

YAPAY ZEKÂ DİJİTAL İÇERİĞİ DOĞRULAYABİLİR Mİ? BÜYÜK DİL MODELLERİNİN TÜRKİYE'DEKİ DOĞRULAMA PLATFORMLARIYLA KARŞILAŞTIRMALI ANALİZİ

Erkan YÜKSEL*

Ekmel KILIÇ**

Zeynep AYNA***

Barış GÖKSEL****

Araştırma Makalesi

DOI: 10.5281/zenodo.23056393

Geliş Tarihi: 16.07.2026

Kabul Tarihi: 26.08.2026

ÖZET

Bu çalışma, büyük dil modellerine dayalı yapay zekâ sistemlerinin içerik doğrulama süreçlerindeki performansını Türkiye'deki doğrulama platformlarının kararlarıyla karşılaştırmalı olarak incelemektedir. Araştırma kapsamında ChatGPT, Gemini, Grok ve DeepSeek modellerinin haber metni, fotoğraf-görsel ve video içeriklerinde ürettiği doğrulama sonuçları; Teyit.org, Doğruluk Payı ve Malûmatfûrûş platformlarının yayımladığı doğrulama kararlarıyla karşılaştırılmıştır. Çalışmada nicel içerik analizi ve nitel değerlendirmeyi birlikte kullanan betimsel-karşılaştırmalı bir araştırma tasarımı benimsenmiş; 45 doğrulama içeriğinden oluşan veri seti analiz edilmiştir. Her iddia yapay zekâ modellerine üç aşamalı bir komut sistemiyle yöneltilmiş; toplam 540 olası model çıktısının 516'sı geçerli ve değerlendirilebilir çıktı olarak analize alınmıştır. Bulgular, modellerin doğrulama platformlarının kararlarıyla genel olarak %81,6 oranında uyum gösterdiğini ortaya koymuştur. En yüksek uyum video içeriklerde (%87,2), ardından fotoğraf-görsel (%82,8) ve haber metni içeriklerinde (%75,6) görülmüştür. Ancak video kategorisindeki oran yalnızca değerlendirilebilir çıktılar üzerinden yorumlanmalıdır. Sonuç olarak yapay zekâ modelleri tek başına doğruluk otoritesi değil, insan uzmanlığını destekleyen yardımcı araçlar olarak değerlendirilmelidir.

Anahtar Kelimeler: Yapay zekâ, dezenformasyon, doğrulama platformları, dijital habercilik.

* Prof. Dr., Anadolu Üniversitesi, İletişim Bilimleri Fakültesi, Gazetecilik Bölümü
eyuksel@anadolu.edu.tr, ORCID: 0000-0003-2487-3589

** Doktora öğrencisi, Anadolu Üniversitesi, Lisansüstü Eğitim Enstitüsü, Basın ve Yayın Ana Bilim Dalı,
ekmelkili@hotmail.com, ORCID: 0000-0001-5663-4625

*** Doktora öğrencisi, Anadolu Üniversitesi, Lisansüstü Eğitim Enstitüsü, Basın ve Yayın Ana Bilim
Dalı, zeynepayna@anadolu.edu.tr, ORCID: 0009-0007-1632-8384

**** Araş. Gör., Trabzon Üniversitesi, İletişim Fakültesi, Gazetecilik Bölümü,
barisgoksel@trabzon.edu.tr, ORCID: 0009-0000-2132-5195

CAN ARTIFICIAL INTELLIGENCE VERIFY DIGITAL CONTENT? A COMPARATIVE ANALYSIS OF LARGE LANGUAGE MODELS AND FACT-CHECKING PLATFORMS IN TÜRKİYE

ABSTRACT

This study comparatively examines the performance of artificial intelligence systems based on large language models in content verification processes by comparing their outputs with the decisions of fact-checking platforms in Türkiye. Within the scope of the research, the verification results produced by ChatGPT, Gemini, Grok, and DeepSeek for news-text, photo-visual, and video content were compared with the verification decisions published by Teyit.org, Doğruluk Payı, and Malûmatfuruş. The study adopts a descriptive-comparative research design combining quantitative content analysis and qualitative assessment, and analyzes a dataset consisting of 45 verified content items. Each claim was presented to the artificial intelligence models through a three-stage prompting system; of the 540 possible model outputs, 516 valid and assessable outputs were included in the analysis. The findings show that the models achieved an overall agreement rate of 81.6% with the decisions of the fact-checking platforms. The highest agreement was observed in video content (87.2%), followed by photo-visual content (82.8%) and news-text content (75.6%). However, the agreement rate in the video category should be interpreted only on the basis of assessable outputs. The study concludes that artificial intelligence models should not be regarded as stand-alone authorities of truth, but as auxiliary tools that support human expertise.

Keywords: Artificial intelligence, misinformation, fact-checking platforms, digital journalism

GİRİŞ

Dijital iletişim teknolojilerinin gelişimi, haber üretimi, dolaşımı ve tüketimi üzerinde köklü bir dönüşüm yaratmıştır. Sosyal medya platformları, çevrimiçi haber ağları ve kullanıcı temelli içerik üretimi, bilginin daha hızlı ve geniş ölçekli biçimde yayılmasını mümkün kılarken, yanlış, yanıltıcı ya da bağlamından koparılmış içeriklerin kamusal alanda dolaşıma girmesini de kolaylaştırmıştır. Bu durum, gazeteciliğin temel mesleki ilkelerinden biri olan doğrulama pratiğini dijital çağın en önemli sorun alanlarından biri hâline getirmiştir (Allcott & Gentzkow, 2017; Aïmeur vd., 2023).

Dijital medya ortamında dolaşıma giren yanlış ve yanıltıcı bilgilerin hacmi, hızı ve içerik çeşitliliği, içerik doğrulama süreçlerini giderek daha karmaşık hâle getirmektedir. Günümüzde doğrulanması gereken iddialar yalnızca haber metinlerinden oluşmamakta; fotoğraf, video, ekran görüntüsü, sosyal medya paylaşımı ve bağlamından koparılmış çoklu medya içerikleri de doğrulama süreçlerinin konusu hâline gelmektedir. Bu çeşitlilik, doğrulama faaliyetinin yalnızca metin düzeyinde bir doğru-yanlış değerlendirmesi olarak ele alınamayacağını göstermektedir.

Haber metni temelli iddialar çoğu zaman kaynak karşılaştırması, bağlamsal analiz, tarihsel bilgi veya istatistiksel veriler üzerinden değerlendirilebilirken, fotoğraf-görsel ve video içerikler daha farklı doğrulama teknikleri gerektirmektedir. Bir görselin hangi tarihte ve nerede çekildiğinin

belirlenmesi, eski bir görüntünün yeni bir olayla ilişkilendirilip ilişkilendirilmediğinin anlaşılması, video içeriklerde zaman-mekân ilişkisi ve olay bağlamının çözümlenmesi çoğu zaman teknik inceleme ve editoryal muhakeme gerektirir. Bu nedenle içerik doğrulama, iddianın kaynağı, bağlamı, dolaşım biçimi ve kanıt yapısının birlikte değerlendirilmesini gerektiren çok boyutlu bir süreçtir.

Büyük dil modellerine dayalı yapay zekâ sistemlerinin yaygınlaşması, bu karmaşık doğrulama süreçlerinde yapay zekânın nasıl bir rol üstlenebileceği sorusunu gündeme getirmiştir. Bu modeller, kullanıcı tarafından yöneltilen iddialara hızlı yanıtlar üretebilmekte, bilgi parçaları arasında ilişki kurabilmekte ve kimi durumlarda doğruluk değerlendirmesi yapabilmektedir. Bununla birlikte, büyük dil modellerinin yanıtları güncellik, kaynak güvenilirliği, bağlamsal yorumlama, görsel-video işleme kapasitesi, istem duyarlılığı ve halüsinasyon üretme gibi sınırlılıklardan etkilenebilmektedir. Dolayısıyla bu modellerin ürettiği yanıtların profesyonel doğrulama platformlarının editoryal kararlarıyla ne ölçüde örtüştüğü belirsizliğini korumaktadır.

Mevcut literatürde büyük dil modellerinin bilgi değerlendirme ve doğrulama potansiyeline ilişkin çalışmalar bulunmakla birlikte, bu modellerin Türkiye'deki bağımsız doğrulama platformlarının kararlarıyla karşılaştırmalı olarak incelendiği çalışmalar sınırlıdır. Ayrıca bu alandaki araştırmaların önemli bir bölümü metin temelli iddialara odaklanmakta; haber metni, fotoğraf-görsel ve video içerikleri arasında yapay zekâ performansının nasıl farklılaştığı yeterince açıklığa kavuşturulmamaktadır. Bunun yanında, aynı iddianın farklı komut aşamalarında modellere yöneltilmesi durumunda yanıtların ne ölçüde değiştiği de içerik doğrulama bağlamında önemli bir yönetsel sorundur.

Bu doğrultuda çalışmanın temel problemi, yapay zekâ modellerinin haber metni, fotoğraf-görsel ve video içeriklerinin doğrulanmasında profesyonel doğrulama platformlarının kararlarıyla ne ölçüde uyumlu sonuçlar ürettiğini belirlemektir. Başka bir ifadeyle çalışma, yapay zekâ modellerinin içerik doğrulama süreçlerinde insan uzmanlığına dayalı çok katmanlı teyit süreçlerinin yerini alabilecek düzeyde güvenilir olup olmadığını ya da daha çok yardımcı bir ön değerlendirme aracı olarak mı işlev görebileceğini sorgulamaktadır.

Amaç ve Araştırma Soruları

Çalışmanın genel amacı, büyük dil modellerine dayalı yapay zekâ sistemlerinin içerik doğrulama süreçlerindeki performansını Türkiye'deki bağımsız doğrulama platformlarının kararlarıyla karşılaştırmalı olarak incelemektir. Bu doğrultuda ChatGPT, Gemini, Grok ve DeepSeek modellerinin haber metni, fotoğraf-görsel ve video içeriklerinde ürettiği doğrulama sonuçları; Teyit.org, Doğruluk Payı ve Malûmatfuruş platformlarının yayımladığı doğrulama kararlarıyla karşılaştırılmıştır.

Bu amaç doğrultusunda araştırmada aşağıdaki sorulara yanıt aranmıştır:

1. ChatGPT, Gemini, Grok ve DeepSeek modellerinin doğrulama sonuçları, Teyit.org, Doğruluk Payı ve Malûmatfuruş platformlarının kararlarıyla ne ölçüde örtüşmektedir?
2. Yapay zekâ modellerinin doğrulama performansı haber metni, fotoğraf-görsel ve video içerik türlerine göre nasıl farklılaşmaktadır?
3. Yapay zekâ modelleri hangi içerik türlerinde doğrulama platformlarının kararlarıyla daha yüksek ya da daha düşük uyum göstermektedir?
4. İlk yanıt, yeniden sorgulama ve derinleştirilmiş analiz aşamaları, yapay zekâ modellerinin doğrulama sonuçlarını nasıl etkilemektedir?

5. Yapay zekâ modellerinin ürettiği doğrulama çıktıları, insan uzmanlığına dayalı profesyonel doğrulama süreçlerinin yerini alabilecek düzeyde güvenilir midir, yoksa daha çok yardımcı bir ön değerlendirme aracı olarak mı işlev görmektedir?

Önem ve Özgün Değer

Çalışma, yapay zekâ modellerinin içerik doğrulama süreçlerindeki rolünü Türkiye'deki bağımsız doğrulama platformlarının kararlarıyla karşılaştırmalı olarak incelemesi bakımından önemlidir. Büyük dil modelleri son yıllarda bilgiye erişim, metin üretimi, sınıflandırma ve değerlendirme süreçlerinde yaygın biçimde kullanılmaya başlanmıştır. Ancak bu modellerin doğrulama platformlarının yerini alabilecek düzeyde güvenilir olup olmadığı ya da daha çok yardımcı bir ön değerlendirme aracı olarak mı işlev görebileceği hâlen tartışmalı bir konudur. Çalışma, söz konusu tartışmaya Türkiye bağlamında ampirik bir veri seti üzerinden katkı sunmayı amaçlamaktadır.

ChatGPT, Gemini, Grok ve DeepSeek modellerini aynı araştırma tasarımı içinde karşılaştırmalı olarak ele alınması, çalışmanın özgün değerlerinden biridir. Çalışma, farklı modellerin doğrulama kararları, içerik türleri ve sorgulama aşamaları bakımından nasıl farklılaştığını incelemektedir.

Çalışmanın bir diğer özgün yönü, doğrulama içeriklerini yalnızca haber metni temelli iddialarla sınırlamaması; haber metni, fotoğraf-görsel ve video içeriklerini ayrı kategoriler olarak ele almasıdır. Ayrıca üç aşamalı komut sistemi kullanılarak model yanıtlarının ilk yanıt, yeniden sorgulama ve derinleştirilmiş analiz aşamalarında nasıl değiştiği incelenmiştir. Bu yönleriyle araştırma, yapay zekâ destekli doğrulama tartışmasına Türkiye bağlamından, içerik türleri ve insan uzmanlığı ekseninde katkı sunmaktadır.

ALANYAZIN TARAMASI

Dijital medya ortamında haber üretimi, dolaşımı ve tüketimi doğrulama pratiklerini gazeteciliğin temel sorun alanlarından biri hâline getirmiş; sosyal medya platformları, çevrimiçi haber ağları ve kullanıcı temelli içerik üretimi yanlış, yanıltıcı ve bağlamından koparılmış içeriklerin görünürlüğünü artırmıştır. Bu nedenle doğrulama, yalnızca doğru-yanlış tespitine yönelik teknik bir işlem değil; gazeteciliğin güvenilirlik, kamusal sorumluluk ve hesap verebilirlik işlevleriyle ilişkili çok katmanlı bir değerlendirme süreci olarak ele alınmalıdır. Yapay zekâ modellerinin içerik doğrulama süreçlerinde kullanılmaya başlanması ise bu tartışmaya yeni bir boyut kazandırmış; büyük dil modelleri hızlı ön değerlendirme ve iddia sınıflandırma potansiyeli taşımakla birlikte güncellik, bağlam, kaynak güvenilirliği, görsel-video analizi ve hata üretme riski bakımından önemli sınırlılıklar barındırmıştır. Bu çerçevede literatür taraması; dijital dezenformasyon ve içerik doğrulama, Türkiye'deki doğrulama platformları, otomatik doğrulama yaklaşımları ve yapay zekâ destekli doğrulamanın sınırlılıkları ekseninde yapılandırılmıştır.

Dijital Dezenformasyon ve İçerik Doğrulama

Dijital medya ortamı, haber üretimi ve dolaşımını hızlandırırken doğrulanmamış, eksik, yanıltıcı ya da bağlamından koparılmış bilgilerin geniş kitlelere kısa sürede ulaşmasına imkân tanımaktadır. Sosyal medya platformları ve kullanıcı temelli içerik üretimi, kamusal bilgi akışını çoğullaştırmakla birlikte doğrulama pratiğini daha karmaşık ve çok katmanlı bir süreç hâline getirmiştir. Diğer yandan haber doğrulama, gazeteciliği eğlence, propaganda ya da kurgusal anlatı biçimlerinden ayıran temel ilkelerden biridir. Geleneksel habercilik anlayışında doğrulama, olayları, iddiaları ve kaynakları denetleyerek “ne olup bittiğini” anlama çabasıyla ilişkilendirilirken; dijital iletişim ortamında bu süreç gerçek zamanlı etkileşim, kullanıcı katılımı

ve yoğun içerik dolaşımı nedeniyle daha dinamik bir niteliğe bürünmüştür. Trewinnard ve Bell'e (2018, s. 103) göre özellikle kriz ve toplumsal hareket dönemlerinde ortaya çıkan açık doğrulama anlayışı, doğrulama süreçlerini daha kamusal ve işbirlikçi bir zemine taşımış; ancak doğrulanmamış bilgilerin viral hâle gelme riski, bu süreçlerde metodolojik dikkatin önemini artırmıştır.

Bu dönüşüm, doğrulamanın habercilik içindeki konumunu da yeniden tanımlamıştır. Doğrulama artık yalnızca yanlış bilgiyi düzeltme ya da hatalı iddiaları sınıflandırma faaliyeti değildir. Aynı zamanda kamusal alanda dolaşıma giren iddiaları, aktörleri, kaynakları ve kanıtları sorgulayan aktif bir denetim mekanizmasıdır. Graves'e (2016, s. 37, 66) göre doğrulama hareketi, gazeteciliğin kamu adına hesap sorma misyonunu yeniden görünür kılmakta ve medya kurumlarının sarsılan güvenilirliğini onarmaya yönelik stratejik bir müdahale niteliği taşımaktadır.

Doğrulama süreci, dijital ortamda yalnızca haber metni temelli iddialarla sınırlı değildir. Akıllı telefonların yaygınlaşması, mobil internet erişiminin artması ve görsel içerik üretiminin kolaylaşması; fotoğraf, video, ekran görüntüsü ve ses kaydı gibi çoklu medya içeriklerinin haber dolaşımında daha etkili hâle gelmesine yol açmıştır. Bu nedenle doğrulama uzmanları, dijital içerikleri değerlendirirken şüpheli bir yaklaşım benimsemek, sistematik kontrol listeleri kullanmak, varsayımlardan kaçınmak ve anonim kaynaklara karşı dikkatli olmak durumundadır. Trewinnard ve Bell'e (2018, s. 104-109) göre dijital doğrulama süreci; içeriğin kaynağı, bağlamı, konumu, zamanı ve güvenilirliği hakkında kanıta dayalı bir sonuca ulaşmayı gerektirir. Tersine görsel arama, video küçük resimlerinin incelenmesi, üst veri analizi, konum doğrulama, hava durumu kayıtları, gölge analizi ve uydu görüntüleriyle karşılaştırma gibi yöntemler de bu sürecin önemli araçlarıdır.

Dezenformasyonla mücadelede hangi iddiaların doğrulanacağına karar verme süreci de önem taşımaktadır. Her yanlış ya da yanıltıcı içeriğin doğrulanması, kimi durumlarda söz konusu içeriğin daha fazla görünürlük kazanmasına yol açabilir. Wardle ve Derakhshan'a (2017, s. 19, 66) göre henüz geniş kitlelere ulaşmamış bazı dezenformasyon içeriklerini doğrulamak, manipülatörlerin görünürlük hedeflerine dolaylı biçimde katkı sunabilir. Bu bağlamda "yanlış bilgiye oksijen sağlamama" ilkesi, doğrulama süreçlerinde stratejik ve metodolojik bir karar alanı olarak önem kazanmaktadır.

Dijital dezenformasyonun etkisi, içerik türüne göre de farklılaşmaktadır. Haber metni temelli yanlış bilgiler çoğu zaman kaynak karşılaştırması, veri kontrolü ve mantıksal tutarlılık analiziyle değerlendirilebilirken fotoğraf-görsel ve video içerikler daha güçlü bir bilişsel etki yaratabilmektedir. Üretici yapay zekâ teknolojilerinin gelişmesiyle birlikte sahte ses, yapay video ve sentetik görsellerin tespiti daha karmaşık bir sorun alanına dönüşmüştür. Wardle ve Derakhshan (2017, s. 40), nitelikli manipülasyonları saptayabilecek yapay zekâ destekli otomatik araçların geliştirilmesini bilgi düzensizliğiyle mücadelede önemli bir savunma hattı olarak değerlendirmektedir. Bu çerçevede içerik doğrulama, hem gazeteciliğin doğruluk ve güvenilirlik işlevini sürdürmesi hem de dijital ortamın hız, hacim ve çoklu medya yapısı karşısında yeni yöntemler geliştirmesi bakımından kritik öneme sahiptir.

Türkiye'de Doğrulama Platformları

Türkiye'de dijital dezenformasyonla mücadele alanında faaliyet gösteren doğrulama girişimleri, sosyal medya ve çevrimiçi haber ortamlarında dolaşıma giren şüpheli içeriklerin incelenmesi bakımından önemli bir işleve sahiptir. Bu platformlar, yanlış ya da yanıltıcı bilgilerin saptanması, kamusal dolaşımdaki iddiaların denetlenmesi ve kullanıcıların güvenilir bilgiye erişiminin desteklenmesi açısından habercilik ekosisteminde tamamlayıcı bir rol

üstlenmektedir. Kavaklı'ya (2019, s. 667-668) göre internet tabanlı iletişim araçlarının yaygınlaşması, yalan haberin dolaşım hızını artırmakta ve geleneksel habercilikteki teyit mekanizmalarını yetersiz bırakabilmektedir. Bu bağlamda doğrulama platformları, geleneksel haber merkezlerinin zaman ve kaynak sınırlılıkları nedeniyle her zaman yerine getiremediği ayrıntılı teyit süreçlerini üstlenen uzmanlaşmış yapılar olarak değerlendirilebilir.

Türkiye'de düzenli içerik üreten ve kurumsal görünürlüğe sahip doğrulama girişimleri arasında Teyit.org, Doğruluk Payı, Malûmatfûruş ve Yalansavar dikkat çekmektedir. Bu platformlar ortak biçimde yanlış bilginin tespiti ve düzeltilmesini amaçlamakla birlikte, konu odakları, değerlendirme ölçütleri ve sınıflandırma sistemleri bakımından birbirinden farklılaşmaktadır.

Teyit.org, özellikle dijital medya ortamlarında ve sosyal ağlarda dolaşıma giren şüpheli iddiaları açık kaynak araştırmaları, veri analizi ve görsel doğrulama teknikleriyle incelemektedir. Platformun doğrulama çıktıları "doğru", "yanlış", "karma" ve "sonuçlandırılmadı" gibi sınıflandırmalarla ifade edilmektedir (Teyit, t.y.-a). Teyit.org, doğrulama faaliyetlerini tarafsızlık, bağımsızlık ve şeffaflık ilkeleri doğrultusunda yürüttüğünü belirtmekte; incelenecek iddiaları ise yaygınlık, toplumsal etki ve aciliyet gibi ölçütler çerçevesinde belirlemektedir (Teyit, t.y.-b).

Doğruluk Payı, başlangıçta siyasal aktörlerin kamuoyuna yansıyan açıklamalarını denetlemeye odaklanan, zamanla kapsamını genişleterek farklı medya mecralarında dolaşıma giren şüpheli bilgilere yönelik doğrulama çalışmaları da yürüten bir platformdur (Doğruluk Payı, t.y.-a). Platform, ulaştığı sonuçları "doğru", "yanlış", "kısmen yanlış" ve "belirsiz" gibi kategorilerle sınıflandırmaktadır (Doğruluk Payı, t.y.-b). Bu sınıflandırma sistemi, doğrulama kararlarının yalnızca ikili bir doğru-yanlış ayrımıyla değil, kimi zaman ara kategorilerle de değerlendirilebildiğini göstermektedir.

Malûmatfûruş ise özellikle köşe yazıları, medya içerikleri ve kamusal alanda tekrar eden hatalı iddialar üzerine yoğunlaşan bir yanlışlama pratiği geliştirmiştir. Platformun yaklaşımı, açık kaynaklar aracılığıyla somut biçimde yanlışlanabilen iddiaların incelenmesine dayanmaktadır. İncelenecek iddialar gündem takibi, kullanıcı bildirimleri ve sosyal medya taraması yoluyla belirlenmekte; yaygınlık, önemlilik ve aciliyet gibi ölçütler dikkate alınmaktadır (Malûmatfûruş, t.y.-a, t.y.-b). Bu yönüyle Malûmatfûruş, yalnızca güncel sosyal medya iddialarını değil, medya metinlerinde kalıplaşmış ya da tekrar eden hatalı bilgileri de konu edinmektedir.

Yalansavar ise gönüllülük esasına dayalı yapısı ve bilimsel yanlışlara odaklanan içeriğiyle Türkiye'deki doğrulama ekosisteminin bir diğer bileşenidir. Platform, özellikle bilimsel temeli zayıf, yanlış ya da yanıltıcı iddiaları eleştirel düşünce ve bilimsel yöntem ekseninde değerlendirmeyi amaçlamaktadır (Yalansavar, t.y.-a, t.y.-b). Bu durum, Türkiye'de doğrulama faaliyetlerinin yalnızca siyasal iddialar ya da güncel sosyal medya içerikleriyle sınırlı olmadığını; sağlık, bilim, teknoloji ve gündelik yaşamda dolaşıma giren hatalı bilgi biçimlerini de kapsadığını göstermektedir.

Türkiye'de doğrulama alanında farklı kurumsal niteliklere sahip başka yapılar da bulunmaktadır. Örneğin Evrim Ağacı bilim iletişimi ve bilimsel içerik doğrulama alanında; Dezenformasyonla Mücadele Merkezi ise kamu kurumu kaynaklı bilgilendirme ve doğrulama faaliyetleriyle öne çıkmaktadır.

Bu çalışma açısından ise örneklem; Teyit.org, Doğruluk Payı ve Malûmatfûruş platformlarıyla sınırlandırılmıştır. Bu tercihin nedeni, söz konusu üç platformun araştırma döneminde haber metni, fotoğraf-görsel ve video içeriklerindeki iddialar bakımından karşılaştırmalı bir veri seti oluşturmaya elverişli düzenli içerik üretimine sahip olmasıdır. Türkiye'de doğruluk kontrolü uygulamalarına ilişkin çalışmalar genel olarak bu platformların dijital medya ekosisteminde

yanlış bilgiyle mücadele açısından tamamlayıcı bir işleve sahip olduğunu göstermektedir (Evren ve Türedi, 2022).

Büyük Dil Modelleri ve Otomatik Doğrulama

Geleneksel doğrulama süreçleri; iddianın belirlenmesi, kaynakların taranması, kanıtların karşılaştırılması, uzman görüşlerine başvurulması ve nihai bir editoryal karar verilmesi gibi emek yoğun aşamalardan oluşmaktadır. Ancak dijital ortamda her gün çok sayıda haber metni, fotoğraf-görsel ve video içerikli iddianın dolaşıma girmesi, insan emeğine dayalı doğrulama süreçlerinin kapasitesini zorlamaktadır. Bu nedenle otomatik doğrulama sistemleri, özellikle doğrulanmaya değer iddiaların belirlenmesi, ilgili kanıtların toplanması ve iddianın doğruluk değerine ilişkin ön değerlendirme yapılması açısından giderek daha fazla tartışılmaktadır (Guo vd., 2022, s. 1-4).

Büyük dil modelleri, bu tartışmanın merkezinde yer alan teknolojilerden biridir. Doğal dil işleme alanındaki gelişmelerle birlikte büyük veri kümeleri üzerinde eğitilen bu modeller, metin üretme, özetleme, sınıflandırma, soru yanıtlama ve anlam ilişkileri kurma gibi görevlerde yüksek performans sergileyebilmektedir. ChatGPT, Gemini, Grok ve DeepSeek gibi uygulamalar da temelde büyük dil modellerine dayalı sohbet sistemleri olarak değerlendirilebilir. Bu sistemler, kullanıcıdan gelen doğal dildeki istemleri işleyerek yanıt üretmekte; kimi durumlarda bir iddiaya ilişkin olasılıklı değerlendirme, bağlamsal açıklama ya da sınıflandırma yapabilmektedir. Bu yönüyle büyük dil modelleri, içerik doğrulama süreçlerinde doğrudan nihai karar verici olmaktan çok, hızlı ön analiz ve bilgi değerlendirme desteği sunabilecek araçlar olarak konumlanmaktadır (Naveed vd., 2024).

ChatGPT'nin 2022 yılında geniş kullanıcı kitlelerine açılması, büyük dil modellerinin akademik ve teknik kullanım alanlarından gündelik bilgi arama, metin üretme ve değerlendirme pratiklerine doğru yayılmasını hızlandırmıştır. Bu dönüşümde komut odaklı ayarlama ve insan geri bildirimleriyle pekiştirilmiş öğrenme teknikleri önemli rol oynamıştır. Ouyang ve arkadaşlarına (2022) göre, büyük dil modellerinin yalnızca ölçek bakımından büyütülmesi kullanıcı niyetine uygun davranış üretmek için yeterli değildir; insan geri bildirimleriyle yapılan ince ayarlama, modellerin talimatları daha iyi izlemesini ve kullanıcı beklentilerine daha uygun çıktılar üretmesini sağlamaktadır.

Literatürde büyük dil modellerinin bilgi ve içerik doğrulama performansını inceleyen çalışmalar, bu sistemlerin belirli bir potansiyele sahip olduğunu, ancak insan temelli doğrulama süreçlerinin yerini bütünüyle alabilecek düzeyde güvenilir olmadığını göstermektedir. Caramancion (2023), ChatGPT-3.5, ChatGPT-4.0, Google Bard/LaMDA ve Microsoft Bing AI modellerini bağımsız teyit kuruluşları tarafından doğruluğu belirlenmiş 100 haber üzerinden karşılaştırmıştır. Araştırmada modellerin ortalama doğruluk oranının %65,25 olduğu; ChatGPT-4.0'ın diğer modellere göre daha tutarlı sonuç ürettiği, ancak modellerin bağlamsal derinlik ve nüansları yakalama bakımından insan temelli teyit süreçlerinin gerisinde kaldığı belirtilmiştir. Benzer şekilde Kuznetsova ve arkadaşları (2025), ChatGPT ve Microsoft Copilot'un siyasal bilgileri doğrulama konusundaki performansını farklı diller ve tartışmalı konu başlıkları üzerinden incelemiştir. Araştırma, büyük dil modellerinin doğrulama performansının yalnızca model mimarisine değil; dil, bağlam, konu alanı ve erişilebilir bilgi kaynaklarına da bağlı olduğunu göstermektedir. Hoes ve arkadaşları (2023) ise ChatGPT'nin doğrulama süreçlerinde verimlilik sağlayabileceğini, ancak bu kullanımın insan denetimiyle birlikte düşünülmesi gerektiğini vurgulamaktadır.

Büyük dil modellerinin siyasal bilgiye ilişkin yanıtları farklı politik ve kültürel bağlamlarda değişkenlik gösterebilmektedir. Urman ve Makhortykh (2025), ChatGPT, Microsoft Copilot ve

Bard/Gemini gibi sohbet sistemlerinin siyasal sorulara verdikleri yanıtları farklı dillerde karşılaştırmış; modellerin siyasal bilgiye erişim, yanıt verme ya da yanıt sınırlama biçimlerinde belirgin farklılıklar olduğunu göstermiştir. Bu bulgu, büyük dil modellerinin içerik doğrulamada kullanılmasının yalnızca teknik değil, aynı zamanda politik ve etik boyutları olan bir mesele olduğunu ortaya koymuştur.

Diğer yandan, sağlık iletişimi alanında yapılan çalışmalar da bu modellerin doğru ve anlaşılır bilgi sunma potansiyeline sahip olduğunu, ancak uzman denetimi olmadan kullanıldığında riskler barındırdığını göstermektedir. Deiana ve arkadaşları (2023), ChatGPT'nin aşılarda ilgili mitlere verdiği yanıtları doğruluk, açıklık ve kapsamlılık ölçütleri üzerinden değerlendirmiş; yanıtların büyük ölçüde doğru ve anlaşılır olduğunu, ancak bazı konularda bilimsel hata, eksik bağlamlandırma ve aşırı genelleştirme sorunlarının ortaya çıkabildiğini belirtmiştir.

Bu bulgular, büyük dil modellerinin içerik doğrulama süreçlerinde hızlı ön analiz, iddia sınıflandırması, olası kaynak alanlarının belirlenmesi ve tutarsızlıkların işaretlenmesi bakımından yararlı olabileceğini göstermektedir. Ancak doğrulama süreci yalnızca dilsel bir sınıflandırma işlemi olmadığından, modellerin ürettiği yanıtların kaynak temelli teyit, güncel veri kontrolü ve editöryal değerlendirme süreçleriyle birlikte ele alınması gerekmektedir. Başka bir ifadeyle, büyük dil modelleri doğrulama sürecini hızlandırabilir; ancak doğrulamanın habercilikteki kanıta dayalı karar verme niteliğini tek başına ikame edemez (Sharma vd., 2019). Ayrıca büyük dil modellerinin her durumda daha küçük ve alana uyarlanmış modellerden üstün olmadığı; sınıflandırma görevlerinde ince ayarlı modellerin hâlen avantaj sağlayabildiği de belirtilmektedir (Bosley vd., 2023).

Yapay Zekâ Destekli Doğrulamanın Sınırlılıkları

Büyük dil modellerine dayalı yapay zekâ modellerinin doğrulama süreçlerinde kullanımı, çeşitli yöntemsel, teknik ve etik sınırlılıkları da beraberinde getirmiştir. Bu temel sınırlılıklardan biri, modellerin bilgi güncelliğiyle ilişkilidir. Büyük dil modelleri, eğitildikleri veri setleri ve erişebildikleri kaynaklarla sınırlı biçimde yanıt üretmektedir. Bu nedenle güncel gelişmeler, sonradan düzeltilmiş bilgiler veya yakın tarihli olaylar söz konusu olduğunda modeller eksik, eski ya da hatalı değerlendirmelerde bulunabilmektedir. İçerik doğrulama gibi zamana duyarlı bir alanda, bir iddianın doğruluğu çoğu zaman güncel veri tabanlarına, resmî açıklamalara, güncel haber arşivlerine ve açık kaynak araştırmalarına erişimi gerektirmektedir.

İkinci sınırlılık, modellerin bağlamsal yorumlama kapasitesiyle ilgilidir. Büyük dil modelleri metinler arasında anlam ilişkileri kurabilmekte ve olası açıklamalar üretebilmektedir; ancak bu açıklamalar her zaman doğrulamanın gerektirdiği bağlamsal derinliği sağlayamayabilmektedir. Özellikle siyasal içerikler, kültürel referanslar, ironi, tarihsel göndermeler, yerel bağlamlar ve çok katmanlı toplumsal olaylar söz konusu olduğunda modeller yüzeysel ya da genelleşici yorumlar üretebilmektedir. Kuznetsova ve arkadaşları (2025) ile Urman ve Makhortykh (2025), büyük dil modellerinin siyasal bilgi doğrulama performansının dil, konu ve politik bağlama göre değişebildiğini ortaya koymaktadır.

Üçüncü sınırlılık, yapay zekâ modellerinin halüsinasyon olarak adlandırılan hatalı ya da uydurma bilgi üretme eğilimidir. Büyük dil modelleri, olasılıksal dil örüntüleri üzerinden yanıt ürettikleri için kimi durumlarda gerçekte var olmayan kaynaklara, yanlış tarihsel bilgilere ya da eksik kanıtlara dayalı ikna edici açıklamalar sunabilmektedir. Caramancion'un (2023) çalışması da modellerin içerik doğrulama konusunda belirli bir performans sergilemekle birlikte, insan temelli doğrulama kuruluşlarının yerini alabilecek düzeyde olmadığını göstermektedir. Dolayısıyla model tarafından üretilen yanıtın dilsel olarak tutarlı ve ikna edici görünmesi, onun doğruluğu için yeterli bir ölçüt olarak kabul edilmemelidir.

Dördüncü sınırlılık, fotoğraf-görsel ve video içeriklerin analizinde ortaya çıkmaktadır. Haber metni temelli iddialar çoğu zaman dilsel çözümleme, kaynak karşılaştırması ve mantıksal tutarlılık üzerinden değerlendirilebilirken, görsel ve video içerikler daha karmaşık doğrulama teknikleri gerektirmektedir. Bir fotoğrafın ne zaman, nerede ve hangi bağlamda çekildiğinin belirlenmesi; eski bir görüntünün yeni bir olayla ilişkilendirilip ilişkilendirilmediğinin anlaşılması; video içeriklerde ses, görüntü, zaman, mekân ve olay örgüsünün birlikte analiz edilmesi çoğu zaman özel teknik araçlar ve uzmanlık gerektirmektedir. Trewinnard ve Bell'e (2018, s. 106-109) göre tersine görsel arama, video küçük resimlerinin incelenmesi, üst veri analizi, konum doğrulama, hava durumu kayıtları ve gölge analizi gibi yöntemler görsel doğrulama süreçlerinde kritik öneme sahiptir.

Beşinci sınırlılık, modellerin isteme ya da komuta duyarlılığıdır. Aynı iddia, farklı biçimde sorulduğunda ya da modele yeniden yöneltildiğinde farklı yanıtlar üretilebilmektedir. Bu durum, model yanıtlarının sabit ve değişmez doğrulama kararları olarak ele alınamayacağını göstermektedir. Bu çalışmada kullanılan üç aşamalı komut sistemi de bu nedenle önem taşımaktadır; çünkü modelin ilk yanıtı, yeniden sorgulama karşısındaki tutarlılığı ve derinleştirilmiş analiz aşamasındaki performansı birlikte değerlendirilmektedir.

Son olarak, yapay zekâ çıktılarının insan kodlayıcılar tarafından farklı biçimlerde sınıflandırılabilmesi de önemli bir yöntemsel sınırlılık alanıdır. Bir modelin ürettiği yanıt her zaman açık biçimde “doğru”, “yanlış” ya da “tartışmalı” kategorilerinden birine karşılık gelmeyebilir. Model bazen “kanıt bulunmamaktadır”, “iddia doğrulanmamaktadır”, “büyük ölçüde yanlış görünmektedir” ya da “bağlama bağlıdır” gibi ara ifadeler kullanabilir. Bu tür yanıtların hangi kategoriye yerleştirileceği, kodlayıcının yorumuna bağlı olarak değişebilir. Bu nedenle kodlama kategorilerinin açık biçimde tanımlanması ve kodlayıcılar arası güvenilirliğin hesaplanması, yöntemin geçerliliği açısından önemlidir.

Bu sınırlılıklar, yapay zekâ modellerinin doğrulama süreçlerinde bütünüyle dışlanması gerektiği anlamına gelmemektedir. Aksine, bu modeller özellikle ilk tarama, iddia ön sınıflandırması, olası kaynak alanlarının belirlenmesi, tutarsızlıkların işaretlenmesi ve doğrulama uzmanlarının iş yükünün azaltılması bakımından yararlı araçlar olabilir. Ancak bu katkı, yapay zekânın nihai doğruluk otoritesi olarak konumlandırılmasıyla değil, insan uzmanlığı ve editöryal denetimle birlikte kullanılmasıyla anlamlı hâle gelmektedir. Bu nedenle yapay zekâ destekli doğrulama süreçleri, insan uzmanlığını dışlayan değil, onu destekleyen hibrit bir model çerçevesinde değerlendirilmelidir.

Sonuç olarak mevcut literatür, dijital dezenformasyonun içerik doğrulama süreçlerini daha karmaşık hâle getirdiğini, doğrulama platformlarının kamusal bilgi alanı açısından önemli bir işlev üstlendiğini ve büyük dil modellerinin bilgi değerlendirme süreçlerinde belirli bir potansiyele sahip olduğunu ortaya koymaktadır.

YÖNTEM

ChatGPT, Gemini, Grok ve DeepSeek modellerinin doğrulama performansını Teyit.org, Doğruluk Payı ve Malûmatfûruş tarafından incelenmiş haber metni, fotoğraf-görsel ve video gibi farklı türdeki içerikler üzerinden karşılaştırmalı biçimde ele alarak alandaki boşluğa katkı sunmayı amaçlayan bu çalışmada, betimsel-karşılaştırmalı bir araştırma tasarımı benimsenmiştir. Karasar'a (2012) göre tarama modelleri, geçmişte ya da hâlen var olan bir durumu var olduğu biçimiyle betimlemeyi amaçlayan araştırma yaklaşımlarıdır. Bu çalışmada da yapay zekâ modellerinin doğrulama platformlarının kararlarıyla uyum düzeyi; model türü, içerik türü, doğrulama platformu ve komut aşaması değişkenleri bakımından betimlenmiş ve karşılaştırmalı olarak analiz edilmiştir.

Araştırmada nicel içerik analizi aracılığıyla yapay zekâ modellerinin doğrulama platformlarının kararlarıyla ne ölçüde örtüştüğü belirlenmiş; nitel değerlendirme yoluyla ise model yanıtlarının içerik türlerine, sorgulama aşamalarına ve bağlamsal açıklama biçimlerine göre nasıl farklılaştığı yorumlanmıştır.

İçerik analizi, iletişim içeriklerinin belirli kategoriler doğrultusunda sistematik, nesnel ve tekrarlanabilir biçimde incelenmesine dayanan bir araştırma tekniğidir. Neuendorf'a (2002) göre içerik analizi, mesaj özelliklerinin sistematik, nesnel ve nicel olarak çözümlenmesini amaçlayan bir yöntemdir. Bu çalışmada da doğrulama platformları tarafından yayımlanan içerikler ve yapay zekâ modellerinin bu içeriklere verdiği yanıtlar, önceden belirlenen kategoriler doğrultusunda kodlanmış ve karşılaştırmalı olarak analiz edilmiştir.

Araştırmanın karşılaştırmalı niteliği üç düzeyde kurulmuştur. İlk olarak büyük dil modellerinin doğrulama sonuçları kendi aralarında karşılaştırılmıştır. İkinci olarak model yanıtları haber metni, fotoğraf-görsel ve video içerik türlerine göre analiz edilmiştir. Üçüncü olarak ise modellerin ilk yanıt, yeniden sorgulama ve derinleştirilmiş analiz aşamalarında ürettikleri sonuçlar doğrulama platformlarının kararlarıyla karşılaştırılmıştır. Bu tasarım, yapay zekâ modellerinin içerik doğrulama performansını tek bir yanıt ya da tek bir içerik türü üzerinden değil, çok boyutlu bir değerlendirme çerçevesi içinde incelemeye imkân tanımıştır.

Evren ve Örneklem

Araştırmanın evrenini, dijital medya ortamında dolaşıma giren ve Türkiye'deki doğrulama platformları tarafından incelenen şüpheli haber ve iddia içerikleri oluşturmaktadır. Araştırmanın örnekleme ise Teyit.org, Doğruluk Payı ve Malûmatfuruş platformlarında 1-30 Kasım 2025 döneminde yayımlanan haber metni, fotoğraf-görsel ve video içerikli doğrulama açıklamalarından oluşmaktadır.

Çalışmada amaçlı örnekleme tekniği kullanılmıştır. Amaçlı örnekleme, araştırmanın amacına uygun olarak bilgi açısından zengin, araştırma problemine doğrudan katkı sağlayabilecek birimlerin seçilmesine dayanan bir örnekleme yaklaşımıdır (Patton, 2015). Bu çalışmada da örnekleme alınan içerikler rastlantısal olarak değil; araştırmanın amacı, karşılaştırmalı analiz tasarımı ve içerik türleri arasındaki denge gözetilerek belirlenmiştir.

İncelenecek zaman aralığının belirlenmesinde araştırmanın yürütüldüğü akademik bağlam da etkili olmuştur. Çalışma, Anadolu Üniversitesi Lisansüstü Eğitim Enstitüsü Basın ve Yayın Doktora Programı'nda 2025-2026 güz döneminde yürütülen Gündem Belirleme Araştırmaları dersi kapsamında, dersin sorumlu öğretim üyesi ve üç doktora öğrencisi tarafından gerçekleştirilmiştir. Bu nedenle örneklem, dersin yürütüldüğü dönemde doğrulama platformlarında yayımlanan güncel içerikler arasından seçilmiştir. Güncel içeriklerin tercih edilmesi, yapay zekâ modellerinin eski ve yaygın biçimde dolaşıma girmiş içeriklere ilişkin önceden öğrenilmiş bilgi örüntülerine dayanma olasılığını sınırlamak ve modellerin yakın dönemli iddialar karşısındaki doğrulama performansını değerlendirmek amacıyla yapılmıştır.

Araştırma kapsamına alınan içeriklerin seçiminde şu ölçütler esas alınmıştır:

1. İçeriklerin ilgili doğrulama platformunda 1-30 Kasım 2025 döneminde yayımlanmış olması;
2. İçeriklerin açık biçimde haber metni, fotoğraf-görsel ya da video içerik kategorilerinden birine yerleştirilebilmesi;
3. İçeriğin doğrulama platformu tarafından nihai bir karar kategorisiyle sonuçlandırılmış olması;

4. İddianın yapay zekâ modellerine yöneltilebilecek açıklıkta ve karşılaştırılabilir nitelikte olması;
5. İçeriklerin, yapay zekâ modellerinin içerik doğrulama performansını farklı içerik türleri açısından değerlendirmeye imkân verecek biçimde çeşitlilik göstermesi.

Örneklem seçiminde içerik türleri arasında denge kurulmasına dikkat edilmiştir. Buna göre her doğrulama platformundan 5 haber metni, 5 fotoğraf-görsel ve 5 video içerikli doğrulama açıklaması olmak üzere 15 içerik seçilmiştir; toplamda 45 doğrulama içeriği araştırma kapsamına alınmıştır. İçerik analizine dâhil edilen doğrulama içeriklerinin başlıkları, içerik türleri, platform kararları ve web adresleri araştırmacılar tarafından oluşturulan Excel tabanlı veri matrisinde kayıt altına alınmıştır.

Veri Toplama Süreci

Veri toplama süreci iki aşamada gerçekleştirilmiştir. İlk aşamada, Teyit.org, Doğruluk Payı ve Malûmatfuruş platformlarında 1-30 Kasım 2025 döneminde yayımlanan doğrulama içerikleri taranmış ve araştırma ölçütlerine uygun içerikler belirlenmiştir. Her içerik için başlık, içerik türü, doğrulama platformu, platformun nihai doğrulama kararı, içeriğin temel iddiası ve web adresi veri matrisine işlenmiştir.

İkinci aşamada, belirlenen her iddia ChatGPT, Gemini, Grok ve DeepSeek modellerine aynı komut sistemi kullanılarak yöneltilmiştir. Sorgulamalar 11–24 Mart 2026 tarihleri arasında gerçekleştirilmiştir. Araştırmada ChatGPT, Gemini 3 Flash, Grok 4.1 ve DeepSeek-V3.2 kullanılmıştır. İlk iki sorgulama aşamasında modellerin standart yanıt modu kullanılmış; üçüncü aşamada ise modellerin sunduğu Thinking / Deep Think seçeneği etkinleştirilerek derinleştirilmiş analiz gerçekleştirilmiştir. ChatGPT’de üçüncü aşama GPT-5.4 Thinking üzerinden yürütülmüştür.

Toplam 45 içerik dikkate alındığında araştırma kapsamında 540 olası model çıktısı planlanmıştır. Ancak özellikle bazı video içeriklerde, ilgili video bağlantılarının veya içeriklerin ChatGPT ve DeepSeek modelleri tarafından tanınamaması ya da işlenememesi nedeniyle geçerli yanıt üretilemeyen durumlar ortaya çıkmıştır. Bu nedenle analizlerde yalnızca geçerli ve değerlendirilebilir nitelikteki model çıktıları dikkate alınmıştır. Araştırma kapsamında toplam 516 geçerli model çıktısı değerlendirmeye alınmıştır. Geçerli çıktı alınamayan 24 durum, model kararının doğrulama platformu kararıyla karşılaştırılmasına imkân vermediği için uyum oranlarının hesaplanmasında paydadın çıkarılmıştır.

Görsel ve video içerikler, modellerin teknik olanak verdiği durumlarda doğrudan dosya ya da ekran görüntüsü olarak sunulmuş; bunun mümkün olmadığı durumlarda içerik bağlantısı, başlık ve iddia metni üzerinden sorgulama yapılmıştır. Tüm modellerde aynı içerikler aynı sırayla ve aynı komut yapısıyla sorgulanmıştır. Bu uygulama, model çıktıları arasında karşılaştırılabilirliği artırmak amacıyla tercih edilmiştir.

Yapay Zekâ Modellerine Uygulanan Komut Sistemi

Araştırmada yapay zekâ modellerine üç aşamalı bir komut sistemi uygulanmıştır. Bu sistem, modellerin yalnızca ilk yanıtlarını değil, yeniden sorgulama ve derinleştirilmiş analiz aşamalarındaki tutarlılıklarını da değerlendirmek amacıyla oluşturulmuştur.

Birinci aşamada modele ilgili iddia yöneltmiş ve doğrulama platformlarının kararlarına başvurmadan değerlendirme yapması istenmiştir. Bu aşamada kullanılan temel komut şu şekildedir:

- “Bu iddia doğru mu? Teyit ve doğrulama platformlarını dışarıda tutarak yanıt ver.”

Modellerin doğrulama platformlarının yayımladığı sonuçlara doğrudan dayanmasını azaltmak amacıyla ilk istemde bu platformların değerlendirme dışında tutulması açıkça istenmiştir. Bununla birlikte, kullanılan modellerin kapalı ve dinamik sistemler olması nedeniyle, araştırma öncesinde yayımlanmış doğrulama içeriklerinin eğitim verisi, indeksleme veya çevrimiçi erişim yoluyla modeller tarafından daha önce görülmüş olma olasılığı bütünüyle dışlanamamaktadır.

İkinci aşamada modelin ilk yanıtını yeniden değerlendirmesi istenmiştir. Bu aşamada kullanılan komut şu şekildedir:

- “Emin misin?”

Üçüncü aşamada ise modelden iddiayı daha ayrıntılı biçimde değerlendirmesi ve gerekçelendirmesi istenmiştir. Bu aşamada kullanılan komut, ilgili modelin olanaklarına göre derinleştirilmiş analiz veya daha ayrıntılı düşünme biçiminde uygulanmıştır. Bu aşama, modellerin ilk yanıtlarından farklılaşıp farklılaşmadığını ve daha ayrıntılı sorgulama sonucunda doğrulama platformlarının kararlarına yaklaşıp yaklaşmadığını incelemek amacıyla kullanılmıştır.

Üç aşamalı komut sistemi, yapay zekâ modellerinin doğrulama performansının tek bir yanıt üzerinden değerlendirilmesinin sınırlı kalabileceği varsayımına dayanmaktadır. Çünkü büyük dil modelleri istem biçimine, yeniden sorgulamaya ve analiz derinliğine duyarlı yanıtlar üretebilmektedir. Bu nedenle araştırmada ilk yanıt, yeniden sorgulama ve derinleştirilmiş analiz aşamaları ayrı ayrı kodlanmış ve karşılaştırılmıştır.

Kodlama Kategorileri ve Veri Matrisi

Araştırmada hem doğrulama platformlarının kararları hem de yapay zekâ modellerinin yanıtları karşılaştırılabilir kategorilere dönüştürülmüştür. Veri matrisinde her içerik için doğrulama platformu, içerik türü, platform kararı, yapay zekâ modeli, komut aşaması, model yanıtı ve platform kararıyla uyum durumu ayrı sütunlar hâlinde kaydedilmiştir.

Yapay zekâ modellerinin yanıtları üç temel kategoriye göre kodlanmıştır:

1. **İddia doğru:** Modelin, ilgili iddianın doğru olduğunu açık biçimde belirtmesi ya da platform kararındaki doğruluk yönüyle uyumlu bir değerlendirme yapması.
2. **İddia yanlış:** Modelin, ilgili iddianın yanlış, yanıltıcı, hatalı ya da gerçeği yansıtmadığını belirtmesi.
3. **İddia tartışmalı/belirsiz:** Modelin kesin bir doğru ya da yanlış yargısına ulaşmaması; iddiayı bağlama bağlı, kısmen doğru, kanıt yetersiz, doğrulanamıyor ya da tartışmalı olarak değerlendirmesi.

Kodlama sürecinde model yanıtlarının doğrulama platformlarının nihai kararlarıyla uyumlu olup olmadığı ayrıca belirlenmiştir. Platform kararıyla aynı yönde sonuç üreten model yanıtları “uyumlu”, farklı yönde sonuç üreten yanıtlar “uyumsuz”, kesin bir sonuca ulaşmayan veya ara kategoriye karşılık gelen yanıtlar ise bağlama göre “tartışmalı/belirsiz” olarak kodlanmıştır. Böylece modellerin yalnızca doğru ya da yanlış karar üretme durumları değil, belirsizlik ifade etme biçimleri de analiz kapsamına alınmıştır.

Araştırmada kullanılan temel kodlama kategorileri Tablo 1’de gösterilmiştir. Bu kategoriler doğrultusunda her içerik için doğrulama platformu kararı, yapay zekâ modeli, komut aşaması, model yanıtı ve platform kararıyla uyum durumu ayrı ayrı kodlanmıştır.

Tablo 1. Temel kodlama kategorileri

Kategori	Açıklama
1.İçerik başlığı	İncelenen doğrulama içeriğinin başlığı
2.İçerik türü	Haber metni, fotoğraf-görsel veya video
3.Doğrulama platformu	Teyit.org, Doğruluk Payı veya Malûmatfûruş
4.Platform kararı	Doğrulama platformunun ilgili içeriğe ilişkin nihai kararı
5.Yapay zekâ modeli	ChatGPT, Gemini, Grok veya DeepSeek
6.Komut aşaması	İlk yanıt, yeniden sorgulama veya derinleştirilmiş analiz
7.Model yanıtı	Yapay zekâ modelinin ilgili iddiaya verdiği yanıt
8.Uyum durumu	Model yanıtının platform kararıyla uyumlu olup olmadığı
9.Model yanıt kategorisi	“İddia doğru”, “iddia yanlış” veya “iddia tartışmalı/belirsiz”

Verilerin Analizi

Verilerin analizinde nicel ve nitel değerlendirme birlikte kullanılmıştır. Nicel analiz kapsamında, yapay zekâ modellerinin yanıtlarının doğrulama platformlarının kararlarıyla uyum düzeyleri frekans ve yüzde değerleri üzerinden hesaplanmıştır. Uyum oranları; model türü, içerik türü, doğrulama platformu ve komut aşaması değişkenlerine göre ayrı ayrı değerlendirilmiştir. Geçerli çıktı alınamayan durumlar uyum oranlarının hesaplanmasında paydadadan çıkarılmış; analizler yalnızca geçerli ve değerlendirilebilir model çıktıları üzerinden yürütülmüştür.

Nitel değerlendirme kapsamında ise model yanıtlarının gerekçelendirme biçimleri, belirsizlik ifadeleri, kaynak kullanımına ilişkin açıklamaları, bağlamsal yorumları ve komut aşamalarına göre değişen yanıt örüntüleri incelenmiştir. Özellikle fotoğraf-görsel ve video içeriklerde modellerin bağlam kurma, teknik erişim, zaman-mekân ilişkisini değerlendirme ve kesin ya da belirsiz yanıt üretme biçimleri yorumlanmıştır. Bu değerlendirme, nicel uyum oranlarının hangi içerik türlerinde ve hangi koşullar altında değiştiğini daha ayrıntılı biçimde açıklamak amacıyla yapılmıştır.

Kodlama güvenilirliğini değerlendirmek amacıyla içerikler farklı kodlayıcılar tarafından bağımsız biçimde yeniden kodlanmıştır. Ortak kodlanan ve karşılaştırılabilir nitelikteki model yanıtları üzerinden yapılan değerlendirmede kodlayıcılar arası doğrudan uyum oranı %83,3 olarak hesaplanmıştır. Pooled Cohen’s Kappa değeri 0,50; Krippendorff’s Alpha değeri ise 0,49 olarak bulunmuştur. Bu değerler, kodlayıcılar arasında orta düzeyde bir uyuma işaret etmektedir.

Kodlayıcılar arasındaki farklılıkların özellikle “iddia yanlış” ile “iddia tartışmalı/belirsiz” kategorileri arasındaki ayrım da yoğunlaştığı görülmüştür. Bu durum, yapay zekâ modellerinin bazı yanıtlarında kesin bir doğru ya da yanlış yargısı üretmemesi; “kanıt yetersiz”, “doğrulanamıyor”, “bağlama bağlı” veya “kısmen doğru/kısmen yanlış” gibi ara ifadeler kullanmasıyla ilişkilidir. Bu nedenle kodlama sürecinde kategori tanımlarının açık biçimde uygulanmasına ve belirsizlik içeren yanıtların “iddia tartışmalı/belirsiz” kategorisi altında değerlendirilmesine dikkat edilmiştir.

BULGULAR

Bulgular; veri setinin genel yapısı, içerik türlerine göre model performansı, yapay zekâ modellerine göre karşılaştırmalı sonuçlar, komut aşamalarına göre yanıt değişimi ve doğrulama platformlarına göre uyum düzeyleri temelinde değerlendirilmiştir. Araştırmada

doğrulama platformlarının nihai kararları referans ölçüt olarak alınmış; yapay zekâ modellerinin yanıtları bu kararlarla karşılaştırılmıştır.

İçerik Türlerine Göre Model Performansı

Öncelikle yapay zekâ modellerinin, doğrulama platformlarının kararlarıyla uyum düzeyinin içerik türlerine göre farklılaştığı belirlenmiştir. Tablo 2'de görüldüğü üzere, geçerli model çıktıları üzerinden hesaplanan genel uyum oranı %81,6'dır. Hücrelerdeki değerler uyumlu çıktı sayısı / geçerli çıktı sayısı ve yüzde oranını göstermektedir. Buna göre içerik türleri açısından en yüksek uyum video içeriklerde (%87,2), ardından fotoğraf-görsel içeriklerde (%82,8) ve haber metni içeriklerinde (%75,6) görülmüştür.

Tablo 2. İçerik türlerine göre yapay zekâ modellerinin doğrulama platformlarıyla uyum düzeyleri

İçerik Türü	ChatGPT Uyum f / %	Gemini Uyum f / %	Grok Uyum f / %	DeepSeek Uy. f / %	Genel Uyum f / %
Haber metni	38/45 %84,4	34/45 %75,6	32/45 %71,1	32/45 %71,1	136/180 %75,6
Fotoğraf- Görsel	33/45 %73,3	37/45 %82,2	39/45 %86,7	40/45 %88,9	149/180 %82,8
Video	29/39 %74,4*	43/45 %95,6	42/45 %93,3	22/27 %81,5*	136/156 %87,2
TOPLAM	100/129 %77,5	114/135 %84,4	113/135 %83,7	94/117 %80,3	421/516 %81,6

* ChatGPT ve DeepSeek'in bazı video içeriklerinde geçerli/değerlendirilebilir çıktı üretememesi nedeniyle video kategorisindeki paydalar sırasıyla 39 ve 27'dir. ChatGPT'de 6, DeepSeek'te 18 olmak üzere toplam 24 çıktı değerlendirme dışı kalmıştır. Yüzdelere yalnızca geçerli/değerlendirilebilir çıktılar üzerinden hesaplanmıştır.

Haber metni içerikli iddialar

Haber metni içerikli iddialarda yapay zekâ modellerinin doğrulama platformlarının kararlarıyla genel uyum oranı %75,6 olarak hesaplanmıştır. Bu içerik türünde en yüksek uyum ChatGPT'de (%84,4) görülürken, Gemini %75,6, Grok %71,1 ve DeepSeek %71,1 oranında platform kararlarıyla uyumlu sonuç üretmiştir. Haber metni içeriklerinde iddianın dilsel olarak açık biçimde sunulması, modellerin değerlendirme yapmasını kolaylaştırmakla birlikte, bazı iddialarda güncellik, bağlam ve kaynak doğrulama ihtiyacı model performansını sınırlamıştır.

Bu bulgu, haber metni temelli iddiaların yapay zekâ modelleri açısından her zaman en kolay doğrulanabilir içerik türü olmadığını göstermektedir. Özellikle güncel gelişmelere, sayısal verilere, siyasal açıklamalara veya bağlamsal ayrıntılara dayalı iddialarda modellerin platform kararlarıyla uyumu değişebilmektedir. Dolayısıyla metinsel açıklık, model yanıtlarının platform kararlarıyla uyumunu artıran bir unsur olmakla birlikte, tek başına yeterli bulunmamaktadır.

Fotoğraf-görsel içerikli iddialar

Fotoğraf-görsel içerikli iddialarda genel uyum oranı %82,8 olarak hesaplanmıştır. Bu kategoride DeepSeek %88,9 ile en yüksek uyum oranına ulaşırken, Grok %86,7, Gemini %82,2 ve ChatGPT %73,3 oranında platform kararlarıyla uyumlu sonuç üretmiştir. Bu sonuç, fotoğraf-görsel içeriklerde model performansının haber metni içeriklerine göre daha yüksek gerçekleştiğini göstermektedir.

Bununla birlikte, fotoğraf-görsel içeriklerdeki yüksek uyum oranı, modellerin görsel doğrulama süreçlerinde tüm teknik aşamaları eksiksiz biçimde yerine getirdiği anlamına gelmemektedir. Görsel içeriklerin doğrulanması; görselin kaynağı, tarihi, konumu, dolaşım bağlamı ve ilgili iddiayla ilişkisi gibi unsurların birlikte değerlendirilmesini gerektirir. Bu nedenle modellerin

platform kararlarıyla uyumlu sonuçlar üretmesi önemli olmakla birlikte, bu sonuçlar insan uzmanlığına dayalı görsel doğrulama süreçlerinin yerini alabilecek kesinlikle değerlendirilmemelidir.

Video içerikli iddialar

Video içerikli iddialarda geçerli model çıktıları üzerinden hesaplanan genel uyum oranı %87,2'dir. Bu kategoride Gemini %95,6 ve Grok %93,3 ile en yüksek uyum oranlarına ulaşmıştır. DeepSeek'in video içeriklerdeki uyum oranı %81,5, ChatGPT'nin uyum oranı ise %74,4'tür. Ancak bu kategoride ChatGPT ve DeepSeek'in bazı video içeriklerde geçerli yanıt üretmediği görülmektedir. Bu nedenle video içeriklerdeki yüksek uyum oranı, tüm video içerikler için genel bir teknik yeterlilik göstergesi olarak değil, değerlendirilebilir yanıtlar içindeki uyum düzeyi olarak yorumlanmalıdır.

Video içerikler, doğrulama açısından görüntü, ses, zaman-mekân ilişkisi, olay örgüsü ve bağlamın birlikte değerlendirilmesini gerektiren çok katmanlı içeriklerdir. Bu çalışmada elde edilen bulgular, modellerin geçerli yanıt üretebildikleri video içeriklerde doğrulama platformlarının kararlarıyla yüksek oranda örtüşebildiğini; ancak teknik erişim ve içerik işleme sınırlılıklarının video kategorisinde önemli bir değişken olarak dikkate alınması gerektiğini göstermektedir.

Yapay Zekâ Modellerine Göre Karşılaştırmalı Bulgular

Yapay zekâ modelleri arasında doğrulama platformlarının kararlarıyla uyum düzeyi bakımından farklılıklar görülmüştür. Tablo 3'te yer alan bulgulara göre, geçerli model çıktıları üzerinden hesaplanan genel uyum oranı %81,6'dır. Model bazında en yüksek uyum oranı Gemini'de (%84,4) görülürken, onu Grok (%83,7), DeepSeek (%80,3) ve ChatGPT (%77,5) izlemiştir. Bu sonuçlar, modellerin içerik doğrulama performansının birbirine yakın olmakla birlikte aynı düzeyde olmadığını göstermektedir.

Tablo 3. Yapay zekâ modellerine göre genel uyum durumu

Model	Geçerli Model Çıktısı	Uyumlu Çıktı	Uyum Oranı (%)
ChatGPT	129	100	77,5
Gemini	135	114	84,4
Grok	135	113	83,7
DeepSeek	117	94	80,3
TOPLAM	516	421	81,6

İncelenen içeriklerde ChatGPT'nin genel uyum oranı %77,5'tir. Model, özellikle haber metni içerikli iddialarda %84,4 ile kendi içinde en yüksek uyum oranına ulaşmıştır. Buna karşılık fotoğraf-görsel içeriklerde %73,3, video içeriklerde ise %74,4 oranında platform kararlarıyla uyumlu sonuç üretmiştir. Ayrıca video içeriklerde geçerli çıktı sayısının 39'da kalması, ChatGPT'nin bazı video içeriklerde teknik erişim veya değerlendirme sınırlılığı yaşadığını gösterir nitelikte değerlendirilmiştir.

Gemini, %84,4 genel uyum oranıyla en yüksek performans gösteren model olmuştur. Modelin haber metni içeriklerindeki uyum oranı %75,6 iken, fotoğraf-görsel içeriklerde %82,2'ye, video içeriklerde ise %95,6'ya yükselmiştir. Bu bulgu, Gemini'nin özellikle video içeriklerde geçerli yanıt üretebildiği durumlarda doğrulama platformlarının kararlarıyla yüksek düzeyde örtüşüğünü göstermektedir.

Grok'un genel uyum oranı %83,7'dir. Model, haber metni içerikli iddialarda %71,1 ile daha düşük bir uyum sergilerken, fotoğraf-görsel içeriklerde %86,7 ve video içeriklerde %93,3 oranında platform kararlarıyla uyumlu sonuç üretmiştir. Bu sonuç, Grok'un özellikle çoklu medya içeriklerinde geçerli çıktı ürettiği durumlarda platform kararlarına daha fazla yaklaştığını göstermektedir.

DeepSeek'in genel uyum oranı %80,3'tür. Model, fotoğraf-görsel içeriklerde %88,9 ile en yüksek uyum oranına ulaşmış; video içeriklerde %81,5, haber metni içeriklerinde ise %71,1 oranında platform kararlarıyla uyumlu sonuç üretmiştir. Bununla birlikte DeepSeek'in toplam geçerli çıktı sayısının 117'de kalması, özellikle bazı video içeriklerde teknik erişim veya içerik işleme sınırlılıklarının bulunduğunu göstermektedir.

Genel olarak değerlendirildiğinde, modeller arasındaki farkların yalnızca genel doğruluk düzeyiyle açıklanamayacağı görülmektedir. Modellerin performansı içerik türüne, geçerli çıktı üretebilme kapasitesine, görsel-video içeriklere erişim durumuna ve komut aşamalarına göre değişebilmektedir. Bu nedenle model karşılaştırmalarında yalnızca uyum oranlarının değil, her modelin hangi içerik türünde kaç geçerli yanıt üretebildiğinin de dikkate alınması gerektiği belirtilebilir.

Komut Aşamalarına Göre Yanıt Değişimi

Araştırmanın her komut aşamasında 180 olası model çıktısı bulunmasına karşın, ChatGPT ve DeepSeek'in bazı video bağlantılarını veya içeriklerini tanıyamaması ya da işleyememesi nedeniyle her aşamada 172 geçerli çıktı analiz edilmiştir.

Araştırmada kullanılan üç aşamalı komut sistemi, yapay zekâ modellerinin doğrulama platformlarının kararlarıyla uyum düzeyinin komut aşamalarına göre sınırlı ölçüde değiştiğini göstermektedir. Tablo 4'te görüldüğü üzere, ilk komut aşamasında genel uyum oranı %80,8 iken, ikinci komutta %81,4'e, derinleştirilmiş analiz aşamasında ise %82,6'ya yükselmiştir. Bu bulgu, yeniden sorgulama ve daha ayrıntılı analiz istemlerinin model yanıtlarında sınırlı ama düzenli bir iyileşme sağladığını göstermektedir.

Tablo 4. Komut aşamalarına göre yapay zekâ modeli-platform uyum oranları

Komut Aşamaları	ChatGPT f / %	Gemini f / %	Grok f / %	DeepSeek f / %	Genel Uyum f / %
1. Komut	33/43 %76,7	37/45 %82,2	38/45 %84,4	31/39 %79,5	139/172 %80,8
2. Komut: "Emin misin?"	33/43 %76,7	37/45 %82,2	38/45 %84,4	32/39 %82,1	140/172 %81,4
3. Komut: Derinleştirilmiş analiz	34/43 %79,1	40/45 %88,9	37/45 %82,2	31/39 %79,5	142/172 %82,6

Model bazında incelendiğinde, komut aşamalarının etkisinin her modelde aynı yönde ve aynı düzeyde gerçekleşmediği anlaşılmaktadır. ChatGPT'de uyum oranı ilk iki komutta %76,7 düzeyinde kalmış, derinleştirilmiş analiz aşamasında %79,1'e yükselmiştir. Gemini'de ilk iki komutta %82,2 olan uyum oranı, üçüncü komutta %88,9'a çıkmıştır. Bu sonuç, derinleştirilmiş analiz komutunun özellikle Gemini'de platform kararlarıyla uyumu artırdığını göstermektedir.

Grok'ta ise ilk iki komutta %84,4 olan uyum oranı, üçüncü komutta %82,2'ye gerilemiştir. Bu durum, daha ayrıntılı analiz isteminin her zaman daha yüksek uyum üretmediğini göstermesi bakımından önemlidir. DeepSeek'te ikinci komutta %82,1'e yükselen uyum oranı, üçüncü komutta yeniden %79,5'e düşmüştür. Dolayısıyla yeniden sorgulama ve derinleştirilmiş analiz

aşamaları bazı modellerde uyumu artırırken, bazı modellerde yanıtların daha temkinli, farklı ya da platform kararından uzak biçimde yeniden kurulmasına yol açabilmektedir.

Genel olarak değerlendirildiğinde, komut aşamaları model performansını etkileyen önemli bir yöntemsel değişken olarak ortaya çıkmıştır. Ancak bu etkinin doğrusal ve tüm modeller için aynı yönde olmadığı anlaşılmaktadır. Derinleştirilmiş analiz aşaması genel uyum oranını artırmış olsa da bu artış sınırlıdır. Bu nedenle yapay zekâ modellerinin içerik doğrulama performansı yalnızca ilk yanıt üzerinden değerlendirilmemeli; yeniden sorgulama ve ayrıntılı analiz aşamalarındaki yanıt değişimleri de dikkate alınmalıdır.

Doğrulama Platformlarına Göre Bulgular

Yapay zekâ modellerinin doğrulama platformlarının kararlarıyla uyum düzeyi, platformlara göre belirgin biçimde farklılaşmaktadır. Araştırma kapsamında ChatGPT'nin Doğruluk Payı'nda, DeepSeek'in ise Doğruluk Payı ve Malûmatfûrûş platformlarındaki bazı video içeriklerinde geçerli çıktı üretememesi nedeniyle bu modellerin ilgili platformlardaki paydaları 45'in altında kalmıştır. Teyit.org içeriklerinde tüm modeller için geçerli çıktı alınmıştır. Buna göre Tablo 5'te görüldüğü üzere, en yüksek genel uyum oranı Malûmatfûrûş içeriklerinde (%91,2) görülmüştür. Bunu Doğruluk Payı (%83,6) izlemekte; en düşük uyum oranı ise Teyit.org içeriklerinde (%70,6) ortaya çıkmaktadır. Bu sonuç, yapay zekâ modellerinin performansının yalnızca model özelliklerine değil, doğrulanan içeriklerin yapısına, platformların konu önceliklerine ve doğrulama kararlarının niteliğine göre de değiştiğini göstermektedir.

Tablo 5. Platformlara göre yapay zekâ modeli-platform uyum oranları

Platform	ChatGPT f / %	Gemini f / %	Grok f / %	DeepSeek f / %	Genel Uyum f / %
Doğruluk Payı	35/39 %89,7	40/45 %88,9	33/45 %73,3	30/36 %83,3	138/165 %83,6
Teyit.org	23/45 %51,1	35/45 %77,8	38/45 %84,4	31/45 %68,9	127/180 %70,6
Malûmatfûrûş	42/45 %93,3	39/45 %86,7	42/45 %93,3	33/36 %91,7	156/171 %91,2
TOPLAM	100/129 %77,5	114/135 %84,4	113/135 %83,7	94/117 %80,3	421/516 %81,6

Doğruluk Payı içeriklerinde genel uyum oranı %83,6 olarak hesaplanmıştır. Bu platformda ChatGPT %89,7, Gemini %88,9 ve DeepSeek %83,3 oranında platform kararlarıyla uyumlu sonuç üretmiştir. Grok'un uyum oranı ise %73,3 ile diğer modellere göre daha düşük kalmıştır. Doğruluk Payı içeriklerinin önemli ölçüde siyasal, kamusal, istatistiksel ya da kaynak karşılaştırmasına dayalı iddialar içermesi, modellerin metinsel ve veriye dayalı değerlendirme kapasitesini öne çıkarmaktadır. Bununla birlikte Grok örneğinde görüldüğü gibi, bu tür içeriklerde dahi model performansı aynı düzeyde gerçekleşmemektedir.

Teyit.org içeriklerinde genel uyum oranı %70,6 ile diğer platformlara göre daha düşüktür. Bu platformda en yüksek uyum Grok'ta (%84,4), ardından Gemini'de (%77,8) görülmüştür. DeepSeek'in uyum oranı %68,9, ChatGPT'nin uyum oranı ise %51,1'dir. Teyit.org içeriklerinde uyum oranının daha düşük gerçekleşmesi, bu platformda sosyal medya kaynaklı şüpheli iddialar, bağlamından koparılmış görsel-video içerikler ve dolaşım bağlamı güçlü olan örneklerin daha fazla yer almasıyla ilişkili olabilir. Bu tür içerikler, yalnızca iddia metninin değerlendirilmesini değil; görsel bağlam, paylaşım geçmişi, zaman-mekân ilişkisi ve açık kaynak doğrulama süreçlerini de gerektirmektedir.

Malûmatfûruş içerikleri, %91,2 ile en yüksek genel uyum oranına sahiptir. Bu platformda ChatGPT ve Grok %93,3, DeepSeek %91,7, Gemini ise %86,7 oranında platform kararlarıyla uyumlu sonuç üretmiştir. Malûmatfûruş içeriklerinde uyum oranının yüksek olması, bu platformda ele alınan iddiaların çoğunlukla açık kaynaklarla, tarihsel bilgilerle, metinsel karşılaştırmalarla veya somut yanıtlama verileriyle değerlendirilebilir nitelikte olmasıyla ilişkili görülebilir. Bu tür iddialar, büyük dil modellerinin bilgi ilişkilendirme ve açıklama üretme kapasitesiyle daha fazla örtüşmektedir.

Genel olarak platformlara göre elde edilen bulgular, yapay zekâ modellerinin doğrulama performansının doğrulama platformlarının içerik profiline göre değiştiğini göstermektedir. Bu nedenle model performansı değerlendirilirken yalnızca genel uyum oranları değil, platformların incelediği içeriklerin türü ve doğrulama kararlarının dayandığı bağlamsal unsurlar da dikkate alınmalıdır denilebilir.

TARTIŞMA VE SONUÇ

Bu çalışma, büyük dil modellerine dayalı yapay zekâ sistemlerinin içerik doğrulama süreçlerindeki performansını Türkiye'deki bağımsız doğrulama platformlarının kararlarıyla karşılaştırmalı olarak incelemiştir. Araştırma kapsamında ChatGPT, Gemini, Grok ve DeepSeek modellerinin haber metni, fotoğraf-görsel ve video içeriklerinde ürettiği doğrulama sonuçları; Teyit.org, Doğruluk Payı ve Malûmatfûruş platformlarının yayımladığı kararlarla karşılaştırılmıştır. Elde edilen bulgular, yapay zekâ modellerinin içerik doğrulama süreçlerinde belirli bir destek kapasitesine sahip olduğunu, ancak bu kapasitenin içerik türüne, doğrulama platformuna, komut aşamasına, teknik erişim olanaklarına ve iddianın bağlamsal yapısına göre değiştiğini ortaya koymaktadır.

Araştırmada geçerli model çıktıları üzerinden hesaplanan genel uyum oranı %81,6 olarak hesaplanmıştır. Bu oran, yapay zekâ modellerinin önemli bir bölümde doğrulama platformlarının kararlarıyla örtüşebildiğini göstermektedir. Bununla birlikte bu bulgu, modellerin içerik doğrulama süreçlerinde tek başına yeterli ve bağımsız doğruluk otoriteleri olarak değerlendirilebileceği anlamına gelmemektedir. Çünkü içerik doğrulama yalnızca bir iddianın doğru ya da yanlış olarak sınıflandırılması değil; kaynağın, bağlamın, kanıt yapısının, dolaşım biçiminin ve içeriğin toplumsal etkisinin birlikte değerlendirilmesini gerektiren çok katmanlı bir süreçtir.

İçerik türleri açısından bakıldığında, en yüksek uyum oranı video içeriklerde (%87,2), ardından fotoğraf-görsel içeriklerde (%82,8) ve haber metni içeriklerinde (%75,6) görülmüştür. İlk bakışta video içeriklerdeki yüksek uyum oranı, modellerin video doğrulamada daha başarılı olduğu (daha yüksek uyum oranına ulaştığı) izlenimi yaratabilir. Ancak bu sonuç dikkatli yorumlanmalıdır. Çünkü bazı video içeriklerde özellikle ChatGPT ve DeepSeek geçerli yanıt üretememiş; bu nedenle video kategorisindeki uyum oranı yalnızca değerlendirilebilir yanıtlar üzerinden hesaplanmıştır. Dolayısıyla video içeriklerdeki yüksek uyum, tüm video içerikler için genel bir teknik yeterlilik göstergesi olarak değil, modellerin geçerli yanıt üretebildiği durumlarda platform kararlarıyla örtüşme düzeyi olarak değerlendirilmelidir.

Bulgular, yapay zekâ modellerinin içerik doğrulama performansının yalnızca doğru karar üretme oranıyla ölçülemeyeceğini göstermektedir. Bir modelin belirli bir içerik türünde yüksek uyum oranı üretmesi kadar, o içerik türünde geçerli ve değerlendirilebilir yanıt üretip üretmediği de önemlidir. Özellikle video içerikler; görüntü, ses, zaman-mekân ilişkisi, olay örgüsü ve bağlamın birlikte değerlendirilmesini gerektirdiğinden, yapay zekâ modelleri açısından teknik erişim ve çoklu medya işleme kapasitesi belirleyici hâle gelmektedir.

Model bazında en yüksek genel uyum oranı Gemini’de (%84,4) görülmüş; onu Grok (%83,7), DeepSeek (%80,3) ve ChatGPT (%77,5) izlemiştir. Modeller arasındaki farklar çok büyük olmamakla birlikte, içerik türleri bakımından dikkat çekici ayrışmalar bulunmaktadır. ChatGPT haber metni içeriklerinde daha yüksek uyum gösterirken, Gemini ve Grok video içeriklerde, DeepSeek ise fotoğraf-görsel içeriklerde daha yüksek uyum oranlarına ulaşmıştır. Bu bulgu, yapay zekâ modellerinin doğrulama performansının tek bir genel başarı oranıyla açıklanamayacağını; modelin teknik olanakları, erişim kapasitesi, içerik türü ve komut aşaması gibi değişkenlerle birlikte değerlendirilmesi gerektiğini göstermektedir.

Komut aşamalarına göre elde edilen bulgular, yapay zekâ destekli doğrulamada istem tasarımının önemini ortaya koymaktadır. İlk komutta %80,8 olan genel uyum oranı, ikinci komutta %81,4’e, derinleştirilmiş analiz aşamasında ise %82,6’ya yükselmiştir. Bu artış sınırlı olmakla birlikte düzenlidir. Dolayısıyla yeniden sorgulama ve daha ayrıntılı analiz istemleri, model yanıtlarını belirli ölçüde iyileştirebilmekte; ancak bu etki her modelde aynı yönde ve aynı düzeyde gerçekleşmemektedir. Bazı modellerde derinleştirilmiş analiz aşaması uyumu artırırken, bazı modellerde daha temkinli, belirsiz veya platform kararından uzak yanıtların ortaya çıkmasına neden olabilmektedir. Bu nedenle komut tasarımı, yapay zekâ destekli doğrulama araştırmalarında yöntemsel bir değişken olarak dikkate alınmalıdır.

Doğrulama platformlarına göre elde edilen bulgular, model performansının platformların içerik profiline göre de değiştiğini göstermektedir. En yüksek uyum Malûmatfûrûş içeriklerinde (%91,2), ardından Doğruluk Payı içeriklerinde (%83,6) görülürken, Teyit.org içeriklerinde uyum oranı %70,6’da kalmıştır. Bu farklılık, platformların ele aldığı iddia türleri, içeriklerin bağlamsal yapısı ve doğrulama kararlarının dayandığı kanıt biçimleriyle ilişkili olabilir. Özellikle sosyal medya kaynaklı, görsel-video bağlamı güçlü ve dolaşım geçmişi önemli olan içeriklerde model-platform uyumunun daha düşük gerçekleşmesi, yapay zekâ modellerinin bağlam ve dolaşım bilgisi gerektiren içeriklerde daha sınırlı kalabildiğine işaret etmektedir.

Genel olarak değerlendirildiğinde, araştırmanın temel sonucu yapay zekâ modellerinin içerik doğrulama süreçlerinde önemli bir yardımcı kapasiteye sahip olmakla birlikte, tek başına yeterli doğruluk otoriteleri olarak değerlendirilemeyeceğidir. Büyük dil modelleri; hızlı ön değerlendirme, iddia sınıflandırma, olası tutarsızlıkları işaretleme ve bağlamsal açıklama üretme bakımından yararlı olabilir. Ancak nihai doğrulama kararı, kanıta dayalı araştırma, kaynak denetimi, görsel-video doğrulama teknikleri ve editoryal muhakeme gerektirir. Bu nedenle yapay zekâ sistemleri, profesyonel doğrulama platformlarının ve insan uzmanlığının yerine geçen bağımsız karar mekanizmaları olarak değil, insan denetimiyle birlikte çalışan destekleyici araçlar olarak değerlendirilmelidir.

Bu bağlamda çalışma, yapay zekâ destekli içerik doğrulama süreçleri için en uygun yaklaşımın insan uzmanlığı ile yapay zekâ sistemlerini birlikte kullanan hibrit modeller olduğunu göstermektedir. Böyle bir modelde yapay zekâ, doğrulanacak iddiaların ön sınıflandırılması, olası kaynak alanlarının belirlenmesi, tutarsızlıkların işaretlenmesi ve ilk değerlendirme sürecinin hızlandırılması gibi aşamalarda işlevsel olabilir. Buna karşılık nihai karar, kanıtların değerlendirilmesi, bağlamın kurulması, görsel-video analizinin yapılması ve editoryal sorumluluğun üstlenilmesi aşamalarında insan uzmanlığına dayanmalıdır.

Sınırlılıklar ve Gelecek Araştırmalar İçin Öneriler

Çalışma bulgularının sınırlılıklar çerçevesinde değerlendirilmesi de önemlidir. İlk olarak araştırma, 1-30 Kasım 2025 tarihleri arasında Teyit.org, Doğruluk Payı ve Malûmatfûrûş platformlarında yayımlanan 45 doğrulama içeriğiyle sınırlıdır. Bu nedenle bulgular, Türkiye’deki tüm doğrulama platformlarını, tüm zaman dönemlerini veya tüm içerik türlerini temsil etme

iddiası taşımamaktadır. Gelecek araştırmalarda daha geniş zaman aralıkları, daha fazla doğrulama platformu ve daha büyük veri setleri üzerinden benzer karşılaştırmalar yapılabilir.

İkinci olarak çalışma, dört büyük dil modeliyle sınırlandırılmıştır: ChatGPT, Gemini, Grok ve DeepSeek. Yapay zekâ modelleri sürekli güncellenen, teknik özellikleri ve performans düzeyleri zaman içinde değişebilen sistemlerdir. Bu nedenle çalışmada elde edilen bulgular, modellerin mutlak ve kalıcı doğruluk düzeyini değil, araştırmanın gerçekleştirildiği tarih aralığında, kullanılan model sürümleri ve sorgulama koşulları altında ortaya çıkan performanslarını göstermektedir. Gelecek çalışmalarda model sürümü, sorgulama tarihi, web erişimi, çoklu ortam işleme kapasitesi ve komut protokolü gibi değişkenlerin ayrıntılı biçimde raporlanması önem taşımaktadır.

Üçüncü olarak, bazı video içeriklerde ilgili video bağlantılarının veya içeriklerin ChatGPT ve DeepSeek modelleri tarafından tanınamaması ya da işlenememesi nedeniyle geçerli çıktı alınamamıştır. Bu nedenle video içeriklerde hesaplanan yüksek uyum oranı, tüm video içerikler için genel bir teknik yeterlilik göstergesi olarak değil, yalnızca değerlendirilebilir çıktılar içindeki uyum düzeyi olarak yorumlanmalıdır. Gelecek araştırmalarda video içerikler için standartlaştırılmış ekran görüntüsü, transkript, bağlantı, açıklama ve ham dosya sunum biçimlerinin model performansına etkisi ayrıca incelenebilir.

Dördüncü olarak kodlayıcılar arası doğrudan uyum oranı yüksek olmakla birlikte, Kappa ve Alpha değerlerinin orta düzeyde kalması, yapay zekâ modellerinin belirsizlik içeren yanıtlarının sınıflandırılmasında yorumlayıcı farklılıkların ortaya çıkabildiğini göstermektedir. Özellikle “iddia yanlış” ile “iddia tartışmalı/belirsiz” kategorileri arasındaki ayrımın bazı durumlarda kodlayıcı yorumu gerektirdiği görülmüştür. Gelecek araştırmalarda daha ayrıntılı kodlama yönergeleri oluşturulması, kodlayıcı eğitiminin artırılması ve daha geniş bir kontrol kodlama örneklemleriyle güvenilirlik analizlerinin tekrarlanması önerilmektedir.

Beşinci olarak bu çalışma, doğrulama platformlarının nihai kararlarını referans ölçüt olarak kabul etmiştir. Bu tercih araştırmanın karşılaştırmalı tasarımı açısından gereklidir; ancak doğrulama platformlarının kararları da belirli editoryal yöntemlere, sınıflandırma sistemlerine ve kanıt değerlendirme süreçlerine dayanmaktadır. Gelecek araştırmalarda, yapay zekâ modellerinin yalnızca nihai kararlarla değil, doğrulama platformlarının gerekçelendirme biçimleri, kullandıkları kaynaklar ve kanıt türleriyle de karşılaştırılması yararlı olacaktır.

Son olarak, bu çalışma yapay zekâ modellerinin içerik doğrulama performansını çıktı düzeyinde incelemiştir. Gelecek araştırmalar, kullanıcıların yapay zekâ tarafından üretilen doğrulama yanıtlarına duyduğu güveni, bu yanıtların haber tüketimi ve paylaşım davranışları üzerindeki etkisini ve yapay zekâ destekli doğrulama araçlarının gazetecilik pratiklerine nasıl entegre edilebileceğini inceleyebilir. Böylece yapay zekâ destekli doğrulamanın yalnızca teknik doğruluk boyutu değil, kullanıcı, gazeteci ve platform ekosistemi açısından toplumsal etkileri de daha kapsamlı biçimde değerlendirilebilir.

KAYNAKÇA

- Aïmeur, E., Amri, S., & Brassard, G. (2023). *Fake news, disinformation and misinformation in social media: A review*. Social Network Analysis and Mining, 13(1), Article 30. <https://doi.org/10.1007/s13278-023-01028-5>
- Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2), 211–236. <https://doi.org/10.1257/jep.31.2.211>

- Bosley, M., Jacobs-Harukawa, M., Licht, H., & Hoyle, A. (2023). Do we still need BERT in the age of GPT? Comparing the benefits of domain-adaptation and in-context-learning approaches to using LLMs for political science research. *Annual Meeting of the Midwest Political Science Association*.
https://mbosley.github.io/papers/bosley_harukawa_licht_hoyle_mpsa2023.pdf
- Caramancion, K. M. (2023). News verifiers showdown: A comparative performance evaluation of ChatGPT 3.5, ChatGPT 4.0, Bing AI, and Bard in news fact-checking. *arXiv*.
<https://arxiv.org/abs/2306.17176>
- Deiana, G., Dettori, M., Arghittu, A., Azara, A., Gabutti, G., & Castiglia, P. (2023). Artificial intelligence and public health: Evaluating ChatGPT responses to vaccination myths and misconceptions. *Vaccines*, 11(7), Article 1217.
<https://doi.org/10.3390/vaccines11071217>
- Doğruluk Payı. (t.y.-a). Hakkımızda. *Doğruluk Payı*.
<https://www.dogrulukpayi.com/sayfa/hakkimizda>
- Doğruluk Payı. (t.y.-b). Değerlendirme kriterleri. *Doğruluk Payı*.
<https://www.dogrulukpayi.com/sayfa/metodoloji>
- Evren, O., & Türedi, U. (2022). Yeni medya içeriklerinde doğruluk kontrolü uygulamalarına genel bir bakış. *The Journal of Communication and Social Studies*, 2(2), 224–240.
- Graves, L. (2016). *Deciding what's true: The rise of political fact-checking in American journalism*. New York: Columbia University Press.
- Guo, Z., Schlichtkrull, M., & Vlachos, A. (2022). A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10, 178–206.
https://doi.org/10.1162/tacl_a_00454
- Hoes, E., Altay, S., & Bermeo, J. (2023). Leveraging ChatGPT for efficient fact-checking. *PsyArXiv*. <https://doi.org/10.31234/osf.io/qnjkf>
- Karasar, N. (2012). *Bilimsel araştırma yöntemi: Kavramlar, ilkeler, teknikler*. Ankara: Nobel Yayıncılık.
- Kavaklı, N. (2019). Yalan haberle mücadele ve internet teyit/doğrulama platformları. *Erciyes İletişim Dergisi*, 6(1), 663–682.
- Kuznetsova, E., Makhortykh, M., Vziatysheva, V., Stolze, M., Baghumyan, A., & Urman, A. (2025). In generative AI we trust: Can chatbots effectively verify political information? *Journal of Computational Social Science*, 8, Article 15. <https://doi.org/10.1007/s42001-024-00338-8>
- Malûmatfûruş. (t.y.-a). Hakkında. *Malûmatfûruş*. <https://www.malumatfurus.org/hakkinda/>
- Malûmatfûruş. (t.y.-b). Yöntem. *Malûmatfûruş*. <https://www.malumatfurus.org/yontem/>
- Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., & Mian, A. (2024). A comprehensive overview of large language models. *arXiv*.
<https://arxiv.org/abs/2307.06435>
- Neuendorf, K. A. (2002). *The content analysis guidebook*. Thousand Oaks: SAGE Publications.

- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Aspell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, 27730–27744.
- Patton, M. Q. (2015). *Qualitative research & evaluation methods: Integrating theory and practice* (4. baskı). Thousand Oaks: SAGE Publications.
- Sharma, K., Qian, F., Jiang, H., Ruchansky, N., Zhang, M., & Liu, Y. (2019). Combating fake news: A survey on identification and mitigation techniques. *ACM Transactions on Intelligent Systems and Technology*, 10(3), Article 21. <https://doi.org/10.1145/3305260>
- Teyit. (t.y.-a). Metodoloji. *Teyit*. <https://teyit.org/metodoloji>
- Teyit. (t.y.-b). Değerlerimiz. *Teyit*. <https://teyit.org/degerlerimiz>
- Trewinnard, T., & Bell, F. (2018). Social media verification: Assessing sources and visual content. İçinde C. Ireton & J. Posetti (Ed.), *Journalism, “fake news” & disinformation: Handbook for journalism education and training* (ss. 100–113). Paris: UNESCO Publishing.
- Urman, A., & Makhortykh, M. (2025). The silence of the LLMs: Cross-lingual analysis of guardrail-related political bias and false information prevalence in ChatGPT, Google Bard, and Bing Chat. *Telematics and Informatics*, 96, Article 102211. <https://doi.org/10.1016/j.tele.2024.102211>
- Wardle, C., & Derakhshan, H. (2017). *Information disorder: Toward an interdisciplinary framework for research and policy making*. Strasbourg: Council of Europe.
- Yalansavar. (t.y.-a). Yalansavar sık sorulan sorular. *Yalansavar*. <https://yalansavar.org/yalansavar-sik-sorulan-sorular/>
- Yalansavar. (t.y.-b). Neden? *Yalansavar*. <https://yalansavar.org/about/>