konusundaki yanlış kavramalarının kaynaklarını belirlemek amaçlı yaptığı çalışmada internet tabanlı medyada birçok yanlış bilginin bulunduğunu belirlemiştir. Hatta bu yanlış bilgilerin, öğrencilerin yanlış kavramalarına sebep olduğunu ifade etmiştir. Clark (2023) da sohbet robotlarının çoklu kavram içeren ve yaratıcı problem çözme becerisi gerektiren sorularda yetersiz kaldığını hatta öğrencilerde belirlenen yanlış kavramaların, sohbet robotlarının verdikleri cevaplarda da bulunduğunu belirlemiştir. Sohbet robotları her ne kadar bir insan gibi düşünemiyor olsa da (Gregorcic ve Pendrill, 2023) kendilerine sorulan sorulara eriştikleri bilgilerden, farklı modellerle (kural tabanlı model, erişim modeli, üretken modeller vb.) yanıt oluşturmaktadır (Hussain,

Ameri-Sianaki ve Ababneh, 2019). Bu yanıtlarda yanlış bilgi, eksik bilgi veya yorumlama hataları da olabilmektedir. Araştırmanın son bulgusu sohbet robotlarının cevaplarının anlaşılırlığının düşük olmadığıdır. Her iki sohbet robotu da %66,7 düzeyinde anlaşılır cevap vermişlerdir. Bu sonuç sohbet robotlarının kimya disiplininin kavramlarını etkili bir şekilde kullanabildiğinin bir göstergesidir.

Sonuç ve Öneriler

Sonuç olarak, araştırma bulgularına göre sohbet robotları kimya ile ilgili soruları cevaplamakta önemli bir performans gösterememişlerdir. Eğitim alanında yapay zekâ araçlarının veya hâlihazırda kullanılan sohbet robotlarının her alanda yüksek başarı gösteremediğinin, yanlış/eksik bilgi içeren ve yanlış kavramalara neden olabilecek cevapları verebildiğinin eğitimciler ve öğrenciler tarafından bilinmesi araştırmacılar tarafından önemli görülmektedir. Belki de bu şekilde sohbet robotlarının öğrenciler tarafından etik dışı davranış gösteren kullanımlarının da önüne geçilebilir.

Yapay zekâ ile çalışan araştırmacılara birkaç öneride bulunulacak olursa; bu araştırmanın önemli bir sınırlılığı sohbet robotlarının kimya alanındaki performansını sadece bir konuya özgü sorularla belirlemeye çalışmasıdır. Yapay zekâ araçlarının performansının zorluk seviyesi kolaydan zora doğru olan farklı konularda belirlendiği çalışmalara ihtiyaç vardır. İlaveten, sohbet robotlarına her bir soru için sadece bir kez istem girilmiştir. Cevapların birkaç kez yenilenmesinin istenmesinin daha doğru ve daha net (anlaşılır) cevapların üretilmesine etki edip etmediğinin araştırıldığı çalışmalara da ihtiyaç vardır. Son olarak sohbet robotlarının cevaplarında yanlış kavramaların olup olmadığı da araştırılabilir. 1990'larda ve 2000'lerde alan yazın öğrencilerin kavram yanılgılarını tespit etmeye odaklanmıştı; belki de sıra artık sohbet robotlarınındır.

Kaynaklar

Acar-Sesen, B. & Ince, E. (2010). Internet as a source of misconception. *Turkish Online Journal of Educational Technology-TOJET, 9*(4), 94-100.

AlAfnan, M. A., Dishari, S., Jovic, M. & Lomidze, K. (2023). Chatgpt as an educational tool: Opportunities, challenges, and recommendations for communication, business writing, and composition courses. *Journal of Artificial Intelligence and Technology, 3*(2), 60-68. https://doi.org/10.37965/jait.2023.0184

Anderson, L. W., Krathwohl, D. R., Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., Raths, J. & Wittrock, M. C. (Ed.). (2001). *A taxonomy for learning, teaching and assessing. A Revision of Bloom's Taxonomy of educational objectives.* United States: Longman Publishing.

Clark, T. M. (2023). Investigating the use of an artificial intelligence chatbot with general chemistry exam questions. *Journal of Chemical Education, 100*(5), 1905-1916. https://doi.org/10.1021/acs.jchemed.3c00027

Das, D., Kumar, N., Longjam, L. A., Sinha, R., Roy, A. D., Mondal, H. & Gupta, P. (2023). Assessing the capability of ChatGPT in answering first- and second-order knowledge questions on microbiology as per Competency-Based Medical Education Curriculum. *Cureus, 15*(3), e36034. https://doi.org/10.7759/cureus.36034

Fergus, S., Botha, M. & Ostovar, M. (2023). Evaluating academic answers generated using ChatGPT. *Journal of Chemical Education, 100*(4), 1672–1675. https://doi.org/10.1021/acs.jchemed.3c00087

Geerling, W., Mateer, G. D., Wooten, J. & Damodaran, N. (2023). Is ChatGPT smarter than a student in principles of economics? *SSRN Electronic Journal*, 1-24. https://doi.org/10.2139/ssrn.4356034

Gilbert, J. K. (2004). Models and modelling: Routes to more authentic science education. *International Journal of Science and Mathematics Education, 2*, 115–130.

Gilbert, J. K. & Watts, D. M. (1983). Concepts, misconceptions and alternative conceptions: Changing perspectives in science education. *Studies in Science Education, 10*(1), 61-98. https://doi.org/10.1080/03057268308559905

Greca, I. M. & Moreira, M. A. (2000). Mental models, conceptual models, and modelling. *International Journal of Science Education, 22*(1), 1–11.

Gregorcic, B. & Pendrill, A. M. (2023). ChatGPT and the frustrated Socrates. *Physics Education, 58*(3), 035021. https://doi.org/10.1088/1361-6552/acc299

Gungordu, N., Yalcin-Celik, A. & Kilic, Z. (2017). Students' misconceptions about the ozone layer and the effect of internet-based media on it. *International Electronic Journal of Environmental Education*, *7*(1), 1-16.

Haenlein, M. & Kaplan, A. (2019). A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. *California Management Review, 61*(4), 5-14. https://doi.org/10.1177/0008125619864

Han, Z., Battaglia, F., Udaiyar, A., Fooks, A. & Terlecky, S. R. (2023). An explorative assessment of ChatGPT as an aid in medical education: Use it with caution. https://www.medrxiv.org/content/10.1101/2023.02.13.23285879v1.full.pdf sayfasından erişilmiştir.

Humphry, T. & Fuller, A. L. (2023). Potential ChatGPT use in undergraduate chemistry laboratories. *Journal of Chemical Education, 100*(4), 1434–1436.

Hussain, S., Ameri-Sianaki, O. & Ababneh, N. (2019). A survey on conversational agents/chatbots classification and design techniques. *Web, Artificial Intelligence and Network Applications: Proceedings of the Workshops of the 33rd International Conference on Advanced Information Networking and Applications (WAINA-2019) 33* içinde (s. 946-956). New York: Springer International Publishing.

Jalil, S., Rafi, S., LaToza, T. D., Moran, K. & Lam, W. (2023). ChatGPT and software testing education: Promises & perils. *2023 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW)* içinde (s. 4130-4137). New York: IEEE.

Koo, T. K. & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155-163.

Korsakova, E., Sokolovskaya, O., Minakova, D., Gavronskaya, Y., Maksimenko, N. & Kurushkin, M. (2022). Chemist bot as a helpful personal online training tool for the final chemistry examination. *Journal of Chemical Education, 99*(2), 1110-1117. https://doi.org/10.1021/acs.jchemed.1c00789

Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., ... & Tseng, V. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digit Health 2*(2), e0000198. https://doi.org/10.1371/journal.pdig.0000198

Lo, C. K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences, 13*(4), 410-455. https://doi.org/10.3390/educsci13040410

Meço, G. & Coştu, F. (2002). Eğitimde yapay zekânın kullanılması: Betimsel içerik analizi çalışması. *Karadeniz Teknik Üniversitesi Sosyal Bilimler Enstitüsü Sosyal Bilimler Dergisi, 12*(23), 171-193.

Mhlanga, D. (2023). Open AI in education, the responsible and ethical use of ChatGPT towards lifelong learning. https://www.researchgate.net/profile/David-Mhlanga-2/publication/368476294_Open_AI_in_Education_the_Responsible_and_Ethical_Use_of_ChatGPT_Towards_Lifelong_Learning/links/63eb1c91bd7860764366f597/Open-AI-in-Education-the-Responsible-and-Ethical-Use-of-ChatGPT-Towards-Lifelong-Learning.pdf sayfasından erişilmiştir.

Miles, M. B. & Huberman, A. M. (1994). *Qualitative data analysis* (2. b.). California: SAGE.

Motlagh, N. Y., Khajavi, M., Sharifi, A. & Ahmadi, M. (2023). The impact of artificial intelligence on the evolution of digital education: A comparative study of openAI text generation tools including ChatGPT, Bing Chat, Bard, and Ernie. https://www.researchgate.net/profile/Negin-Yazdani-Motlagh/publication/373800938_The_Impact_of_Artificial_Intelligence_on_the_Evolution_of_Digital_Education_A_Comparative_Study_of_OpenAI_Text_Generation_Tools_

including_ChatGPT_Bing_Chat_Bard_and_Ernie/links/64fd0144d6fa5c5bc46cfdbd/The-Impact-of-Artificial-Intelligence-on-the-Evolution-of-Digital-Education-A-Comparative-Study-of-OpenAI-Text-Generation-Tools-including-ChatGPT-Bing-Chat-Bard-and-Ernie.pdf sayfasından erişilmiştir.

Nakhleh, M. B. (1992). Why some students don't learn chemistry: Chemical misconceptions. *Journal of Chemical Education, 69*(3), 191-195.

Rahaman, M. S., Ahsan, M. M., Anjum, N., Rahman, M. M. & Rahman, M. N. (2023). The AI race is on! Google's Bard and OpenAI's ChatGPT head to head. http://dx.doi.org/10.2139/ssrn.4351785

Rudolph, J., Tan, S. & Tan, S. (2023). War of the chatbots: Bard, Bing Chat, ChatGPT, Ernie and beyond. The new AI gold rush and its impact on higher education. *Journal of Applied Learning and Teaching, 6*(1), 364-389. https://doi.org/10.37074/jalt.2023.6.1.23

Shawar, B. A. & Atwell, E. (2007). Chatbots: Are they really useful? *Journal for Language Technology and Computational Linguistics, 22*(1), 29–49.

Steenbergen-Hu, S. & Cooper, H. A. (2014). Meta-analysis of the effectiveness of intelligent tutoring systems on college students' academic learning. *Journal of Educational Psychology, 106*(2), 331–347.

Susnjak, T. (2022). ChatGPT: The end of online exam integrity? *Cornell University arXiv.* https://doi.org/10.48550/arXiv.2212.09292

Suta, P., Lan, X., Wu, B., Mongkolnam, P. & Chan, J. H. (2020). An overview of machine learning in chatbots. *International Journal of Mechanical Engineering and Robotics Research, 9*(4), 502-510. https://doi.org/10.18178/ijmerr.9.4.502-510

Taber, K. (2002). *Chemical misconceptions: prevention, diagnosis and cure* (c. 1). Londra: Royal Society of Chemistry.

Taber, K. (2009). Challenging misconceptions in the chemistry classroom: resources to support teachers. *Educació Química*(4), 13-20.

Tamer, H. Y. & Övgün, B. (2020). Yapay zekâ bağlamında dijital dönüşüm ofisi. *Ankara Üniversitesi SBF Dergisi, 75*(2), 775-803. https://doi.org/10.33630/ausbf.691119

Wood, D. A., Achhpilia, M. P., Adams, M. T., Aghazadeh, S., Akinyele, K., Akpan, M., ... & Kuruppu, C. (2023). The ChatGPT artificial intelligence chatbot: How well does it answer accounting assessment questions? *Issues in Accounting Education, 38*(4), 81-108. https://doi.org/10.2308/ISSUES-2023-013

Yıldırım, A. & Şimşek, H. (2018). *Sosyal bilimlerde nitel araştırma yöntemleri* (11. b.). Ankara: Seçkin Yayıncılık.

Zhu, J. J., Jiang, J., Yang, M. & Ren, Z. J. (2023). ChatGPT and environmental research. *Environmental Science & Technology* [Special Issue], A-D. https://doi.org/10.1021/acs.est.3c01818

Extended Summary

As of 2022, chatbots have been an artificial intelligence tool used in education (Suta, Lan, Wu, Mongkolman and Chan, 2020). With the use of artificial intelligence-based chatbots in education, new issues have emerged for educators (Gregorcic and Pendrill, 2023). As a chat tool, chatbots answer students' questions, prepare assignments, or answer exam questions (Fergus et al., 2023). However, there is criticism that the answers prepared by chatbots are not always correct, that they make comments based on incorrect or incomplete data, or that they produce false or untrue information (Han, Battaglia, Udaiyar, Fooks and Terlecky, 2023, p. 32). Therefore, this research focuses on determining the performance of chatbots in chemistry. This research aims to determine the performance of two different AI-based chatbots at the university level on questions prepared on surface tension, taking into account Bloom's cognitive domain taxonomy. The aim of this research, which involves using two chatbots, is to provide teachers, students and educators with information about how successful they can be in chemistry by revealing the level of performance of chatbots in chemistry.

The research was conducted using a case study design, which is one of the qualitative research methods (Yıldırım and Şimşek, 2018). The research used ChatGPT, introduced by OpenAI, and Bard chatbots, introduced by Google. These two chatbots were preferred because of being easily accessible and having Turkish language support. The study used a scale with six open-ended questions about surface tension as a data collection tool. The questions were prepared by considering Bloom's revised cognitive domain taxonomy. Two experts in the field of science were consulted on the appropriateness of the questions. The research was carried out in three stages. The first stage was preparing the questions to be posed to the chatbots. The second stage was the stage of receiving the answers from the chatbots. In this stage, a user account was opened for one of the researchers on both chatbots, and each prompt was written one by one in the same chat section. The prompts in this study were the questions in the data collection tool used by the chatbot to generate responses. The questions were asked to the chatbots in Turkish during the same line of questioning, in the same chat section, on the same day, and only once, and the robots were not asked to generate the answers again or different drafts of the answers. The third stage was the stage where the performance of the chatbots was determined. At this stage, three experts who taught physical chemistry and physical chemistry laboratory courses and who studied the subject were asked to rate the answers of the chatbots, starting from the first question and giving each question a score between 0 and 10. To determine the inter-rater

reliability from the scores given by the experts, the intraclass correlation coefficient was calculated as 0.67 (moderate level) and 0.78 (good level) for the chatbots, respectively (Koo and Li, 2016).

According to the research results, chatbot 1 scored 35 out of 60 (58 out of 100) and chatbot 2 scored 38 out of 60 (64 out of 100). When analyzing the chatbots' performance at Bloom's Cognitive Domain Taxonomy levels, it was found that both chatbots received similar average scores at the same taxonomy level. The question for which the chatbots received the highest score was the question belonging to the 'creating' level, while the question for which they received the lowest score was the question belonging to the 'analyzing' level. It was also found that the answers given by both chatbots were mostly partially correct (83.33% and 66.67%, respectively). Another data point used in the study to determine the performance of the chatbots was the reasons for incorrect or partially incorrect answers given by the chatbots. According to the results of the analyses, the most influential factor in the failure of chatbots to give correct answers was misinformation. It was followed by insufficient information. An interpretation error was also responsible for the incorrect answers of the chatbot 1. Finally, it was found that chatbots mostly gave 'clear explanations' (66.7%), and some of their answers were unclear (33.3%).

Taken together, the performance of the chatbots was around 60%, and most of their answers were accepted as 'partially correct', suggesting that the reason for these two findings may be related to the discipline of chemistry. The discipline of chemistry, like physics and biology, is a science that involves abstract concepts and is based on scientific models and theories (Nakhleh, 1992; Taber, 2002, 2009). Many studies in sciences such as chemistry have focused on students' modelling of basic concepts and their understanding of phenomena and events and have shown that students have many misconceptions (Gilbert, 2004; Gilbert and Watts, 1983; Greca and Moreira, 2000). If even students who can think, reason and comment have difficulties in chemistry courses, it is normal for artificial intelligence to make mistakes. The inability of the chatbots to model the concepts/events/phenomena related to chemistry may have led them to give incorrect answers to the questions. According to Gregorcic and Pendrill (2023), although chatbots appear to "think like humans", this is not the case; they only use "concepts" very well and appear to "think", use concepts, follow a reasoning process, but arrive at wrong and incomplete results because they cannot think and understand. Humphry and Fuller (2023) also found that chatbots were not very good at chemistry-related topics. These findings are consistent with the results of the current study.

It was also found that both chatbots answered the question incorrectly at the 'analyze' level and with the highest accuracy at the 'create' level. This is an unexpected situation. Fergus et al. (2023) also found that chatbots failed on questions requiring interpretation skills. Chatbots perform web-based research while answering questions. Not all information on the internet is correct (Acar-Sesen

and Ince, 2010). Therefore, chatbots may have given answers that contained incorrect information. Gungordu et al. (2019), in their study to determine the sources of students' misconceptions about the environment, stated that there was much misinformation in internet-based media, which causes students' misconceptions. Clark (2023) also found that chatbots were inadequate for questions that involved multiple concepts and required creative problem-solving skills, and even that the misconceptions identified in students were also found in the answers given by chatbots. Although chatbots cannot think like a human (Gregorcic and Pendrill, 2023), they form answers to the questions posed to them using different models (rule-based model, access model, generative models, etc.) from the information they access (Hussain et al., 2019). These responses may contain incorrect information, incomplete information, or misinterpretations.

As a result, according to the research findings, chatbots could not perform significantly in answering chemistry-related questions. We believe that educators and students must be aware that artificial intelligence tools or chatbots currently used in education may not be highly successful in all areas. They may sometimes provide answers that contain incorrect/completed information or even misconceptions. Perhaps in this way, we can prevent using chatbots by students who behave unethically.

As a suggestion, it is possible to investigate whether there are misconceptions in the answers given by chatbots. In the 1990s and 2000s, the literature focused on identifying students' misconceptions; perhaps it is now the chatbots' turn.

Ekler

Yapay Zekâ Tabanlı Sohbet Robotlara Sorulan Yüzey Gerilimi Konusu ile İlgili Sorular

1) Yüzey gerilimi nedir? Tanımlar mısın?

2) A sıvısının, 25 ^{o}C'daki yüzey geriliminin aynı sıvının 125 ^{o}C'daki yüzey geriliminden daha yüksek olmasını nasıl açıklarsın?

3) Aynı stalogmometreden koparak düşen benzen ve bilinmeyen bir sıvı damlalarının 20 ^{o}C'daki ortalama kütleleri sırasıyla 0,0520 g ve 0,0420 g olarak ölçülmüştür. Benzenin deney sıcaklığındaki yüzey gerilimi 28,9 dyn/cm olduğuna göre diğer sıvının yüzey gerilimi nedir?

4) Aynı ortamda bulunan 0,2 M izobütil alkol ile 0,2 M amil alkol çözeltilerinin yüzey gerilimlerini kıyaslar mısın?

5) Gökten alkol yağacak olsa, aynı derişimde amil alkol, izobütil alkol ve bütil alkolden hangisinin yağmasını istersin? Bu soruya alkollerin yüzey gerilimlerini ve bir damlanın kütlesini dikkate alarak cevap verebilir misin? Bu alkol yağmurunda bir insan olsan hangisini tercih ederdin?

6) 0,1 M bütil alkol çözeltisi, 0,1 M izobütil alkol çözeltisi ve 0,1 M amil alkol çözeltilerinin yüzey gerilimini, suyun yüzey geriliminden yararlanarak kıyaslamak istiyorum. Bütil alkol, izobütanol, amil alkol çözeltilerinin yüzey gerilimini kıyaslayabilmek için nasıl bir deney tasarısı önerirsin?

Araştırmacıların Katkı Oranı Beyanı

Bu araştırmanın planlanması, yürütülmesi ve yazılı hale getirilmesinde araştırmacılar eşit oranda katkı sağlamıştır.

Destek ve Teşekkür Beyanı

Bu araştırmada herhangi bir kurum, kuruluş ya da kişiden destek alınmamıştır.

Çatışma Beyanı

Araştırmacıların, araştırma ile ilgili diğer kişi ve kurumlarla herhangi bir kişisel ve finansal çıkar çatışması yoktur.

Etik Kurul Beyanı

Bu araştırma verileri güncel, açık kaynaklı yapay zekâ tabanlı sohbet robotlarından elde edildiği için etik kurul izni gerektirmemektedir.