

CHATGPT'NİN FARKLI BÜYÜK DİL MODELLERİ PERFORMANSLARININ TÜRKÇEDEKİ EŞ ADLI KELİMELER ÜZERİNDEN İNCELENMESİ

Çiğdem AYTEKİN
Marmara Üniversitesi, Türkiye
cigdem.aytekin@marmara.edu.tr
<https://orcid.org/0000-0002-1385-9864>

Talha Bedir KARABİNA
Marmara Üniversitesi, Türkiye
talhakarabina@marun.edu.tr
<https://orcid.org/0009-0001-9730-5083>

<i>Atf</i>	AYTEKİN, Ç.; KARABİNA, T., B. (2024). CHATGPT'NİN FARKLI BÜYÜK DİL MODELLERİ PERFORMANSLARININ TÜRKÇEDEKİ EŞ ADLI KELİMELER ÜZERİNDEN İNCELENMESİ. <i>Istanbul Aydın Üniversitesi Sosyal Bilimler Dergisi</i> , 16(3), 365-390.
------------	---

ÖZ

Son zamanların popüler konusu ChatGPT ve gerçekleştirdiği başarılı işler, yapay zekânın ne kadar geliştiğini ve ilerleyen yıllar için vadettiklerini bizlere göstermektedir. ChatGPT'nin hâlihazırda kullanılan Büyük Dil Modelleri arasındaki farklılıklar bu çalışmanın konusunu oluşturmaktadır. ChatGPT-3.5 ve ChatGPT-4'ün performansları Türkçedeki eş adlı kelimeler üzerinden incelenmiştir. Büyük Dil Modelleri oluşturulurken kullanılan Doğal Dil İşleme sistemlerinde aşılması en büyük zorluklardan birisi de bu sistemlerin kelime-anlam belirsizliğini ayırt edebilme becerileridir. Bu belirsizlikleri tespit etmek amacıyla Türkçede en yaygın olarak kullanılan 200 eş adlı kelime örneklem olarak seçilmiştir. Ardından tek bir eş adlı kelimenin, aynı cümle içerisinde iki farklı anlama da gelecek şekilde iki kez kullanılmasıyla cümleler oluşturulmuş ve öncelikle ChatGPT-3.5'den sonra ChatGPT-4'den farklı anlamları tespit etmesi istenmiştir. ChatGPT'ler her iki anlamdan birini bilemediği ve bazen iki anlamı da bilemediği çıktılar üretmiştir. Amaç doğrultusunda ChatGPT-3.5 ve ChatGPT-4 modellerinden alınan çıktılar karşılaştırılmıştır. ChatGPT 3.5'e kıyasla daha fazla parametreye ve veri setine sahip olan ChatGPT-4, beklendiği gibi çok daha iyi bir performans göstermiştir. Başarı oranı dağılım analizi, eş adlı kelimeye göre performans değişikliği, eş adlı kelimenin karakter sayısı ve başarı oranı, istatistiksel testler yapılan diğer analizlerdir.

Anahtar Kelimeler: ChatGPT-3.5, ChatGPT-4, Büyük Dil modeli, Eş Adlı Kelime, Dil Bilimsel Belirsizlik.

ANALYZING THE PERFORMANCE OF DIFFERENT LARGE LANGUAGE MODELS OF CHATGPT ON TURKISH HOMONYMS

ABSTRACT

ChatGPT, the popular topic in recent periods, and its achievements show us how much artificial intelligence has developed and what it promises for the coming years. This study focuses on the differences between ChatGPT and its currently used Large Language Models. The performances of ChatGPT-3.5 and ChatGPT-4 are analyzed on Turkish homonyms. One major challenge faced by Natural Language Processing systems used in the generation of Large Language Models is identifying word-sense ambiguity. In order to detect these ambiguities, the 200 most commonly used synonyms in Turkish were selected as the sample. Then, sentences were formed by using a single homonym twice in the same sentence to convey two different meanings, and ChatGPT-3.5 and then ChatGPT-4 were asked to detect the different meanings. ChatGPTs generated outputs in which they could not know either of the two meanings and sometimes could not know both meanings. In line with the objective, the outputs from ChatGPT-3.5 and ChatGPT-4 models were compared. As expected, ChatGPT-4, with its larger parameters and datasets, outperformed ChatGPT-3.5. Success rate distribution analysis, performance variation based on the homonym, the number of characters of the homonym and the success rate are the other statistical tests carried out.

Keywords: *ChatGPT-3.5, ChatGPT-4, Large Language Model, Homonym, Linguistic Ambiguity.*

GİRİŞ

Yapay zekâ teknolojilerindeki gelişmeler günümüzde her ne kadar daha çok bilgisayar mühendisliği alanı ile ilişkilendirilse de kavram aslında başta iletişim olmak üzere sağlık, endüstri, dil bilim, eğitim, psikoloji vb. gibi daha birçok alanla ilgilidir. Dolayısı ile her bilim alanı kendi bakış açısı ile araştırmalar yapmakta ve insanlık için olası çeşitli senaryoları uzun bir süredir tartışmaktadır.

ChatGPT, OpenAI tarafından geliştirilen bir yapay zekâ sohbet robotudur. Kullanıma sunulduğu günden beri de üzerinde pek çok araştırma yapılmıştır. Bunların önemli bir kısmı onun başarılarını (örneğin, farklı versiyonların performanslarını karşılaştırma suretiyle vb.) ele alarak deterministik bir yaklaşım sunmakta, bir kısmı da (örneğin, cevabını bilmediği sorularla karşılaştığında halüsinasyon görme eğilimini olumsuzlayarak veya etik üzerine etkisini tartışarak veya eğitim alanında yaratacağı hazırcılığı tenkit ederek vb.) eleştirel bir bakış açısı getirmektedir. Bu çalışma tekno-iyimser bir bakış açısına sahiptir.

GPT'nin Büyük Dil Modelleri (Large Language Models-LLM) metin üretme, soru-cevap uygulamaları ve duygu analizi gibi birçok doğal dil işleme görevinde başarı göstermişlerdir. Büyük Dil Modelleri oluşturulurken kullanılan Doğal Dil İşleme sistemlerinde aşılması en büyük zorluklardan birisi de bu sistemlerin kelime-anlam belirsizliğini ayırt edebilme becerileridir. Bu çalışmada kelime-anlam belirsizliğinin ayırt edilebilme başarısı ChatGPT-3.5 ve ChatGPT-4'ün performansları üzerinden incelenmiştir. Bu doğrultuda Türkçedeki eş adlı kelimeler (söylenişleri ve yazılışları aynı, anlamları farklı kelimeler) örneklem olarak seçilmiş ve her iki modelin bu farklı anlamları tespit etmeleri istenerek ürettikleri çıktılarının doğruluk oranları karşılaştırılmıştır. ChatGPT 3.5'e kıyasla daha fazla parametreye ve veri setine sahip olan ChatGPT-4, beklendiği gibi çok daha iyi bir performans göstermiştir.

Çalışmanın inşasındaki kuramsal çerçeve ChatGPT'nin gelişimi ve büyük dil modelleri arasındaki farklar ile eş adlılık ve belirsizlikler üzerine kuruludur ve kendi alan yazımına bu doğrultuda başvurmaktadır. Çalışma, Türkçe dilinde eş adlı kelimeler üzerinden Büyük Dil Modellerinin performans farkını ortaya koymaya öncülük etmektedir.

CHATGPT'NİN GELİŞİMİ VE BÜYÜK DİL MODELLERİ ARASINDAKİ FARKLAR

ChatGPT'nin yaygın bir şekilde kullanılmaya başladığı zamanlarda, yapay zekânın insanların yerini alıp almayacağı konusu yine gündeme gelmişti. Bu durum hakkındaki tartışmalardan birisi de Elon Musk ve eski Google CEO'su Larry Page arasında 2015 yılında daha ortada ChatGPT yokken gerçekleşmişti. Bu tartışmanın dolaylı bir sonucu olarak OpenAI kurulmuş ve başarısını ChatGPT üzerinde göstermiştir (Metz vd., 2023).

ChatGPT'nin halihazırda kullanımda olan sohbet robotu özelliği uzun bir çalışma sürecinin sonucunda ortaya çıkan bir üründür. GPT 3.5'den önce OpenAI şirketi, GPT-3 gibi sürümleri kendi internet sayfasında yazılım geliştiricilerinin kullanımı için program tasarlama arayüzü veya API olarak sunuyordu. Sohbet robotu çıkmadan önceki sürümler geniş kitleler tarafından bilinmiyordu. 2022'nin Kasım ayında sohbet robotu piyasaya sürüldüğünde OpenAI'ın bile beklemediği bir ilgiyle karşılandı (Heaven, 2023).

ChatGPT'nin bu kadar popüler olmasının sebeplerinden birisi de o güne kadar yapılmış en kapsamlı ve başarılı olarak çalışan yapay zekâ sohbet robotu olmasıydı. Robot, kullanıcıların bir sohbet robotuyla değil de bir insanla veya iş arkadaşıyla konuşuyormuş hissine kapılmasını sağlamıştı. Bu başarısının ardında şüphesiz yıllar süren bir çalışma vardı. Bu çalışmalar kullanıcı bildirimleri ve yeni teknolojik gelişmeler sayesinde halen daha devam etmektedir (Makridakis vd., 2023, s. 536).

OpenAI GPT-1, GPT-2, GPT-3 ve GPT-4 gibi farklı modeller yayınlamıştır. Her birinin eğitimi için kullanılan gittikçe artan veri seti bulunmaktadır. ChatGPT, "RLHF" (Reinforcement Learning from Human Feedback) adı verilen İnsan Geri Bildiriminden Destekli Öğrenme metoduyla geliştirilmiştir. İlk modeller yapay zekâyı oluşturan kişilerin hem kullanıcı hem de eğitici rolünde sohbet ettiği biçimde denetimle fine-tuning (*ince ayar*) adı verilen işlemle oluşturulmuştur. Bunun üzerine yine GPT mimarisini kullanan InstructGPT'yi de tanıtan OpenAI, onun ChatGPT'den farkını farklı odak noktaları ve eğitimleri üzerinden belirtmiştir. InstructGPT adından da anlaşılacağı üzere daha çok yönerge metni üretmeye odaklanmış ve adım adım rehberlik sağlamak amacıyla oluşturulmuştur. Buna istinaden ChatGPT daha genel amaçlı ve çeşitli metin tabanlı istemler ile sohbet üzerine kurulmuştur (Alawida vd., 2023, s. 3).

OpenAI, bir süre sonra GPT-4 mimarisini kullanan ve ücretsiz modeli GPT-3.5'ten daha gelişmiş ve güncel bir sürüm olarak hizmet veren ChatGPT Plus'ı tanıttı. Bu hizmet ile birlikte artık kullanıcılar daha fazla seçeneğe sahip olarak ChatGPT'yi eğitebiliyor ve bunun üzerine belirli alanlarda ChatGPT'yi amaçlarına uygun olarak kullanabiliyorlar. Yapabildiği görevler kendisine sorulduğunda ise cevabı şu şekildedir: Otomatik müşteri desteği ve hizmeti sağlamak, kullanıcı verisini analiz ederek kişisel içerik sunmak, bir dilden diğerine çeviri yapmak, yaratıcı içerik üretmek ve büyük veri setlerini inceleyip öngörüler sağlayarak araştırma ve analize yardımcı olmak (Makridakis vd., 2023, s. 545).

Karmaşık sorulara cevap vermesi bakımından GPT-4, GPT-3.5'e göre çok daha iyi bir performans göstermektedir. Çünkü GPT-4 aynı anda birçok farklı kaynaktan veri çekerek bunları sentezlemekte ve böylece karmaşık durumlarla başa çıkabilmektedir. GPT-4 multimodal (çok modlu) bir model üzerine kurulmuştur.

Yani, metin tabanlı olmanın yanı sıra görseller, videolar veya Excel dosyaları işleyebilmekte ve bunları üretebilmektedir. Bu özellikleri sayesinde kurumlar, reklam metni yazarları ve özel ya da profesyonel işlerini yapan kişiler tarafından kullanılmaktadır (Alawida vd., 2023, s. 9).

GPT-4 işlem kapasitesi, kelime haznesi ve eğitiminde kullanılan parametre sayısı gibi faktörlerle kıyas yapıldığında GPT-3'den nicel olarak öndedir. GPT-4 eğitilirken önceki modellere oranla fine-tuning yapılmış veri setleri ve işlemleri yüksek düzeyde kullanılmıştır. Kullanılan parametre sayısı da GPT-3 için 175 milyar iken GPT-4 için 300 milyardır (Alawida vd., 2023, s. 4).

Bu özellikler ışığında GPT-4, daha uzun metin istemlerini anlayabilmekte ve bağlamdan kopmadan cevap verebilmektedir. Fine-tuning sayesinde çeşitli konularda ve belirli alanlarda daha net sonuç sunmaktadır. Belirsizliği daha iyi anlayabilmekte ve daha net çıktılar üretebilmektedir. Aynı zamanda belirli güvenlik önlemleriyle kişilere zarar verebilecek veya etik değerlere karşı gelebilecek cevaplar vermesinin de önüne geçilmiştir (Plevris vd., 2023, s. 954).

Her ne kadar modeller arasındaki bu gözle görülür fark artmış olsa da GPT-3 ve GPT-4 güncelleme aldıkça bazı alanlarda iyileşme göstermiş, ancak bazı alanlarda da bilgisinin azaldığı veya eskisi kadar iyi cevap veremediği ortaya konmuştur. Chen vd. (2023), yaptıkları çalışmada ChatGPT'ye güncelleme yapılan zamanların belirli olmadığından bahsetmişlerdir. Bu doğrultuda 2023 Mart ve Haziran aylarında her iki modelde de aynı testleri uygulamışlardır. Onlar matematik problemleri, hassas sorular, kod yazma gibi farklı alanlarda deneyler yapmışlardır. Çalışmanın sonunda vardıkları husus ise zaman içerisinde her iki modelin de belirli alanlarda iyileşme gösterdiği, fakat bazı alanlarda da gerilediğidir. Onlara göre, Büyük Dil Modellerinin performansının eşit şekilde iyileştirilmesi zorludur. Modellerin belirli alanlardaki performansını iyileştirmek için yapılan fine-tuning işlemleri diğer alanlardaki davranış üzerinde beklenmedik yan etkiler ortaya çıkarabilmektedir.

Google da kendi yapay zekâ sohbet robotu Bard'ı 7 Şubat 2023'te kullanıma sunmuştur. ChatGPT'ye benzer şekilde sohbet sistemli çalışan Bard, kısa süre içerisinde 100 milyon kullanıcıya ulaşmıştır. LaMDA (Diyalog Uygulamaları için Dil Modeli) ile çalışan Bard, ChatGPT'ye rakip olarak ortaya çıkmıştır. ChatGPT'nin ücretsiz modeli olan GPT-3.5, 2021'e kadar sınırlı olan bilgisi ve web üzerinde arama yapmadan veri tabanındaki bilgileri kullanarak cevap üretirken, Bard anlık olarak web araması yapabilmektedir. Aynı zamanda ChatGPT'nin parametre sayısının Bard'a oranla fazla olmasından dolayı da ChatGPT çeşitli görevleri daha iyi yapabilmekte ve bağlamı koruyarak cevaplar verebilmektedir. Önemli olan bir diğer faktör de Bard'ın ücretsiz olarak kullanılabilmesi, ancak GPT-4'e ancak ücretli üyelik ile erişilebilmesidir (Ram & Pratima Verma, 2023, ss. 260-261).

Lim vd. (2023), yapay zekâ sohbet robotlarının tıp alanındaki bilgisini test etmek üzere internette sıkça aranan bir konu olan miyopluk ile ilgili ChatGPT-3.5, ChatGPT-4 ve Bard'ın verdiği cevapları karşılaştırmışlardır. Buldukları sonuçlara göre ChatGPT-4 çıktılarının “iyi” olarak değerlendirilme oranı %80,6 iken, ChatGPT-3.5 için bu oran %61,3 ve Google Bard için ise %54,8'dir.

Tıp alanında bir başka çalışma da Teebagy vd. (2023) tarafından yapılmıştır. Göz Hekimliği Bilgisi Değerlendirme Programı'nın (OKAP) farklı alt kategoriler içeren sınavını hem ChatGPT-3.5 hem de ChatGPT-4'e uygulayarak elde ettikleri sonuçlara göre GPT-4, “%81'e %57; $p < 0.001$ ” oranla ChatGPT-3.5'den daha üstün gelmiştir. Bunun yanı sıra ChatGPT-4, tüm alt kategorilerde 3.5'den daha iyi performans göstermiş, sadece bir kategoride ikisinin de sonuçları eşit çıkmıştır.

ChatGPT'nin belirli vakalardaki intihar girişimi ihtimallerini değerlendiren Levkovich ve Elyosef (2023), ChatGPT-4'ün intihar girişimi ihtimallerini ruh sağlığı hekimlerinin değerlendirmeleriyle kıyasladıklarında sonuçların benzer olduğunu ortaya koymuşlardır. Aynı çalışmada ChatGPT-3.5, bu ihtimalleri hekimlerin değerlendirmelerine kıyasla göz ardı etmiştir. Bulgulara göre, intihar etme düşüncesini tespit etme konusunda ChatGPT-4 daha iyi çalışmaktadır. Buna rağmen, “psychache” olarak adlandırılan psikolojik ağrıyı tanıma konusunda ChatGPT-4'ün abartılı tahminlerde bulunduğu gözlenmiş ve bu konuda daha fazla çalışma yapılması gerektiği belirtilmiştir.

ChatGPT-4'ün birbirinden farklı alanlarda gösterdiği başarılarından birisi de hukuktur. ChatGPT için simüle edilen bir baro giriş sınavında ChatGPT-4 sınava giren kişiler arasında puan olarak ilk %10'luk dilime girerken, ChatGPT-3.5 son %10'luk dilimde kalmıştır (Ray, 2023, s. 124).

Görüldüğü gibi ChatGPT-3.5 ve ChatGPT-4 karşılaştırması konulu çeşitli alanlara yönelik birçok araştırma yapılmıştır. Araştırmaların büyük bir kısmı doğruluk oranının yüksekliğini konu ederken, burada sözü edilmeyen bir kısım araştırmalar ise onların insanlara göre yanlılık durumunu vb. problem olarak ele almışlardır.

EŞ ADLİLİK VE BELİRSİZLİKLER

“Eş sesli (*sesteş*)” ve “eş adlı” kavramlarını tanımlama konusunda kesin bir fikir birliği bulunmamaktadır. Türk Dil Kurumu (TDK) Sözlüğüne göre “*sesteş*” kelimesi, “*söylenişleri aynı, anlam ve kökleri ayrı olan (kelimeler); eş sesli, homonim*” olarak tanımlanmaktadır (*Türk Dil Kurumu Sözlükleri*, 2022).

Yaygın olarak bilinen *sesteş* (eş sesli) terimi her ne kadar Türkçede belirli bir ölçüde söylenişleri aynı, anlamları farklı kelimeleri ifade etse de İngilizcede ve kimi dillerde eş seslilik kavramı, bir bakıma eş adlılığın alt başlığında bulunmaktadır. Eş adlılık kavramında önemli olan biçimin aynı, anlamın farklı

olmasıdır (Erkinay Tamtamış, 2021, s. 622). Bir diğer yazara göre de eş adlılık yerine kapsayıcı bir diğer terim eş gösterenliliktir. Dil, gösteren olduğu için somut bir niteliktedir ve eş gösterenlilik eş seslilik, eş adlılık, eş yazımlılık terimlerinin hepsini kapsayan bir üst başlık olmalıdır (Atmaca, 2013, s. 119).

Eş adlılık veya eş gösterenlilik altındaki kavramlar şu şekilde tanımlanmaktadır: Eş adlılık (*homonymy*), genel olarak yazılışı ve söylenişi aynı fakat anlamları farklı olan sözcükleri ifade etmektedir. İngilizce *bat (yarasa) – bat (sopa)*, Türkçe *ocak (eşya) – ocak (ay)* buna örnek olarak verilebilir. Eşyazımlılık (*homograph*), yazılışı aynı fakat söylenişi farklı olan sözcükleri tanımlamaktadır. İngilizce *tear (gözyaşı) – tear (yırtmak)*, Türkçe *seri (hızlı) – seri (dizi)* buna örnek olarak verilebilir. Burada her ne kadar biçimler aynı olsa da söylenişlerde vurgu veya ses farkı vardır. Eş seslilik (*homophone*) ise yazılışı farklı lakin söylenişi aynı olan kelimelerdir. İngilizce *brake (fren) – break (kırmak)*, Türkçe *grup (ekip) – gurup (güneşin batması)* da buna örnek olarak verilebilir (Hazar & Tarhan, 2013, s. 65). Bu durumda *ocak – ocak* örneğinde olduğu gibi eş adlı olan bir kelime hem eş yazımlı hem de eş seslidir. Fakat eş yazımlı olan bir kelime, eş sesli olmayabilir veya tam tersi durumda eş sesli olan bir kelime eş yazımlı olmayabilir. Her bir kelime biriminde bu durumlar dikkatle kontrol edilmelidir (Atmaca, 2013, s. 119).

Önemli olan bir diğer konu da yazılışları neredeyse aynı olan lakin küçük farklılıklarla birbirinden ayrılan ve aynı zamanda birbirlerinden farklı anlamlara gelen kelimelerin eş adlı kelime olarak kabul edilmez oluşudur. Bu hususta dikkat edilecek olan kural, seslerin üstünde bulunan düzeltme işaretidir (^). Örneğin, *mani (ruh hastalığı) – mâni (engel); hala (babanın kız kardeşi) – hâlâ (henüz)* kelimeleri düzeltme işaretiyle birbirinden farklılaşmaktadır.

Eş sesli kelimeler, konuşmayı metne çeviren sistemlerde önemli bir unsurdur. Neticede bu sistemler konuşulan metni yazıya dökmekte ve söylenişi aynı olan kelimeler arasında ayırım yapmakta zorlanmaktadır. Bu durumda ChatGPT gibi metin istemiyle çalışan bir sistemde yalnızca eş sesli (*homophone*) kelimeleri test etmek manasız olacaktır. Bunun yerine eş adlı, yani hem eş sesli hem de eş yazımlı kelimeler test edilmiştir.

Sözcükler arası anlam ilişkilerinden kaynaklı Doğal Dil İşleme sistemlerinde *belirsizlik (ambiguity)* durumu ortaya çıkmaktadır. Belirsizlikler bir sözcüğün, söz diziminin veya cümlenin birden fazla manaya gelebileceği durumlarda görülmektedir. Belirsizlikler birkaç farklı türde gerçekleşir ve bu belirsizlikleri giderme konusunda farklı yöntemler uygulanır. Bunlar dil bilimsel belirsizlik, sözcüksel belirsizlik, anlamsal belirsizlik ve söz dizimsel belirsizlik olmak üzere üç veya daha fazla şekilde gerçekleşebilir (Eyecioğlu Özmutlu, 2021, ss. 133-134).

Sözcüksel belirsizlik (*lexical ambiguity*), bir kelimenin birden çok anlama geldiği çok anlamlılık (*polysemy*) durumunda veya kelimenin eş adlı olması durumunda ortaya çıkar (Ortega-Martín vd., 2023, s. 2). Örneğin, “hayır” kelimesi tek başına sözcüksel belirsizlik oluşturmaktadır. “Hayır kurumuna hayır dedi.” ifadesi, cümle içerisinde eş adlı kelimenin farklı anlamlarda iki kere kullanılması nedeniyle sözcüksel belirsizlik oluşturmaktadır. Bazı yazarlar belirsizliğin, kelimenin biçiminde veya gramatik niteliğinde bulunmayıp anlamında bulunduğunu, bu durumun aynı zamanda anlamsal belirsizlik de oluşturduğunu vurgulamaktadır (Harsha vd., 2022, s. 1553).

Anlamsal belirsizlik (*semantic ambiguity*), cümle içerisindeki bağlam eksikliğinin birden fazla manaya yol açması durumunda gerçekleşir (Ortega-Martín vd., 2023, s. 2). Örneğin, “adam, evde otururken ayvayı yediğini fark etti.” cümlesi kişinin gerçekten bir meyve olarak ayva yediğini veya argo anlamıyla kötü bir duruma düştüğünü ve işinin bozulduğunu belirtebilir.

Söz dizimsel belirsizlik (*syntactic ambiguity*), cümle içerisindeki belirli bir söz grubunun birden fazla manaya gelecek şekilde cümle yapısında kullanılmasından dolayı oluşmaktadır (Ortega-Martín vd., 2023, s. 2). Örneğin, “adam kızı teleskopla gördü.” cümlesinden adamın kızı teleskopla görmüş olabileceği veya adamın teleskop taşıyan bir kızı görmüş olabileceği şeklinde iki farklı yorum çıkmaktadır (Eyecioğlu Özmutlu, 2021, s. 134). Bu durumda, cümle öğelerinin kullanıldığı yere bağlı olarak bir belirsizlik oluşmaktadır.

Eş adlı kelimelerin her biri sözcüksel belirsizlik (*lexical ambiguity*) oluşturmaktadır. Sözcüksel belirsizliği gidermenin yollarından birisi de cümle içerisinde belli bir bağlamla kullanılmasıdır. Aksi takdirde eş adlı bir kelimenin hangi anlamda olduğu tam olarak bilinemez. İnsanların dil bilgisi ve genel kültür becerileri, sözcüksel belirsizliği anlamada ve bunları gidermede yardımcı olmaktadır. Sözcüksel belirsizlik, özellikle belirli durumlarda kelime oyunları ve şakaları yapmak için bilerek kullanılmaktadır (Harsha vd., 2022, s. 1553).

Doğal Dil İşleme sistemlerinde sözcüksel belirsizlik “Word-Sense Disambiguation” (*WSD*) adı verilen ve “Kelime Anlam Belirsizliği Giderme” manasına gelen uygulama tarafından algılanır. Kelime Anlam Belirsizliği Giderme uygulamasının temel amacı, çok anlamlı veya eş adlı kelimeler sebebiyle cümlenin anlamının Doğal Dil İşleme sistemleri tarafından yanlış yorumlanmasını engellemektir (Harsha vd., 2022, s. 1554).

Kelime Anlam Belirsizliği Giderme uygulamalarında farklı algoritmalarla çözümler üretilmeye çalışılmıştır. Bu türler bilgi tabanlı, derlem tabanlı ve hibrit yöntemler olarak sınıflandırılabilir (Çetiner vd., 2021, s. 1). Doğal Dil İşleme sistemleri tarafından sözcüksel belirsizliğin ne kadar başarılı bir şekilde tespit

edilebildiğine dair çalışmalar bulunmaktadır. Ne var ki, bu çalışmaların birçoğu yüksek kaynaklı bir dil olan İngilizcede yapılmıştır. Türkçede yapılan çalışmalar oldukça kısıtlıdır.

Ortega-Martín vd. (2023), yaptıkları deneysel çalışmada ChatGPT'nin dil bilimsel belirsizlikleri tespit etme konusundaki güçlü ve zayıf yanları belirtmişlerdir. Ayrıca, çalışmayı yüksek kaynaklı olan İngilizce dili özelinde yapmayı bir sınırlandırma olarak değerlendirip farklı bir yüksek veya düşük kaynaklı dil kullanılması durumunda sonuçların değişebileceğini vurgulamışlardır.

Liu ve Fang (2023), eş adlı sözcükleri bulunan özel isimlerle yaptıkları çalışmada çıkardıkları sonuçta, özel olarak bunun için eğitilmiş bir Büyük Dil Modeline kıyasla ChatGPT'nin henüz bu ayrımı başarılı bir şekilde yapma konusunda sınırlı olduğunu belirtmişlerdir. Bunun sebepleri arasında genel amaçlı kullanılan Büyük Dil Modelinin alanla ilgili sınırlı bilgisini göstermişlerdir. Ayrıca Büyük Dil Modelinin belirli bir alanla ilgili muhakeme yeteneğinin, özellikle eğitilmiş bir modele kıyasla daha az olmasından dolayı bu modellerin ilgili ayrımları tam olarak yakalayamayacağını açıklamışlardır. Yine de ChatGPT'nin ve Büyük Dil Modellerinin metinsel muhakeme gerektiren istemler için veya etiketli verilerin az olduğu durumlarda bilgi grafiği oluşturmaya yardımcı olmada gelecek vaat eden bir yönünün olduğunu vurgulamışlardır.

ARAŞTIRMA

Araştırmanın Amacı ve Önemi

Araştırmanın amacı, ChatGPT-3.5 (ücretsiz) ve ChatGPT-4 (ücretli) Büyük Dil Modellerinin Türkçedeki eş adlı kelimelerin anlamlarını ayırt etmedeki başarılarını ayrı ayrı ölçmek ve bu yolla iki model arasındaki performans farkını ortaya koymaktır. Çalışma eş adlı kelimeler üzerinden bu farkı ortaya koymaya öncülük etmekte ve ChatGPT-4'ün gelecek vaat eden potansiyelinin altını çizmektedir.

Araştırma Verilerinin Hazırlanması ve Örneklem

Eş adlı kelimelerin tespiti ile ilgili internet üzerinden yapılan araştırmada bu kelimelerin yayımlandığı pek çok web sitesine rastlanmıştır (<https://www.dilbilgisi.net/>, <https://turkcealesi.net/>, <https://www.turkdilbilgisi.com/> gibi). Ayrıca bu kelimeleri içeren Yuva Yayınları tarafından yayımlanan “Resimli, Eş Anlamlı, Eş Sesli, Zıt Anlamlı Kelimeler Sözlüğü” (Eren, 2014), Parıltı Yayınları tarafından yayımlanan “Eşanlamlı Zıt Anlamlı Eşsesli Kelimeler Sözlüğü” (Kolektif, 2017) ve Ata Yayınları tarafından yayımlanan “Eşanlamlı Zıt Anlamlı ve Eş Sesli Kelimeler Sözlüğü” (Komisyon, 2023) olmak üzere üç kitap da bulunmaktadır. Başka bir çalışmada ise derli toplu, bilimsel bir sesteş (bu çalışmadaki karşılığı eş adlı) kelimeler sözlüğünün henüz mevcut olmadığı belirtilmektedir (Doğan, 2020, s. 229). Bu doğrultuda, araştırmada ele alınacak eş adlı kelimeler bahsedilen tüm kaynakların içinde en çok tekrar edilmiş

olanlardan seçilmiştir. Daha sonra bu eş adlı kelimelerin kaynaklarda verilen anlamları TDK Sözlüğündeki anlamlar ile karşılaştırılmıştır. Eş adlı kelimelerin kaynaklarda verilen bazı anlamlarının sözlükte geçmediği tespit edildiğinden bu kelimeler araştırmaya dâhil edilmemiştir. Bu yolla 200 eş adlı kelimededen oluşan bir veri tabanı elde edilmiştir. Dolayısıyla örneklem, olasılıklı olmayan örneklem türlerinden biri olan amaçlı (kasıtlı) örneklemidir.

ChatGPT'ye doğru istem girmek çıktının başarısı açısından önemli bir görevdir. Liu ve Fang'a göre (2023), iyi yapılandırılmış istemler oluşturmak GPT-3 gibi modellerin performansını önemli ölçüde artırarak daha doğru ve tutarlı çıktılar üretmelerini sağlayabilir. İstem ifadesinin, bağlamının ve biçiminin dikkatlice değerlendirilmesi, modelin anlayışını ve çıktı üretimini büyük ölçüde etkileyerek daha iyi bir görev performansı elde edilmesini sağlayabilir. OpenAI ise ChatGPT çıktılarında daha iyi sonuçlar elde etmek için gereken altı stratejiyi şu şekilde sıralamaktadır: *1-daha alakalı yanıtlar almak için sorgunuza ayrıntıları ekleyin, 2-modelden bir kişiliği benimsemesini isteyin, 3-girişin farklı bölümlerini açıkça belirtmek için sınırlayıcıları kullanın, 4-bir görevi tamamlamak için gereken adımları belirtin, 5-örnekler sağlayın, 6-çıktının istenen uzunluğunu belirtin (OpenAI, 2023).*

ChatGPT'nin başarı ölçümü için kullanılan istem kalıbı, farklı alternatifler denendikten sonra aşağıdaki biçimde tespit edilmiş ve tüm eş adlı kelimeler için yapılandırılmıştır:

“[cümle]” cümlesindeki “[kelime]” kelimeleri hangi anlamlara geliyor?

Örnek olarak, hayır kelimesine ilişkin istem:

“hayır kurumuna hayır dedi.” cümlesindeki “hayır” kelimeleri hangi anlamlara geliyor? biçimindedir.

ChatGPT-3.5'in bu istem için verdiği çıktı ise aşağıdaki gibidir:

“hayır kurumuna hayır dedi.” cümlesindeki “hayır” kelimeleri şu anlamlara gelebilir:

1. İlk “hayır”: Bu “hayır,” olumsuz bir yanıtı ifade eder. Yani kişi bu belirli bir kuruma veya teklife “hayır” demiş, kabul etmemiş veya reddetmiş demektir.

2. İkinci “hayır”: Bu “hayır,” muhtemelen bir hayır işi yapan kuruluşu veya yardım amaçlı bir kurumu ifade eder. Yani kişi, bu tür bir kuruma yapılan bağışı veya yardımı reddettiği anlamını taşır.

Cümlenin anlamı, bağlama ve kullanılan dile bağlı olarak değişebilir; ancak genel olarak “hayır” kelimesi olumsuz bir yanıtı veya bir kuruma yapılan bağışın reddedildiğini ifade etmek için kullanılır.

“hayır” eş adlı kelimesinin TDK Sözlüğündeki anlamları ise aşağıdaki şekildedir:

- “Yok, öyle değil, olmaz” anlamlarında onamama, inkâr bildiren bir söz.
- İyilik, karşılık beklenmeden yapılan yardım.

Görüldüğü gibi “hayır” kelimesine ilişkin istemin ChatGPT-3.5 çıktısı, onun iki farklı anlamını -her ne kadar ters sırada belirtmiş olsa da- doğru olarak değerlendirdiğini göstermektedir.

Bu yolla yazarlar 200 eş adlı kelimenin her biri için ayrı olmak üzere, eş adlı kelimenin her iki anlamını da aynı cümlede kullanan kasıtlı cümleler oluşturmuşlardır. Oluşturulan cümleler bağlam, biçim, anlam gibi özellikler bakımından uzman bir dil bilimci tarafından kontrol edilmiş ve eksiklikler giderildikten sonra istem ekranına taşınmıştır (örnek cümleler: “*mahalledeki tavşan ustasından bana tavşan motifli bir tepsi yapmasını istedim.*”, “*dedem ona aldığı mestleri görünce mest oldu.*”, “*papağanıma verdiğim katı yiyeceklerin onun katısına dokunmaması için dikkat etmeliyim.*”, “*ülkenin emiri vergilerin 3 akçe olacağına dair yeni emir yayınladı.*”, “*asmalardan toplanan üzümleri koruk yapmak için dama asma işini bana verdiler.*”).

Daha sonra bu cümleler ChatGPT-3.5 ve 4.0’e ayrı ayrı istem olarak girilmiş ve çıktılarda yer alan anlamlar doğru bilinenler için “1” ve yanlış bilinenler için “0” olmak üzere veri tabanına işlenmiştir. Bu değerler veri tabanının "ChatGPT-3.5'in 1. anlam başarısı", "ChatGPT-3.5'in 2. anlam başarısı", "ChatGPT-4'ün 1. anlam başarısı" ve "ChatGPT-4'ün 2. anlam başarısı" başlıklı sütunlarında yer almışlardır.

Bu noktada şu hususları da belirtmek gerekir:

Kasıtlı cümlelerde eş adlı kelimeler ek de alabilmişlerdir. Örnek olarak, “*hayır kurumuna hayır dedi.*” istemindeki hayır kelimelerinin eki yokken, “*ocak tanıtımı ocakta yapılacak.*” istemindeki ocak kelimelerinin birinde -de/da eki bulunmaktadır. Bazı cümlelerde ise her iki eş adlı kelime de ek alabilmiştir. Diğer yandan, kullanılan tüm ekler çekim ekidir (isim/fiil çekim ekleri). Zira yapım ekleri eklendiği kelimenin hem anlamını hem de türünü değiştirmektedir.

ChatGPT’ye istem olarak girilen 200 cümledeki bazı eş adlı kelimeler, cümle içindeki kullanımda ses düşmesine uğrayabilmiştir. Örnek olarak, “*rahman ve rahim olan Allah ana rahmindeki tüm bebeklerin koruyucusudur.*” istemindeki “rahim” eş adlı kelimesi “rahmindeki” biçiminde kullanılmış ve böylelikle ses düşmesine uğramıştır.

ChatGPT’ye istem olarak girilen 200 cümledeki eş adlı kelimeler, cümle içindeki kullanımda ünsüz yumuşamasına uğrayabilmiştir. Örnek olarak, “*dededen kalan bozuğu çalmak istedim, ancak akordu bozuktu.*” istemindeki “bozuk” eş adlı kelimesi “bozuğu” biçiminde kullanılmış ve böylelikle ünsüz yumuşamasına uğramıştır.

Tablo 1’de veri tabanından örnek bir kesit görülmektedir.

Tablo 1.

Veri Tabanından Örnek Bir Kesit

Eş adlı kelime	ChatGPT-3.5'in 1. anlam başarısı	ChatGPT-3.5'in 2. anlam başarısı	ChatGPT-4'ün 1. anlam başarısı	ChatGPT-4'ün 2. anlam başarısı
açık	1	1	1	1
açma	1	1	1	1
ağ	1	1	1	1
ağız	0	1	1	1
ahu	1	0	1	1
akar	0	1	1	1
akide	1	1	1	1
akrep	1	1	1	1
alay	1	1	1	1
arı	1	1	1	1

Araştırmanın Sınırlılıkları

ChatGPT'ye ilişkin sınırlılıklar

- İstem çıktıları ChatGPT-3.5 için 16.10.2023-17.12.2023 tarihleri arasında istenmiştir. ChatGPT-4 için ise bu tarihler 09.01.2024-15.01.2024 aralığındadır. Dolayısı ile her iki model için bu zaman aralıklarındaki iyileştirmeler araştırmanın sınırlılığı kapsamında değerlendirilebilir.
- Bir diğer sınırlılık ise ChatGPT-3.5'in bazen ayrıntılı (doyurucu) bir çıktı üretmesi, ancak kimi zaman da aşağıda görüldüğü gibi çok sınırlı bir çıktı üretmesi ile ilgili durumlardan oluşmaktadır. Örneğin, burada "bere" eş adlı kelimesinin 2. anlamı hiç açıklanmamıştır.

İstem:

"bere ve eldiven takmış olmama rağmen ormanda ağaç toplarken hem üşüdüm hem de ellerim yara bere içinde kaldı." cümlesindeki "bere" kelimeleri hangi anlamlara geliyor?

Çıktı:

"bu cümlede geçen "bere" kelimesi, başı örtmek ve ısıtmak amacıyla kullanılan, genellikle kış aylarında giyilen bir tür başlık anlamına gelmektedir."

İstem cümlelerine ilişkin sınırlılıklar

- İstem cümleleri "*sedir ağacından kendimize bir sedir yaptık.*" örneğinde olduğu gibi yüklemının türüne göre eylem cümlesi olabildiği gibi, "*bayramlarda çocuklara akide şekeri dağıtmak bizim dini akidelerimizdendir.*" örneğinde olduğu gibi isim cümlesi de olabilmektedir.
- İstem cümlelerinin hepsi "*traktörle ormandan odun getirmeye çalışırken*

yolda biriken mile saplanınca direksiyon milim kırıldı.” örneğinde olduğu gibi yüklemimin yerine göre kurallı cümledir. Devrik cümle veya eksilteli cümle kullanılmamıştır.

- İstem cümleleri *“heybetiyle meydana giren şah, atını şaha kaldırarak halkı selamladı.”* örneğinde olduğu gibi anlamına göre olumlu cümle olabildiği gibi, *“sağcı çatışmacılardan sağ kalan olmadı.”* örneğinde olduğu gibi olumsuz cümle de olabilmektedir. Ayrıca istem cümleleri *“hayatım farı biraz abartmışsın, benim arabanın farı o kadar parlamıyor.”* örneğinde olduğu gibi ünlem cümlesi de olabilmektedir. Soru cümlesi kullanılmamıştır.
- İstem cümleleri *“bahçedeki setin arkasındaki düzlükte birkaç set maç yaptık.”* örneğinde olduğu gibi yapısına göre basit cümle olabildiği gibi, *“biz piknikte mangalda palamut pişirirken sincaplar da palamut topluyordu.”* örneğinde olduğu gibi birleşik cümle de olabilmektedir. Ayrıca *“arkadaşım çok tetik bir çocuktü, tetiğe herkesten önce basardı.”* örneğinde olduğu gibi sıralı cümle veya *“kaşkas boyundan gelenler genellikle renkli gözlü ve orta boyda olurlar.”* örneğinde olduğu gibi bağılı cümle de olabilmektedir.
- İstem cümleleri *“güreşte rakibim kulağıma pes bir tonla pes dedi.”* örneğinde olduğu gibi anlatımına göre özne anlatımlı cümle olabildiği gibi, *“üfleme sazlarından bazılarında saz kullanılır.”* örneğinde olduğu gibi nesnel anlatımlı cümle de olabilmektedir. Bu konuda bir ayrıma gidilmemiştir.

Eş adlı kelimelere ilişkin sınırlılıklar

- Veri tabanında yer alan bazı eş adlı kelimeler peşlerine aldıkları kelimelerle yeni bir kelime haline de gelebilmektedir. Örneğin, “ser” kelimesi için serasker ve “dar” kelimesi için de darağacı biçimindeki kelimeler TDK Sözlüğünde yer almaktadır. Böyle kelimelere istem cümlelerinde yer verilmemiştir.
- Veri tabanında yer alan eş adlı kelimelerin kullanılan bütün anlamları TDK Sözlüğünde düzeltme işareti olmadan tanımlanmaktadır. Ancak “mani” eş adlı kelimesi bu durumun bir istisnadır. Yine de bu kelimenin kullanılan her iki anlamında da TDK Sözlüğüne göre düzeltme işareti bulunmaktadır. Bu durum çalışmanın literatür bölümünde tartışılan eş seslilik durumuna işaret etmekle birlikte, çalışmada eş adlılık üst başlığı kullanılmamıştır.
- Veri tabanında yer alan eş adlı kelimelerin bazıları önelerine aldıkları kelimelerle TDK Sözlüğüne göre birleşik kelime oluşturabilmektedir. Örneğin, “bez” eş adlı kelimesi cümlede “sargı bezi”, “bere” eş adlı kelimesi de “yara bere” biçiminde birleşik kelime yapısında kullanılabilmektedir (birleşik kelime yapısında kelimenin hem önüne hem de sonuna ikinci kelime gelebilir).
- Veri tabanında yer alan eş adlı kelimelerin bazıları cümle içinde kullanılırken TDK Sözlüğünün verdiği Atasözleri, Deyimler, Birleşik Fiiller veya Kalıp Sözler kapsamında kullanılabilmektedir. Örnek olarak, “öğretmenim bana cebiri

sevdirmek için cebir kullanmak zorunda kaldı.” cümlesinde “cebir” eş adlı kelimesi “cebir kullanmak” birleşik fiili biçiminde kullanılmıştır.

- Veri tabanında yer alan eş adlı kelimeler cümle içinde kullanılırken anlam bakımından farklı türlere ait olabilmektedir. Örnek olarak, “bere” eş adlı kelimesinin “yara bere” biçimindeki kullanımı veya “ser” eş adlı kelimesinin “ser verip sır vermemek” biçimindeki kullanımı deyim anlama, “cebir” ve “baş” eş adlı kelimeleri terim anlama (matematik, anatomi), “pusula” eş adlı kelimesi somut anlama, “akide” eş adlı kelimesinin inanç anlamındaki kullanımı soyut anlama vurgu yapar. Argo anlam ve mecaz anlama yer verilmemiştir.
- Veri tabanında yer alan eş adlı kelimeler cümle içinde kullanılırken özel isim olarak kullanılmamıştır. Örnek olarak, “han” eş adlı kelimesi “*karsta ziyaret ettiğim tarihi hanı 1200’lü yıllarda ziyaret edenler arasında önceki Türk devletlerinin hanlarının da bulunduğunu öğrendim.*” cümlesinde örneğin “Cengiz Han” biçiminde kullanılmamıştır.

Araştırmanın Yöntemi

Yazarlar tarafından oluşturulan kasıtlı cümlelerin farklı özellikler bakımından uzman bir dil bilimci tarafından kontrol edilmesi ve eksiklikler giderildikten sonra istem ekranına taşınması araştırmanın güvenilirliğine ilişkin bir ölçüttür. Ayrıca ChatGPT-3.5 ve ChatGPT-4’ün 1. ve 2. anlam başarılarının her iki yazar tarafından da aynı sonuçla veri tabanına işlenmesi “tekrarlı olduğunda da aynı sonuçlara ulaşmak” ilkesine göre yine güvenilirlik kapsamında değerlendirilmiştir. Geçerlilik ise burada ChatGPT-3.5 ve ChatGPT-4’ün 1. ve 2. anlam başarı oranlarına karşılık gelmektedir.

Araştırmada ChatGPT-4 tarafından önerilen yöntemlerle (başarı oranı dağılım analizi, eş adlı kelimeye göre performans değişikliği, eş adlı kelimenin karakter sayısı ve başarı oranı, istatistiksel testler) sağlanan çıktılar, yazarlar tarafından hesaplanan gerçek değerlerle karşılaştırılmıştır. Ayrıca başlangıçta öne sürülen hipotez ChatGPT-4’ün başarısının daha yüksek olacağı yönünde idi. Bu hipotez araştırma sonucunda doğrulanmış olup, onun amaç bakımından keşfedici olduğuna vurgu yapar.

Araştırmanın Bulguları ve Değerlendirme

İstem cümleleri ChatGPT-3.5 ve ChatGPT-4’e Türkçe dilinde, analiz istemleri ise ChatGPT-4’e İngilizce dilinde girilmiştir. ChatGPT-4 çıktıları aşağıda Türkçe diline çevrilerek verilmiştir (şekiller hariç). Bulgular ChatGPT-4 tarafından önerilen analizler -8 tane- doğrultusunda gerçekleştirilmiş olup amaç doğrultusunda 4 tanesi kullanılmıştır.

Başarı oranı dağılım analizi (Success rate distribution analysis)

İstem: *Can you visualize the success rates both for not considering the success rates of first and second words and the first one you did?*

Çıktı: Verilen çıktı Tablo 2'ye uyarlanmıştır.

Tablo 2.

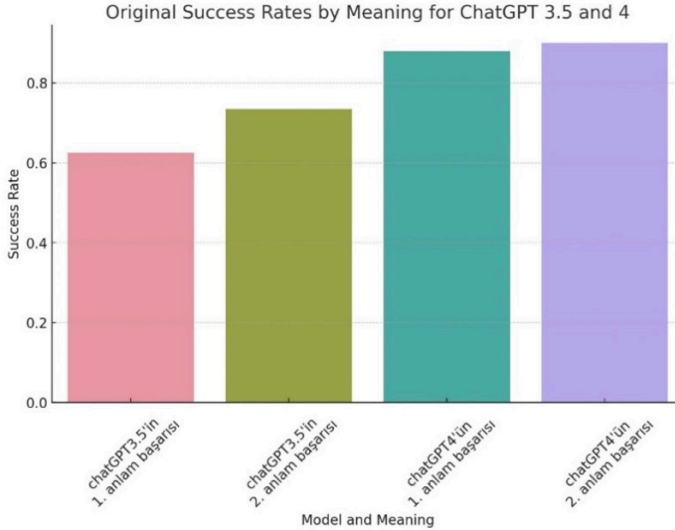
ChatGPT-3.5 ve ChatGPT-4'ün Anlam Başarı Oranları

ChatGPT-3.5 ve ChatGPT-4'ün Anlam Başarıları	Oranlar (%)
ChatGPT-3.5'in 1. anlam başarısı	62,5
ChatGPT-3.5'in 2. anlam başarısı	73,5
ChatGPT-4'ün 1. anlam başarısı	88,0
ChatGPT-4'ün 2. anlam başarısı	90,0

Tablo 2'de görüldüğü gibi ChatGPT-3.5'e göre daha gelişmiş bir Büyük Dil Modeli olan ChatGPT-4'ün her iki anlam başarı oranı da daha yüksektir. Ayrıca her ikisinde de eş adlı kelimenin cümledeki 2. kullanımları 1.'ye göre daha doğru olarak bilinmiştir. Bu durumun sebebi, istem cümlelerinin İngilizce dilinde olduğu gibi cümlenin sonundan başına doğru anlamlandırılmaya çalışma çabası olabilir. ChatGPT-4 tarafından verilen çıktı görselleri ise aşağıdaki gibidir.

Şekil 1.

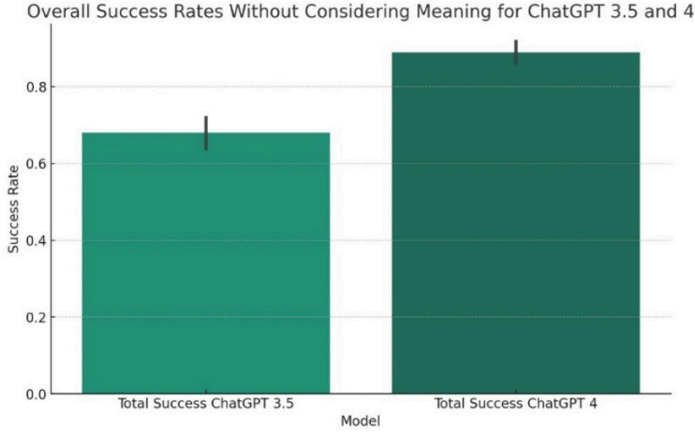
3.5 ve 4'ün Ayrı Ayrı Başarı Değerleri



(ChatGPT-4'ten alınan istem çıktısı)

Şekil 2.

3.5 ve 4'ün Toplam Başarı Oranları



(ChatGPT-4'ten alınan istem çıktısı)

Gerçek değer: Yazarlar tarafından da manuel değerlendirmeye ulaşılan sonuçlar Tablo 2, Şekil 1 ve Şekil 2'nin aynısıdır.

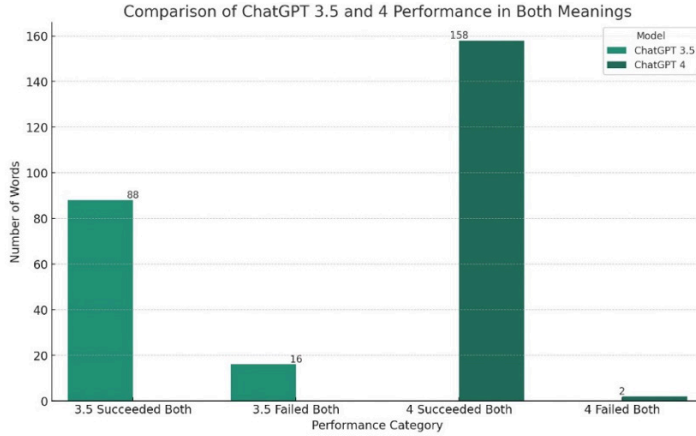
Eş adlı kelimeye göre performans değişikliği (Performance variation by word)

İstem: *Can you visualize where ChatGPT-3.5 succeeded in both meanings and where it failed in both meanings? Can you visualize where ChatGPT 4 succeeded in both meanings and where it failed in both meanings? Can you combine these two graphics into one?*

Çıktı: Verilen çıktı Şekil 3'te görülmektedir. Buna göre, ChatGPT-3.5 her iki anlamda da 200 eş adlı kelimedenden 88'inin anlamını doğru bir biçimde bilmiş, oysa her iki anlamda da 200 eş adlı kelimenin 16'sında anlamı bilmede başarısız olmuştur. ChatGPT-4 ise her iki anlamda da 200 eş adlı kelimedenden 158'inin anlamını doğru bir biçimde bilmiş, her iki anlamda da 200 eş adlı kelimenin sadece 2'sinde anlamı bilmede başarısız olmuştur. Görüldüğü gibi ChatGPT-4 ChatGPT-3.5'e göre her iki anlamı bilmede %80, bilememede ise %80 oranlarında daha başarılı olmuştur.

Şekil 3.

3.5 ve 4'ün Her İki Anlamı Bilme ve Bilememe Değerleri



(ChatGPT-4'ten alınan istem çıktısı)

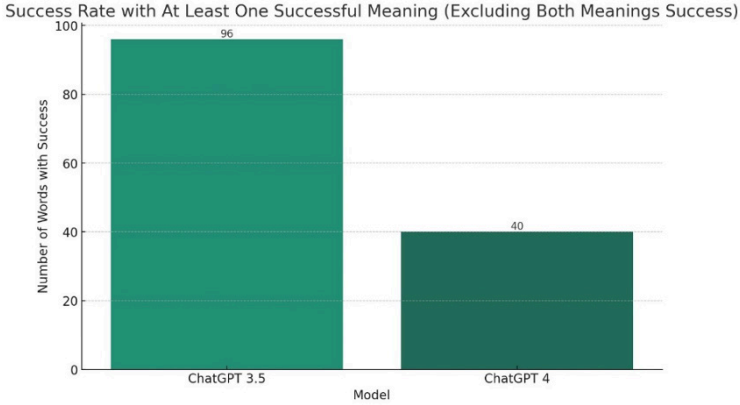
Gerçek değer: Yazarlar tarafından da manuel değerlendirmeye ulaşılan sonuçlar Şekil 3'ün aynıdır.

İstem: *Regardless of 1st meaning and 2nd meaning, can you visualize the success rate where they have at least 1? Can you exclude where they succeeded in both meanings?*

Çıktı: Verilen çıktı Şekil 4'te görülmektedir. Bu grafik, 1. anlam veya 2. anlam fark etmeksizin eş adlı kelimelerden en az birinin anlamını bilen ChatGPT-3.5 ve ChatGPT-4'e ait değerleri görselleştirmektedir. Bu sorgu bir önceki sorgunun tersi biçimindedir. Buna göre, her ne kadar ChatGPT-3.5'in değerleri ChatGPT-4'e göre daha yüksek gözükse de aslında bu durum yanıltıcıdır. Zira ChatGPT-4'ün her iki anlamı da bildiği değerler çok yüksek olduğu için buradaki herhangi bir anlamdan sadece bir tanesini bilmeye ilişkin değerler daha düşük kalmıştır.

Şekil 4.

3.5 ve 4'ün Her İki Anlamdan En Az Birini Bilme Değerleri



(ChatGPT-4'ten alınan istem çıktısı)

Gerçek değer: Yazarlar tarafından da manuel değerlendirmeye ulaşılan sonuçlar Şekil 4'ün aynısıdır.

Eş adlı kelimenin karakter sayısı ve başarı oranı (Word length and success rate)

İstem: *Can you make this Word Length vs Success Rate analysis for the cases where both meanings are 1? Can you give me the rates for this graph?*

Çıktı: Verilen çıktı Tablo 3'e uyarlanmıştır.

Tablo 3.

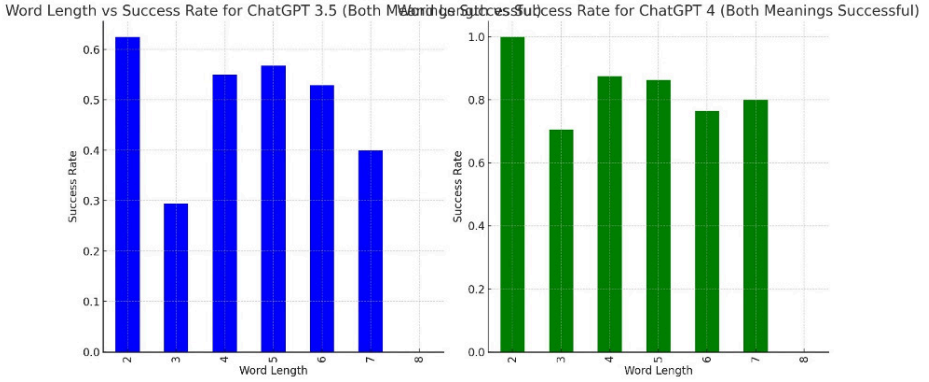
Eş Adlı Kelimenin Karakter Sayısına Göre ChatGPT-3.5 ve ChatGPT-4 Başarı Oranları

Eş Adlı Kelimenin Karakter Sayısı	ChatGPT-3.5 Başarı Oranları (%)	ChatGPT-4 Başarı Oranları (%)
2	62.5	100.0
3	29.4	70.6
4	55.0	87.5
5	56.8	86.4
6	52.9	76.5
7	40.0	80.0
8	0.0	0.0

ChatGPT-4'ün ChatGPT-3.5'e göre olan genel yüksek başarı oranı, eş adlı kelimelerin karakter sayılarında da belirgin bir biçimde kendini göstermektedir. Diğer yandan, her ikisinde de başarı oranları kelimelerin karakter sayısındaki artışa/azalışa göre düzgün bir dağılım göstermemektedir. Son olarak, eş adlı kelimenin karakter sayısı 8'e çıktığında hem ChatGPT-3.5 hem de ChatGPT-4 kelimenin her iki anlamını da bilememişlerdir. İlgili grafik Şekil 5'te görülmektedir.

Şekil 5.

3.5 ve 4'ün Eş Adlı Kelimenin Karakter Sayısına Göre Anlamı Bilme Değerleri



(ChatGPT-4'ten alınan istem çıktısı)

Gerçek değer: Yazarlar tarafından da manuel değerlendirmeye ulaşılan sonuçlar Şekil 5'in aynısıdır.

İstatistiksel Testler (Statistical Testing)

Bu noktada çıkarımsal istatistiksel testlerin yapılabilmesi için veri tabanına birinci olarak "eş adlı kelime türü 1" ve "eş adlı kelime türü 2" olmak üzere 2 bağımsız değişken eklenmiştir. Buna göre, ChatGPT'ye istem olarak girilen 200 cümledeki her bir eş adlı kelime, kelime türü bakımından incelenmiş ve veri tabanına kaydedilmiştir. Örnek olarak, "hayır kurumuna hayır dedi." istemindeki hayır kelimelerinin türleri sırasıyla isim ve ünlem biçimindedir. Kelime türü tespitinde TDK Sözlüğü esas alınmıştır.

Çıkarımsal testlerden ki-kare testi uygulanmıştır. Bu test aynı örnekteki kategorik değişkenleri karşılaştırmak için kullanılmaktadır.

ChatGPT-3.5 için ki-kare testi (kelime türü):

İstem: *Can you make chi-square test on this? First meaning and second meaning columns should be lines. Eş adlı kelime türü 1 and Eş adlı kelime türü 2 should be columns. I want crosstab.*

Çıktı: Tüm kategorilerin dahil edildiği ki-kare testinin sonuçları aşağıdaki gibidir:

- ki-kare istatistiği: 6.00
- p-değeri: 0.307

Şekil 6.

Eş Adlı Kelime Türlerine Göre Ki-Kare Testi (ChatGPT-3.5)

Success Rate	İsim	Sıfat	Zarf	Zamir	Ünlem	Fiil
0 (Failure)	112	10	1	0	2	3
1 (Success)	218	37	7	2	2	6

P değeri yaklaşık 0,307'dir ve bu istatistiksel anlamlılık için kullanılan tipik eşik olan 0,05'ten daha yüksektir. Bu, çeşitli kelime türleri arasındaki (hem birinci hem de ikinci anlamlar için) başarı oranlarındaki farklılıkların istatistiksel olarak anlamlı olmadığını göstermektedir. Daha basit bir ifadeyle, başarı oranlarının farklı kelime türleri arasında önemli ölçüde farklılık gösterdiğini gösteren güçlü bir kanıt yoktur.

Gerçek değer: Yazarlar tarafından ChatGPT-3.5 ve ChatGPT-4'ün 1. ve 2. anlam başarılarına ilave olarak eş adlı kelime türü 1 ve eş adlı kelime türü 2 sütunlarının eklendiği tablo SPSS programına aktarılmış ve ki-kare testi yapılmıştır. Yapılan analiz sonucunda p değerinin 0,05'den büyük olması nedeniyle anlamlı bir ilişki bulunamamıştır. Diğer bir ifade ile manuel değerlendirme sonucu ChatGPT-4 çıktısıyla aynı olmuştur.

ChatGPT-4 için ki-kare testi (kelime türü):

İstem: *Can you make chi-square test on this? First meaning and second meaning columns should be lines. Eş adlı kelime türü 1 and Eş adlı kelime türü 2 should be columns. I want crosstab.*

Çıktı: Tüm kategorilerin dahil edildiği ki-kare testinin sonuçları aşağıdaki gibidir:

- ki-kare istatistiği: 1.90

- p değeri: 0.863

Şekil 7.

Eş Adlı Kelime Türlerine Göre Ki-Kare Testi (ChatGPT-4)

Success Rate	Fiil	İsim	Sıfat	Zamir	Zarf	Ünlem
0 (Failure)	1	37	6	0	0	0
1 (Success)	8	293	41	2	8	4

P değeri yaklaşık 0,863 olup istatistiksel anlamlılık için kullanılan standart eşik olan 0,05'ten çok daha yüksektir. Bu, çeşitli kelime türleri arasındaki başarı oranlarındaki farklılıkların (hem birinci hem de ikinci anlamlar için) istatistiksel olarak anlamlı olmadığını göstermektedir. Daha basit bir ifadeyle, başarı oranlarının farklı kelime türleri arasında önemli ölçüde farklılık gösterdiğini gösteren güçlü bir kanıt yoktur.

Gerçek değer: Yazarlar tarafından ChatGPT-3.5 ve ChatGPT-4'ün 1. ve 2. anlam başarılarına ilave olarak eş adlı kelime türü 1 ve eş adlı kelime türü 2 sütunlarının eklendiği tablo SPSS programına aktarılmış ve ki-kare testi yapılmıştır. Yapılan analiz sonucunda p değerinin 0,05'den büyük olması nedeniyle anlamlı bir ilişki bulunamamıştır. Diğer bir ifade ile manuel değerlendirme sonucu ChatGPT-4 çıktısıyla aynı olmuştur.

Yine bu noktada çıkarımsal istatistiki testlerin yapılabilmesi için veri tabanına ikinci olarak "eş adlı kelimelerin ek durumu" olmak üzere 1 bağımsız değişken daha eklenmiştir (eş adlı kelime türleri yerine). ChatGPT'ye istem olarak girilen 200 cümledeki her bir eş adlı kelime ek alıp almama durumuna göre incelenmiş ve ekli/eksiz biçiminde veri tabanına kaydedilmiştir.

Çıkarımsal testlerden ki-kare testi uygulanmıştır. Bu test aynı örneklemdaki kategorik değişkenleri karşılaştırmak için kullanılmaktadır.

ChatGPT-3.5 için ki-kare testi (ek durumu):

İstem: *Can you make chi-square test on this? First meaning and second meaning columns should be lines. Eş adlı kelimelerin ek durumu should be column. Can you make the test separately for each meaning?*

Çıktı: Her anlam için ayrı ayrı yapılan ki-kare testlerinin sonuçları aşağıdaki gibidir:

1. anlam

- ki-kare istatistiği: 5.95

- p değeri: 0.015

2. anlam

- ki-kare istatistiği: 0.35

- p değeri: 0.554

Şekil 8.

Eş Adlı Kelimelerin Ek Alma Durumlarına Göre Ki-Kare Testi (ChatGPT-3.5)

1st Meaning:		
Success Rate	ek var	ek yok
0 (Failure)	62	13
1 (Success)	82	43

2nd Meaning:		
Success Rate	ek var	ek yok
0 (Failure)	36	17
1 (Success)	108	39

İlk anlam için p değeri yaklaşık 0,015'tir. Bu, istatistiksel anlamlılık testleri için kullanılan ortak alfa düzeyi olan 0,05'ten düşüktür. Bu durum, eş adlı kelimelerin ek durumunun birinci anlama ilişkin başarı oranları arasındaki dağılım farkının istatistiksel olarak anlamlı olduğunu göstermektedir. İkinci anlam için p değeri yaklaşık 0,554 olup 0,05'ten büyüktür. Bu da eş adlı kelimelerin ek durumunun ikinci anlam başarı oranları arasındaki dağılım farkının istatistiksel olarak anlamlı olmadığını göstermektedir.

Gerçek değer: Yazarlar tarafından ChatGPT-3.5 ve ChatGPT-4'ün 1. ve 2. anlam başarılarına ilave olarak eş adlı kelimelerin ek durumu sütununun eklendiği tablo SPSS programına aktarılmış ve ki-kare testi yapılmıştır. Yapılan 1. ve 2. anlam analiz sonuçlarına göre, 1. anlamda p değeri 0,05'den küçük çıkmıştır. Dolayısıyla bu ilişki anlamlıdır. Oysa 2. anlamdaki p değeri 0,05'ten büyüktür ve ilişki anlamsızdır. 1. anlamdaki ilişki şunu ifade etmektedir: ChatGPT-3.5'de kelimelerin ek almama durumunda başarı oranı ek alma durumuna göre 3,3 kattan daha yüksektir. Aynı durum ek alma durumu için de geçerlidir, ancak bu oran daha düşüktür (1,3 kat). Sonuç olarak, manuel değerlendirme sonucu ChatGPT-4 çıktısıyla aynı olmuştur.

ChatGPT-4 için ki-kare testi (ek durumu):

İstem: *Can you make chi-square test on this? First meaning and second meaning columns should be lines. Eş adlı kelimelerin ek durumu should be column. Can you make the test separately for each meaning?*

Çıktı: Her anlam için ayrı ayrı yapılan ki-kare testlerinin sonuçları aşağıdaki gibidir:

1. anlam
 - ki-kare istatistiği: 1.16
 - p değeri: 0.282
2. anlam
 - ki-kare istatistiği: 1.22
 - p değeri: 0.270

Şekil 9.*Eş Adlı Kelimelerin Ek Alma Durumlarına Göre Ki-Kare Testi (ChatGPT-4)*

1st Meaning:		
Success Rate	ek var	ek yok
0 (Failure)	20	4
1 (Success)	124	52

2nd Meaning:		
Success Rate	ek var	ek yok
0 (Failure)	17	3
1 (Success)	127	53

Birinci ve ikinci anlamlar için p değerleri sırasıyla 0,282 ve 0,270'tir. Bunların her ikisi de istatistiksel anlamlılık için kullanılan 0,05'lik standart eşikten daha yüksektir. Bu durum, eş adlı kelimelerin ek durumunun hem birinci hem de ikinci anlamlara ilişkin başarı oranları arasındaki dağılım farklılıklarını istatistiksel olarak anlamlı olmadığını göstermektedir. Daha basit bir ifadeyle, başarı oranlarının her iki anlam için de kelimelerin ek alma durumuna göre önemli ölçüde değiştiğini gösteren güçlü bir kanıt yoktur.

Gerçek değer: Yazarlar tarafından ChatGPT-3.5 ve ChatGPT-4'ün 1. ve 2. anlam başarılarına ilave olarak eş adlı kelimelerin ek durumu sütununun eklendiği tablo SPSS programına aktarılmış ve ki-kare testi yapılmıştır. Yapılan 1. ve 2. anlam analiz sonuçlarına göre, her iki anlamda da p değeri 0,05'den büyük çıkmıştır. Dolayısıyla ilişkiler anlamsızdır. Sonuç olarak, manuel değerlendirme sonucu ChatGPT-4 çıktısıyla aynı olmuştur.

Bu noktada şu hususu vurgulamak gerekir ki, kelime türleri ve ek alma durumlarına göre analiz yapılmasının sebebi, bu durumların bağımsız bir değişken olamayacağını göstermekti. Zira ChatGPT-4 eğitiminin arka planında bir sonraki öğeyi tahmin etmeye dayalı bir model bulunmaktadır.

SONUÇ

Bu çalışmada ChatGPT-3.5 ve ChatGPT-4 Büyük Dil Modellerinin performansları Türkçedeki eş adlı kelimeler üzerinden incelenmiştir. ChatGPT-4'ün eşadlı kelimelerin anlamlarını bilme başarısı ortalamada ChatGPT-3.5'e göre 1,3 kat daha fazla oranda gerçekleşmiştir. Daha fazla parametreye ve veri setine sahip olan ChatGPT-4 için bu sonuç "beklenen" değerdir.

ChatGPT-4 ayrıca verilen tabloya göre yapılabilecek analizler konusunda

önerilerde de bulunmuştur. Bunlardan birisi eş adlı kelimenin karakter sayısına göre her iki Büyük Dil Modelinin performanslarının karşılaştırılması olmuştur. Buna göre, 8 karakter uzunluğundaki eş adlı kelimenin her iki anlamını her iki dil modelinin de bilememiş olmasına karşın, ChatGPT-4 diğer karakter sayısına sahip tüm eş adlı kelimelerde ChatGPT-3.5' e göre üstün performans göstermiştir. Bu oranlar 1,44 ile 2,4 arasında değişmektedir.

ChatGPT-4 analiz önerilerinden bir diğeri de verilen tablo üzerinde yapılabilecek istatistiksel testlerdir. Bunlardan ki-kare testi uygulanmıştır. İlk ki-kare testi eş adlı kelime türlerine göre yapılan testlerdir. Burada hem ChatGPT-3.5 hem de ChatGPT-4'de kelime türleri arasındaki başarı oranlarındaki farklılıklar istatistiksel olarak anlamlı değildir. Bir diğer deyişle, kelime türünün isim, sıfat, zarf, ünlem vb. olması başarı oranını etkileyen değişkenler değildir. Manuel değerlendirmeye de bu durum ispatlanmıştır.

İkinci ki-kare testi eş adlı kelimenin ek alma durumuna göre yapılan testlerdir. Burada ChatGPT-3.5 için yapılan test anlamlı sonuçlara ulaşırken, ChatGPT-4 için bu durum söz konusu olmamıştır. Başka bir deyişle, bir kelime ek aldığımda ChatGPT-3.5'de buna göre başarı oranı farklılık göstermekte, ancak ChatGPT-4'de göstermemektedir. Yine manuel değerlendirmeye bu durum da ispatlanmıştır. Diğer yandan, bu farklılık Büyük Dil modellerinin değişen performansından kaynaklanmıştır. Parametre setinin genişliği, daha yeni eğitim verileri, araştırma geribildirimleri bu performans değişikliğinin nedenleri arasında sayılabilir. Ayrıca burada ChatGPT-3.5 için elde edilen anlamlı sonuç olumsuz bir performans olarak değerlendirilmektedir. Zira ek alma durumuna göre bağımlılık, Büyük Dil Modellerinin “bir sonraki veriye bakarak çıktı üretme” prensibi ile örtüşmemektedir.

Elde edilen sonuçlar, görüldüğü gibi ChatGPT-4'ün eş adlı kelimelerin farklı anlamları ile ilgili istemlere doğru çıktı üretme potansiyelini vurgulamakta ve buna ilişkin değerli içgörüler sunmaktadır. Ayrıca bu karşılaştırmalı analiz, farklı Büyük Dil Modelleri arasındaki doğruluğun detaylı bir şekilde anlaşılmasını sağlamaktadır. Dahası ChatGPT-4 tarafından önerilen analizleri de ortaya koymakta ve bunların doğruluğunu manuel değerlendirme ile hesaplanan gerçek değerlerle karşılaştırarak göstermektedir.

Son olarak, ChatGPT-4'ün Türkçe eş adlı kelimeleri anlamada ChatGPT-3.5'e göre daha iyi performans göstermesi, daha kapsamlı ve bağlama duyarlı yapay zekâ sistemlerine doğru hızla ilerlendiğini göstermektedir. Bu ilerleme sadece yapay zekânın dil işleme yeteneklerini geliştirmekle kalmamakta aynı zamanda ayrıntılı dil anlayışı gerektiren çeşitli alanlardaki potansiyel kullanımını da artırmaktadır.

KAYNAKÇA

Alawida, M., Mejri, S., Mehmood, A., Chikhaoui, B., & Isaac Abiodun, O. (2023). A comprehensive study of chatgpt: advancements, limitations, and ethical considerations in natural language processing and cybersecurity. *Information*, 14(8). <https://doi.org/10.3390/info14080462>

Atmaca, E. (2013). Eski oğuz türkçesinden günümüz türkiye türkçesine eşgösterenlilik. *Dil Araştırmaları*, 12(12), 117-138.

Chen, L., Zaharia, M., & Zou, J. (2023). How is ChatGPT's behavior changing over time? <https://doi.org/10.48550/arXiv.2307.09009>

Çetiner, M., Yıldırım, A., Onay, B., & Öksüz, C. (2021). KeNet ile kelime anlam belirsizliği giderme. 29th Signal Processing and Communications Applications Conference (SIU).

Doğan, E. (2020). Türkiye Türkçesinin bilinen ilk sesteş kelimeler sözlüğü lisânımızda elfâz-i müteşâbihe (1924) adlı eser üzerine. *Uluslararası Türk Lehçe Araştırmaları Dergisi (Türklad)*, 4(2), 229-249. <https://doi.org/10.30563/turklad.794981>

Eren, U. (2014). *Resimli, eş anlamlı, eş sesli, zıt anlamlı kelimeler sözlüğü*. Yuva yayınları.

Erkınay Tamtamış, H. K. (2021). Türkçede eş adlılık. *Akademik Dil ve Edebiyat Dergisi*, 5(1), 619-653. <https://doi.org/10.34083/akaded.899583>

Eyecioğlu Özmutlu, A. (2021). Doğal dil işleme. In *Bilgisayar Bilimlerinde Teorik Ve Uygulamalı Araştırmalar* (ss. 129-154). Efe Akademi Yayınları.

Harsha, N. S., Kumar, C. N., Sonthi, V. K., & Amarendra, K. (2022). Lexical ambiguity in natural language processing applications. 2022 International Conference on Electronics and Renewable Systems (ICEARS),

Hazar, M., & Tarhan, O. (2013). *Türk anlam bilimi terimleri sözlüğü*. Eğitim Yayınevi.

Heaven, W. D. (2023). *The inside story of how ChatGPT was built from the people who made it: MIT Technology Review*. <https://www.technologyreview.com/2023/03/03/1069311/inside-story-oral-history-how-chatgpt-built-openai/>

Kolektif. (2017). *Eş anlamlı zıt anlamlı eşsesli kelimeler sözlüğü*. Parıltı Yayınları.

Komisyon. (2023). *Eş anlamlı zıt anlamlı ve eş sesli kelimeler sözlüğü*. Ata Yayınları.

Levkovich, I., & Elyoseph, Z. (2023). Suicide risk assessments through the eyes of chatgpt-3.5 versus chatgpt-4: vignette study. *JMIR Ment Health*, 10. <https://doi.org/https://doi.org/10.2196/51232>

Lim, Z. W., Pushpanathan, K., Yew, S. M. E., Lai, Y., Sun, C.-H., Lam, J. S. H., Chen, D. Z., Goh, J. H. L., Tan, M. C. J., & Sheng, B. (2023). Benchmarking large language models' performances for myopia care: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard. *EBioMedicine*, 95. <https://doi.org/10.1016/j.ebiom.2023.104770>

Liu, S., & Fang, Y. (2023). Use large language models for named entity disambiguation in academic knowledge graphs. Proceedings of the 2023 3rd International Conference on Education, Information Management and Service Science (EIMSS 2023),

Makridakis, S., Petropoulos, F., & Kang, Y. (2023). Large language models: their success and impact. *Forecasting*, 5(3), 536-549. <https://doi.org/10.3390/forecast5030030>

Metz, C., Weise, K., Grant, N., & Isaac, M. (2023, December 3). *Ego, fear and money: how the a.i. Fuse Was Lit*. <https://www.nytimes.com/2023/12/03/technology/ai-openai-musk-page-altman.html?smid=url-share>

OpenAI. (2023). *Prompt engineering*. <https://platform.openai.com/docs/guides/prompt-engineering>

Ortega-Martín, M., García-Sierra, Ó., Ardoiz, A., Álvarez, J., Armenteros, J. C., & Alonso, A. (2023). Linguistic ambiguity analysis in ChatGPT. <https://doi.org/10.48550/arXiv.2302.06426>

Plevris, V., Papazafeiropoulos, G., & Jiménez Rios, A. (2023). Chatbots put to the test in math and logic problems: a comparison and assessment of chatgpt-3.5, chatgpt-4, and google bard. *AI*, 4(4), 949-969. <https://doi.org/10.3390/ai4040048>

Ram, B., & Pratima Verma, P. V. (2023). Artificial intelligence AI-based Chatbot study of ChatGPT, Google AI Bard and Baidu AI. *World Journal of Advanced Engineering Technology and Sciences*, 8(1), 258-261. <https://doi.org/10.30574/wjaets.2023.8.1.0045>

Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, 121-154. <https://doi.org/10.1016/j.iotcps.2023.04.003>

Teebagy, S., Colwell, L., Wood, E., Yaghy, A., & Faustina, M. (2023). Improved performance of chatgpt-4 on the okap exam: a comparative study with chatgpt-3.5. *medRxiv*. <https://doi.org/10.1101/2023.04.03.23287957>

Türk Dil Kurumu Sözlükleri. (2022, 04/12/2023). <https://sozluk.gov.tr/>