Black Sea Journal of Artificial Intelligence
doi: 10.5281/zenodo.18213660

Open Access Journal
e-ISSN: 3108-4281

Araştırma Makalesi (Research Article)
Cilt 1 - Sayı 1: 11-14 / Haziran 2025
(Volume 1 - Issue 1: 11-14 / June 2025)

PHISHING ATAKLARININ BÜYÜK DİL MODELLERİ ÜZERİNE ETKİLERİ

Hakan Can ALTUNAY[1]*

[1]*Ondokuz Mayıs University, Çarsamba Chamber of Commerce Vocational School, Department of Computer Technologies, 55200, Samsun, Türkiye*

Özet: Büyük dil modellerinin (LLM) yaygın olarak uygulanması, önemli akademik ve toplumsal ilgiyi çeken yeni güvenlik zorluklarını ve etik endişeleri artırmıştır. LLM'lerin güvenlik açıklarının ve siber suçlarda kötüye kullanımının analizi, gelişmiş metin oluşturma yeteneklerinin kişisel gizlilik, veri güvenliği ve bilgi bütünlüğü için ciddi tehditler oluşturduğunu ortaya koymaktadır. Ayrıca, mevcut LLM tabanlı savunma stratejilerinin etkinliği incelenmiş ve değerlendirilmiştir. Bu çalışma, LLM'lerin phishing ataklarına karşı sosyal etkilerini incelemekte ve alanın gelişimini bilgilendirmeyi amaçlayan güvenlik uygulamaları ile etik yönetişimlerini geliştirmek için gelecekteki uygulamaları önermektedir.

Anahtar kelimeler: LLM tabanlı tehdit azaltma, Kimlik avı atakları, Yapay zeka tabanlı kimlik avı

Effects of Phishing Attacks on Large Language Models

Abstract: The widespread application of large language models (LLMs) has raised new security challenges and ethical concerns that have attracted significant academic and societal attention. Analysis of LLMs' vulnerabilities and their misuse in cybercrime reveals that their advanced text generation capabilities pose serious threats to personal privacy, data security, and information integrity. Furthermore, the effectiveness of existing LLM-based defense strategies is examined and evaluated. This study examines the social implications of LLMs against phishing attacks and suggests future applications to enhance their security practices and ethical governance, aiming to inform the development of the field.

Keywords: LLM-based threat mitigation, Phishing attacks, AI-enabled phishing

***Sorumlu yazar (Corresponding author):** Ondokuz Mayıs University, Çarsamba Chamber of Commerce Vocational School, Department of Computer Technologies, 55200, Samsun, Türkiye
E mail: hakancan.altunay@omu.edu.tr (H. C. ALTUNAY)
Hakan Can ALTUNAY https://orcid.org/0000-0002-0175-239X

Gönderi: 08 Nisan 2025 **Received:** April 08, 2025
Kabul: 01 Mayıs 2025 **Accepted:** May 01, 2025
Yayınlanma: 15 Haziran 2025 **Published:** June 15, 2025
Cite as: Altunay, H. C. (2025). Effects of phishing attacks on large language models. *Black Sea Journal of Artificial Intelligence*, *1*(1), 11–14.

1. Giriş

Son yıllarda, büyük dil modelleri (LLM'ler) doğal dil işlemede önemli ölçüde ilerleme kaydetti. Önemli örnekler arasında üretken önceden eğitilmiş dönüştürücü (GPT), BERT (Dönüştürücülerden Çift Yönlü Kodlayıcı Gösterimleri) ve T5 (Metinden Metne Aktarım Dönüştürücüsü) yer almaktadır (Annepaka ve Pakray, 2025). Bu modeller, metin oluşturma, makine çevirisi ve soru-cevap sistemleri gibi bir dizi görevde etkileyici performans göstermektedir. Ancak, bunların artan yaygınlığı güvenlik ve etik endişeleri artırarak hem akademiden hem de endüstriden dikkat çekmektedir (Zheng vd., 2025).

LLM'lerin ve üretken yapay zekanın yükselişi birçok endüstriyi derinden etkilemiştir. Eğitim, sağlık ve siyaset gibi sektörlerde, LLM'ler geleneksel iş süreçlerini dönüştürmektedir. Dijital dönüşümü ve akıllı gelişmeyi yönlendirirken verimliliği artırıyor ve maliyetleri düşürmektedirler. Ancak, bu hızlı ilerleme zorluklarla birlikte gelmektedir. Bilgi güvenliği ile ilgili olarak bakıldığında, LLM'ler yanlış bilgi üretebilir (Altunay, 2024a). Dikkat çekici bir örnek, CNET'in LLM tarafından oluşturulan makaleleri açık bir açıklama yapmadan yayınlamasıydı ve bu da potansiyel yanlış bilgiye ve şeffaflık eksikliğine yol açtı. Dahası, LLM'ler siber saldırılar için kötüye kullanılabilir. Bu karmaşık düşmanca saldırılar, zararlı içerikler üreterek yapay zeka güvenlik mekanizmalarından kaçınabilir. Bu, kimlik hırsızlığı ve toplumsal düzeni ve kişisel gizliliği tehdit eden kötü niyetli sosyal medya gönderileri de dahil olmak üzere önemli riskler oluşturur (Das vd., 2025). Dahası, LLM karar alma sürecindeki olası önyargılar ve adaletsizlik kapsamlı sosyo-etik tartışmaları ateşlemiştir. Eğitim verilerindeki önyargılar adaletsiz sonuçlara yol açabilir ve klişeleri güçlendirebilir. Ayrıca LLM'lerin Windows 11 seri numaraları gibi hassas bilgileri ifşa etmeye nasıl yönlendirilebileceğini göstermiştir. Bu sorunlar teknolojinin istikrarlı gelişimini engeller ve toplumsal uyum ve istikrara meydan okur (Zhu vd., 2025).

Bu çalışma, 2020'den Nisan 2025'e kadar bilgi güvenliği çerçevesinde LLM'lere gerçekleştirilen phishing

ataklarının kapsamlı bir genel görünümünü ve karşılaştırmasını sunmaktadır. Bu alandaki en son eğilimleri ve zorlukları açıklığa kavuşturmayı amaçlamaktadır. İlgili literatürü sentezleyerek ve analiz ederek, son yıllarda bu alanda bilgi güvenliği ve sosyal etikle ilişkili gizli riskler ortaya çıkarılmıştır. Tespit ve savunma teknikleri kategorize edilerek, hem kamu hem de profesyonel kuruluşlar için büyük dil modellerin bilgi ve etik güvenliği ve savunma mekanizmaları hakkında daha net bir anlayış sağlanmıştır. Bu çalışma, LLM'lerin güvenlik etiğine derinlemesine bir bakış sunarak, bilgi güvenliğine yönelik tehditlerin yanı sıra sosyal etik bağlamında savunma ve tespit tekniklerini vurgulamaktadır. Etik güvenliğin hayati bir parçası olarak bilgi güvenliği, teknolojinin güvenli ve sorumlu bir şekilde kullanılması için hayati öneme sahiptir. Birincil amacı, bilgileri ve sistemleri yetkisiz erişim, kullanım, hasar, müdahale veya yıkımdan korumaktır. Makalede, LLM'lerin karşılaşabileceği olası sorunlar, işlevlerinin kötüye kullanılması ve kötü niyetli saldırılarla ilişkili riskler de dahil olmak üzere ele alınmaktadır. Bu tehditlere yanıt olarak, savunma ve tespit teknikleri ana hatlarıyla belirtilmiş ve model dağıtımından önce uygulanan stratejiler ve dağıtımdan sonra uygulanan acil durum önlemleri olarak kategorize edilmiştir.

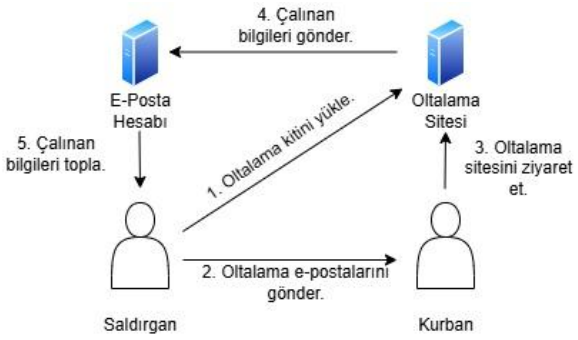
2. Büyük Dil Modelleri ve Phishing Atakları

Günümüz dijital çağında, LLM'ler yapay zekada önemli bir teknolojidir. Hayatımızın birçok alanında yaygınlaşmıştır (Geren vd., 2025). Ancak, kullanımları genişledikçe bilgi güvenliği sorunları giderek daha önemli hale gelmektedir. Bu bölüm, LLM'lerin oluşturduğu başlıca güvenlik tehditlerini inceleyerek iki temel alana odaklanmaktadır: birincisi, LLM işlevlerinin kötüye kullanılmasından kaynaklanan güvenlik sorunları; ikincisi, LLM'lere yönelik kötü niyetli saldırılardan kaynaklanan endişeler.

Phishing ataklarının sayısı ChatGPT'nin kullanılmaya başlanması ile artmıştır. Çok yüksek meblağlarda zararlar veren phishing saldırılarından korunmak için kurum ve kuruluşlar personellerini eğitmeli ve doğru siber güvenlik araçları kullanmalıdır (Khonji vd., 2013). Polimorfik kimlik avı, bir e-postanın bileşenlerini rastgele hale getirerek aynı saldırının birçok farklı versiyonunu oluşturan gelişmiş bir kimlik avı biçimidir. Amaç, kara listelere veya imza tabanlı tespite dayanan e-posta filtrelerini atlatmaktır. Bir e-postanın bir öğesi kötü amaçlı olarak tespit edilirse, bu öğe saldırının diğer sürümlerinde farklı olacak ve tanınmayacaktır (Hadnagy, 2010).

Gelişmiş tespit araçları olmadan, polimorfik saldırılara karşı savunma yapmak zordur. Saldırının bir sürümü tespit edilse bile, varyasyonları çalışanların gelen kutularına ulaşabilir. Bu saldırılar, suç-hizmet-olarak-satılan kimlik avı kitlerinde bulunan otomasyon yöntemleri sayesinde giderek yaygınlaşmaktadır (Alabdan, 2020). Özetle, polimorfik saldırıların oluşturulması kolaydır ve imza tabanlı tespit yazılımlarının tespit etmesi zordur (Bossetta, 2018).

Kimlik avı saldırıları, kullanıcı adları, parolalar ve kredi kartı bilgileri gibi hassas kullanıcı bilgilerini aldatma yoluyla elde etmeyi amaçlayan yaygın bir siber suç taktiğidir (Altunay, 2024b). Phishing ataklarının genel yapısı Şekil 1'de gösterilmiştir. Bu saldırılar iki türe ayrılabilir: büyük ölçekli kimlik avı ve hedefli kimlik avı (Xiao vd., 2025). Yapay zekadaki son gelişmeler, özellikle LLM'lerde, saldırganların son derece kişiselleştirilmiş ve ikna edici web tabanlı e-postalar oluşturmasını sağlamıştır. Bu e-postalar genellikle geleneksel, kişiselleştirilmemiş e-postaların sınırlamalarını aşar (Altunay ve Albayrak, 2024). Ek olarak, LLM'lerin kültürler arası yetenekleri, saldırganların çabalarını belirli bölgelere veya dil gruplarına göre uyarlamalarına olanak tanır ve böylece aldatıcı taktiklerini güçlendirir (Wang vd., 2025). LLM tarafından oluşturulan metinler, güvenilir kuruluşların iletişim stillerini taklit ederek kullanıcı güvenini artırır (Quinn ve Thompson, 2024). Bu yetenek, spam filtrelerini ve güvenlik sistemlerini atlatmaya yardımcı olarak kimlik avı saldırılarının başarı oranlarını önemli ölçüde artırır (Kulkarni vd., 2025). Sonuç olarak, bu saldırıların ölçeğini ve otomasyonunu genişletir. Dahası, LLM'ler kimlik avı web sitelerinin oluşturulmasını ve başlatılmasını kolaylaştırarak saldırganların kapsamlı teknik becerilere ihtiyaç duymadan etkili dolandırıcılıklar gerçekleştirmesine olanak tanır. LLM'lerin yinelemeli öğrenme yetenekleri, saldırı stratejilerinin sürekli iyileştirilmesini sağlar, kimlik avı saldırılarını daha yaygın hale getirir ve ağ güvenliği için önemli tehditler oluşturur (Afane vd., 2024). Phishing saldırılarının genel yapısı Şekil 1'de verilmiştir.



Şekil 1. Phishing saldırılarının genel yapısı.

Hızlı Enjeksiyon Güvenlik Açığı, kullanıcı istemlerinin LLM'nin davranışını veya çıktısını istenmeyen şekillerde değiştirmesi durumunda ortaya çıkar. Bu girdiler, insanlar tarafından algılanamasa bile modeli etkileyebilir; bu nedenle, içerik model tarafından ayrıştırıldığı sürece, hızlı enjeksiyonların insanlar tarafından görülebilir/okunabilir olması gerekmez (Salt, 2009).

Hızlı Enjeksiyon güvenlik açıkları, modellerin istemleri nasıl işlediğinde ve girdilerin modeli, istem verilerini modelin diğer bölümlerine hatalı bir şekilde iletmeye

nasıl zorlayabileceğinde mevcuttur. Bu da, modelin kuralları ihlal etmesine, zararlı içerik oluşturmasına, yetkisiz erişime olanak sağlamasına veya kritik kararları etkilemesine neden olabilir. Geri Alma Artırılmış Üretimi (RAG) ve ince ayar gibi teknikler, LLM çıktılarını daha alakalı ve doğru hale getirmeyi amaçlasa da, araştırmalar bunların hızlı enjeksiyon güvenlik açıklarını tamamen azaltmadığını göstermektedir (Boyd ve Keromytis, 2004). Hızlı enjeksiyon ve jailbreak, LLM güvenliğinde ilişkili kavramlar olsa da, genellikle birbirinin yerine kullanılırlar. Hızlı enjeksiyon, güvenlik önlemlerini atlatmayı da içerebilen davranışını değiştirmek için model yanıtlarını belirli girdiler aracılığıyla manipüle etmeyi içerir. Jailbreak, saldırganın modelin güvenlik protokollerini tamamen göz ardı etmesine neden olan girdiler sağladığı bir tür hızlı enjeksiyondur. Geliştiriciler, hızlı enjeksiyon saldırılarını azaltmak için sistem komutlarına ve girdi işleme süreçlerine güvenlik önlemleri ekleyebilirler, ancak jailbreak'in etkili bir şekilde önlenmesi, modelin eğitim ve güvenlik mekanizmalarının sürekli güncellenmesini gerektirir (Zheng vd., 2024).

İnsanlar tarafından hazırlanan phishing saldırıları ile LLM'ler tarafından hazırlanan phishing saldırlarının etkilerinin karşılaştırılması Tablo 1'de gösterilmiştir.

Tablo 1. İnsan ve LLM kullanılarak hazırlanan phishing saldırılarının etkileri (Bethany vd., 2025)

E-Posta Yazarı	E-Posta Konusu	E-Posta Alıcıları	Açılan E-Posta Oranı	Tıklanan Bağlantı Oranı	Veri Girişi Oranı
İnsan	Kurum Amirinden Rapor	1541	%61,97	%21,67	%10,64
LLM	Kurum Amirinden Rapor	1490	%61,68	%21,34	%11,07
İnsan	Süreli Oltalama	1591	%36,08	%6,66	%2,51
LLM	Süreli Oltalama	1479	%40,37	%10,34	%4,19
İnsan	Genel Oltalama	1520	%54,67	%4,47	%1,18
LLM	Genel Oltalama	1508	%47,41	%4,84	%0,93

3. Tartışma ve Sonuç

LLM'ler çeşitli alanlarda giderek daha fazla kullanıldıkça, ilişkili bilgi güvenliği ve etik zorlukları daha belirgin hale geliyor. Bu makale, bilgi güvenliğindeki en son akademik gelişmelerden yararlanarak LLM'lerle ilgili güvenlik tehditlerini, savunma tekniklerini ve sosyo-etik sorunları analiz ediyor. Ayrıca, LLM'lerin oluşturduğu yeni güvenlik tehditlerini ortaya çıkarıyor. Bunlara kimlik avı saldırıları, kötü amaçlı yazılım tehditleri, bilgisayar korsanlığı olayları, sosyal mühendislik saldırıları, yanlış bilgilendirme ve LLM yeteneklerini istismar eden diğer kötüye kullanımlar dahildir. Ek olarak, model ters çevirme saldırıları, zehirleme saldırıları, arka kapı saldırıları, ipucu enjeksiyonları ve jailbreak saldırıları gibi riskler de bulunmaktadır. Makale ayrıca LLM güvenliğini artırmak için çeşitli stratejiler sunmaktadır. Teknolojinin sosyal etik üzerindeki etkisini ele alıyor ve otomatik düşmanca eğitim tekniklerinin geleceğini, çok dilli ortamlarda saldırı uyarlamalarının incelenmesini ve etik ve yasal çerçevelere olan ihtiyacı tartışıyor. Bu içgörüler, gelecekteki güvenlik uygulamaları ve LLM teknolojisinin gelişimi için güçlü bir teorik temel ve kapsamlı bir bakış açısı sağlar.

Katkı Oranı Beyanı

Yazarın katkı yüzdeleri aşağıda verilmiştir. Yazar makaleyi incelemiş ve onaylamıştır.

%	H.C.A.
K	100
T	100
Y	100
VTI	100
VAY	100
KT	100
YZ	100
GR	100
PM	100

K= kavram, T= tasarım, Y= yönetim, VTI= veri toplama ve/veya işleme, VAY= veri analizi ve/veya yorumlama, KT= kaynak tarama, YZ= Yazım, GR= gönderim ve revizyon, PY= proje yönetimi.

Çatışma Beyanı

Yazar bu çalışmada hiçbir çıkar ilişkisi olmadığını beyan etmektedirler.

Etik Onay Beyanı

Bu araştırmada hayvanlar ve insanlar üzerinde herhangi bir çalışma yapılmadığı için etik kurul onayı alınmamıştır.

Kaynaklar

Afane, K., Wei, W., Mao, Y., Farooq, J., & Chen, J. (2024). Next-generation phishing: How LLM agents empower cyber attackers. In *2024 IEEE International Conference on Big Data (Big Data)* (pp. 2558–2567). IEEE. https://doi.org/10.1109/BigData62323.2024.10825316

Alabdan, R. (2020). Phishing attacks survey: Types, vectors, and technical approaches. *Future Internet*, *12*(10), 168. https://doi.org/10.3390/fi12100168

Altunay, H. C. (2024a). Detection of SQL Injection attacks using machine learning algorithms based on NLP-based feature extraction. In *9th International Conference on Computer Science and Engineering (UBMK)* (pp. 468–472). IEEE. https://doi.org/10.1109/UBMK62933.2024.10756715

Altunay, H. C. (2024b). Analysis of cyber attacks using honeypot. *Black Sea Journal of Engineering and Science*, *7*(5), 954–959. https://doi.org/10.52704/bsscience.1481075

Altunay, H. C., & Albayrak, Z. (2024). SMS spam detection system based on deep learning architectures for Turkish and English messages. *Applied Sciences*, *14*(24), 11804. https://doi.org/10.3390/app142411804

Annepaka, Y., & Pakray, P. (2025). Large language models: A survey of their development, capabilities, and applications. *Knowledge and Information Systems*, *67*(3), 2967–3022. https://doi.org/10.1007/s10115-024-02264-0

Bethany, M., Galiopoulos, A., Bethany, E., Karkevandi, M. B., Beebe, N., Vishwamitra, N., & Najafirad, P. (2025). Lateral phishing with large language models: A large organization comparative study. *IEEE Access*, *13*, 60684–60701. https://doi.org/10.1109/ACCESS.2025.3526685

Bossetta, M. (2018). The weaponization of social media: Spear phishing and cyberattacks on democracy. *Journal of International Affairs*, *71*(2), 97–106.

Boyd, S. W., & Keromytis, A. D. (2004). SQLrand: Preventing SQL injection attacks. In *Applied Cryptography and Network Security (ACNS 2004)* (pp. 292–302). Springer. https://doi.org/10.1007/978-3-540-24852-1_23

Das, B. C., Amini, M. H., & Wu, Y. (2025). Security and privacy challenges of large language models: A survey. *ACM Computing Surveys*, *57*(6), 1–39. https://doi.org/10.1145/3690650

Geren, C., Board, A., Dagher, G. G., Andersen, T., & Zhuang, J. (2025). Blockchain for large language model security and safety: A holistic survey. *ACM SIGKDD Explorations Newsletter*, *26*(2), 1–20. https://doi.org/10.1145/3706497.3706500

Hadnagy, C. (2010). *Social engineering: The art of human hacking*. John Wiley & Sons.

Khonji, M., Iraqi, Y., & Jones, A. (2013). Phishing detection: A literature survey. *IEEE Communications Surveys & Tutorials*, *15*(4), 2091–2121. https://doi.org/10.1109/SURV.2013.032213.00001

Kulkarni, A., Balachandran, V., Divakaran, D. M., & Das, T. (2024). From ML to LLM: Evaluating the robustness of phishing webpage detection models against adversarial attacks. *Digital Threats: Research and Practice*, *6*(2), 1–25. https://doi.org/10.1145/3696455

Quinn, T., & Thompson, O. (2024). *Applying large language model (LLM) for developing cybersecurity policies to counteract spear phishing attacks on senior corporate managers* [Preprint]. Research Square. https://doi.org/10.21203/rs.3.rs-4405206/v1

Salt, C. (2009). *SQL injection attacks and defense*. Elsevier (Syngress).

Wang, S., Zhao, Y., Hou, X., & Wang, H. (2025). Large language model supply chain: A research agenda. *ACM Transactions on Software Engineering and Methodology*, *34*(5), 1–46. https://doi.org/10.1145/3702995

Xiao, X., Zhang, Y., Xu, J., Ren, W., & Zhang, J. (2025). Assessment methods and protection strategies for data leakage risks in large language models. *Journal of Industrial Engineering and Applied Science*, *3*(2), 6–15.

Zheng, J., Qiu, S., Shi, C., & Ma, Q. (2025). Towards lifelong learning of large language models: A survey. *ACM Computing Surveys*, *57*(8), 1–35. https://doi.org/10.1145/3703155

Zheng, X., Pang, T., Du, C., Liu, Q., Jiang, J., & Lin, M. (2024). Improved few-shot jailbreaking can circumvent aligned language models and their defenses. *Advances in Neural Information Processing Systems (NeurIPS)*, *37*, 32856–32887.

Zhu, X., Zhou, W., Han, Q. L., Ma, W., Wen, S., & Xiang, Y. (2025). When software security meets large language models: A survey. *IEEE/CAA Journal of Automatica Sinica*, *12*(2), 317–334. https://doi.org/10.1109/JAS.2024.410762