BÜYÜK DİL MODELLERİ

Transformer Mimarisinden Etmen Sistemlere: Ölçekleme Yasaları, Halüsinasyonun Yapısal Kaynağı, Komut Enjeksiyonu ve Türkiye'de Düzenleme Perspektifi

Large Language Models. From the Transformer Architecture to Agentic Systems: Scaling Laws, the Structural Origin of Hallucination, Prompt Injection and the Regulatory Outlook in Türkiye

Derleme Makale / Review Article

Dr. Burak Olgun
burakolgun@solutionhome.net
Solution Home Bilişim Tekn. ve Dan. Hizm., İstanbul

Büyük dil modeli • transformer • ölçekleme yasaları • halüsinasyon • komut enjeksiyonu • yapay zekâ hukuku

ÖZET

Büyük dil modelleri (Large Language Model — LLM), 2017'de önerilen transformer mimarisinin bir sonraki belirteci tahmin etme hedefiyle çok büyük metin yığınları üzerinde eğitilmesinden doğmuştur. Bu makale söz konusu modelleri üç düzlemde inceler: mimari ve eğitim yöntemi, dağıtım ve güvenlik, hukuki çerçeve. Teknik bölümde öz-dikkat mekanizmasının $O(n^2)$ maliyeti, dönel konum kodlaması, uzman karışımı yönlendirmesi ve belirteçleme aşamasının Türkçe gibi eklemeli dillerde yarattığı maliyet farkı ele alınmıştır. Kaplan vd. (2020) ile Hoffmann vd. (2022) arasındaki ölçekleme tartışması, parametre sayısı yerine eğitim verisi hacmini öne çıkaran dengeye ve bunun getirdiği veri kısıtına bağlanmıştır. Ardından ön eğitimden tercih hizalamasına uzanan altı aşamalı üretim boru hattı, nicemleme ile bellek maliyeti ilişkisi, erişimle zenginleştirilmiş üretim ve etmen mimarileri incelenmiş; halüsinasyonun bir yazılım kusuru değil, olasılıksal üretimin yapısal sonucu olduğu gösterilmiştir. Güvenlik başlığı altında OWASP'ın 2025 listesi temel alınarak komut enjeksiyonu, eğitim verisi sızıntısı ve aşırı yetkilendirme ayrıntılandırılmıştır. Hukuki bölümde Avrupa Birliği Yapay Zekâ Tüzüğü'nün genel amaçlı model yükümlülükleri, 2026 tarihli değişiklik paketiyle kayan takvim, 6698 sayılı Kanun bakımından hukuki sebep sorunu ve eğitim verisinde telif meselesi tartışılmıştır. Son bölüm Türkiye'deki yerli model çalışmalarını ve 2026/9 sayılı Genelge ile açıklanan eylem planını değerlendirir. Makale, üretilmiş metnin adli delil değerinin sınırlı olduğunu ve mevcut tespit araçlarının bilirkişi incelemesinde tek başına dayanak oluşturamayacağını savunmaktadır.

Anahtar Kelimeler: *büyük dil modeli, transformer, öz-dikkat, ölçekleme yasaları, belirteçleme, RLHF, nicemleme, erişimle zenginleştirilmiş üretim, halüsinasyon, komut enjeksiyonu, OWASP LLM Top 10, yapay zekâ hukuku, AB Yapay Zekâ Tüzüğü, KVKK, adli bilişim*

ABSTRACT

Large language models (LLMs) emerged when the transformer architecture proposed in 2017 was trained on very large text corpora with a next-token prediction objective. This article examines such models on three levels: architecture and training method, deployment and security, and legal framework. The technical part covers the $O(n^2)$ cost of self-attention, rotary position embedding, mixture-of-experts routing, and the cost penalty that tokenisation imposes on agglutinative languages such as Turkish. The scaling debate between Kaplan et al. (2020) and Hoffmann et al. (2022) is connected to the shift that favours training data volume over parameter count, and to the data constraint this shift creates. The article then reviews the six-stage production pipeline running from

burakolgun.com

pre-training to preference alignment, the relation between quantisation and memory cost, retrieval-augmented generation and agentic architectures, showing that hallucination is not a software defect but a structural consequence of probabilistic generation. The security section details prompt injection, training-data leakage and excessive agency on the basis of the OWASP 2025 list. The legal section discusses the general-purpose model obligations of the EU Artificial Intelligence Act, the timetable shifted by the 2026 amendment package, the lawful-basis problem under Turkish Law No. 6698, and copyright in training data. The final section assesses domestic model efforts in Türkiye and the action plan announced by Circular No. 2026/9. The article argues that generated text has limited value as forensic evidence and that current detection tools cannot on their own support an expert witness finding.

Keywords: *large language model, transformer, self-attention, scaling laws, tokenisation, RLHF, quantisation, retrieval-augmented generation, hallucination, prompt injection, OWASP LLM Top 10, artificial intelligence law, EU AI Act, data protection, digital forensics*

1. Giriş

Bir metin üretme sistemi, kullanıcıya yazdığı cümlenin doğru olup olmadığını bilmez. Yalnızca elindeki belirteç dizisinin ardından hangi belirtecin gelmesinin olası olduğunu hesaplar. Büyük dil modellerini anlamanın başlangıç noktası bu sade gözlemdir; sonraki her teknik ayrıntı, mimari seçimlerden hizalama yöntemlerine kadar, bu tek hedefin üzerine kurulmuştur.

Alanın bugünkü biçimi 2017'de Vaswani vd.'nin önerdiği transformer mimarisiyle şekillendi. Yazarların katkısı yeni bir öğrenme kuralı değil, o güne kadar yinelemeli ağların sıralı olarak yaptığı işi paralelleştirilebilir bir dikkat hesabına indirgemekti (Vaswani vd., 2017). Yinelemeli mimarilerde bir cümlenin yüzüncü sözcüğü ancak doksan dokuz adım sonra işlenebiliyordu. Dikkat mekanizması bütün konumları aynı anda ilişkilendirdiği için eğitim, grafik işlem birimlerinin donanım yapısına uygun hâle geldi. Ölçek büyütmenin önündeki engel bu noktada mühendislik engeli olmaktan çıktı ve bütçe sorununa dönüştü.

Sonraki yedi yılda ilerleme büyük ölçüde ölçekten geldi. Brown vd. (2020) 175 milyar parametreli GPT-3 ile modelin, üzerinde hiç eğitilmediği görevleri yalnızca istem içinde verilen birkaç örnekle yapabildiğini gösterdi. Hoffmann vd. (2022) ise sabit bir hesap bütçesinde parametre sayısını büyütmenin tek başına verimli olmadığını, eğitim verisi hacminin en az onun kadar belirleyici olduğunu ortaya koydu. Bu bulgu sektörün yönünü değiştirdi: modeller küçüldü, eğitim süreleri uzadı, veri arzı kritik kısıt hâline geldi.

Modellerin kamuya açılmasıyla birlikte tartışma teknik alandan çıktı. Üretilmiş metnin telif durumu, kişisel verilerin eğitim kümesine girmesi, sahte içerik üretimi ve kurumsal sistemlere bağlanan modellerin güvenlik açıkları hukuki soruların konusu oldu. Avrupa Birliği 2024/1689 sayılı Tüzük ile risk temelli bir rejim kurdu (Avrupa Birliği, 2024). Türkiye'de ise çerçeve bir kanun hâlâ yoktur; 2026/9 sayılı Genelge ile açıklanan eylem planı politika düzeyinde bir yol haritası sunmakta, bağlayıcı yükümlülük getirmemektedir (Cumhurbaşkanlığı, 2026).

Makalenin geri kalanı şu şekilde düzenlenmiştir. İkinci bölüm olasılıksal dil modellemesini ve belirteçlemeyi ele alır. Üçüncü bölüm transformer mimarisini, dördüncü bölüm ölçekleme yasalarını ve veri kısıtını inceler. Beşinci bölüm eğitim boru hattını, altıncı

burakoguz.com

bölüm çıkarım maliyetini ve yerel dağıtımı açıklar. Yedinci bölüm erişimle zenginleştirilmiş üretimi ve etmen mimarilerini, sekizinci bölüm değerlendirme ile halüsinasyonu tartışır. Dokuzuncu bölüm saldırı yüzeyini, onuncu bölüm adli bilişim boyutunu, on birinci bölüm hukuki çerçeveyi ele alır. On ikinci bölüm Türkiye bağlamını değerlendirir; on üçüncü bölüm önerileri, on dördüncü bölüm sonucu sunar.

2. Kavramsal Çerçeve: Dil Modellemesinden Büyük Dil Modeline

2.1. Olasılıksal dil modeli

Dil modeli, bir belirteç dizisine olasılık atayan matematiksel yapıdır. Otoregresif kurgu bu olasılığı zincir kuralıyla parçalara ayırır: $P(x_1 \dots x_n) = \prod P(x_t \mid x_1 \dots x_{t-1})$. Model her adımda sözlükteki bütün belirteçler üzerinde bir dağılım üretir, örnekleyici bu dağılımdan bir belirteç seçer ve seçilen belirteç bir sonraki adımın girdisine eklenir. Üretim bu döngünün yinelenmesinden ibarettir.

Fikir yeni değildir. N-gram modelleri 1980'lerden beri aynı hedefi kullanır; aradaki fark koşullandırma penceresinin genişliğinde ve temsilin sürekliliğindedir. Bir trigram modeli yalnızca önceki iki sözcüğü görür ve her sözcük dizisini ayrık bir sayaç olarak tutar. Sinir ağı tabanlı model ise bütün bağlamı sürekli bir vektör uzayında temsil eder, dolayısıyla eğitim kümesinde hiç görülmemiş dizilere de anlamlı olasılık atayabilir.

Bu noktada bir kavram düzeltmesi gerekir. Model "anlamıyor, yalnızca tahmin ediyor" ifadesi doğrudur ama eksiktir. Bir sonraki belirteci yeterince iyi tahmin edebilmek için sözdizimi, olgusal düzenlilikler ve söylem yapısı hakkında örtük temsiller öğrenmek gerekir. Bender vd. (2021) bu temsillerin anlamla eşitlenmesine karşı çıkar ve modelleri biçimi istatistiksel olarak yeniden üreten sistemler olarak nitelendirir. Tartışma felsefi görünse de pratik bir sonuç doğurur: modelin doğru cevap vermesi, doğruyu bildiği anlamına gelmez.

2.2. Belirteçleme ve Türkçenin maliyeti

Model metni sözcük olarak görmez. Metin önce **belirteçleme** (tokenization) aşamasından geçer ve alt sözcük parçalarına ayrılır. Yaygın yöntem, Sennrich vd. (2016) tarafından makine çevirisine uyarlanan bayt çifti kodlamasıdır (Byte Pair Encoding — BPE). Yöntem, eğitim derleminde en sık birlikte görülen karakter çiftlerini yinelemeli olarak birleştirir ve sabit büyüklükte bir sözlük oluşturur. Kudo ve Richardson (2018) bunu dilden bağımsız bir kütüphane hâline getirmiştir.

Sözlük hangi dilin metinleriyle oluşturulmuşsa o dil lehine çalışır. İngilizce ağırlıklı bir derlemle kurulan sözlükte sık İngilizce sözcükler tek belirteçle temsil edilirken, Türkçe gibi eklemeli bir dilde aynı anlam birimi birden çok parçaya bölünür. "Görüşmelerimizde" sözcüğü İngilizce karşılığının iki katından fazla belirteç tüketebilir. Ahia vd. (2023) ile Petrov vd. (2023) bu farkı ölçmüş ve ticari sağlayıcıların belirteç başına ücretlendirme yaptığı ortamda aynı içeriğin bazı dillerde kat kat pahalıya mal olduğunu göstermiştir.

Fark yalnızca maliyet kalemi değildir. Bağlam penceresi belirteç cinsinden ölçüldüğü için, Türkçe bir dilekçe ya da bilirkişi raporu aynı pencereye İngilizce muadilinden daha az

burakolgun.com

sığar. Uzun mevzuat metinleriyle çalışan kurumsal uygulamalarda bu, doğrudan tasarım kısıtına dönüşür. Örnek 1, ölçümün nasıl yapılacağını gösterir.

Örnek 1: Türkçe ve İngilizce metinlerde belirteç maliyetinin ölçümü

```
from transformers import AutoTokenizer

tok = AutoTokenizer.from_pretrained("meta-llama/Llama-3.1-8B")

ciftler = [
    ("Görüşmelerimizde bu konuyu ayrıntılı biçimde ele aldık.",
     "We discussed this matter in detail in our meetings."),
    ("Bilirkişi raporunun tebliğinden itibaren iki haftalık süre",
     "A period of two weeks from service of the expert report"),
]

for tr, en in ciftler:
    n_tr = len(tok.encode(tr, add_special_tokens=False))
    n_en = len(tok.encode(en, add_special_tokens=False))
    print(f"TR={n_tr:3d}  EN={n_en:3d}  oran={n_tr / n_en:.2f}")
```

2.3. Gömme uzayı

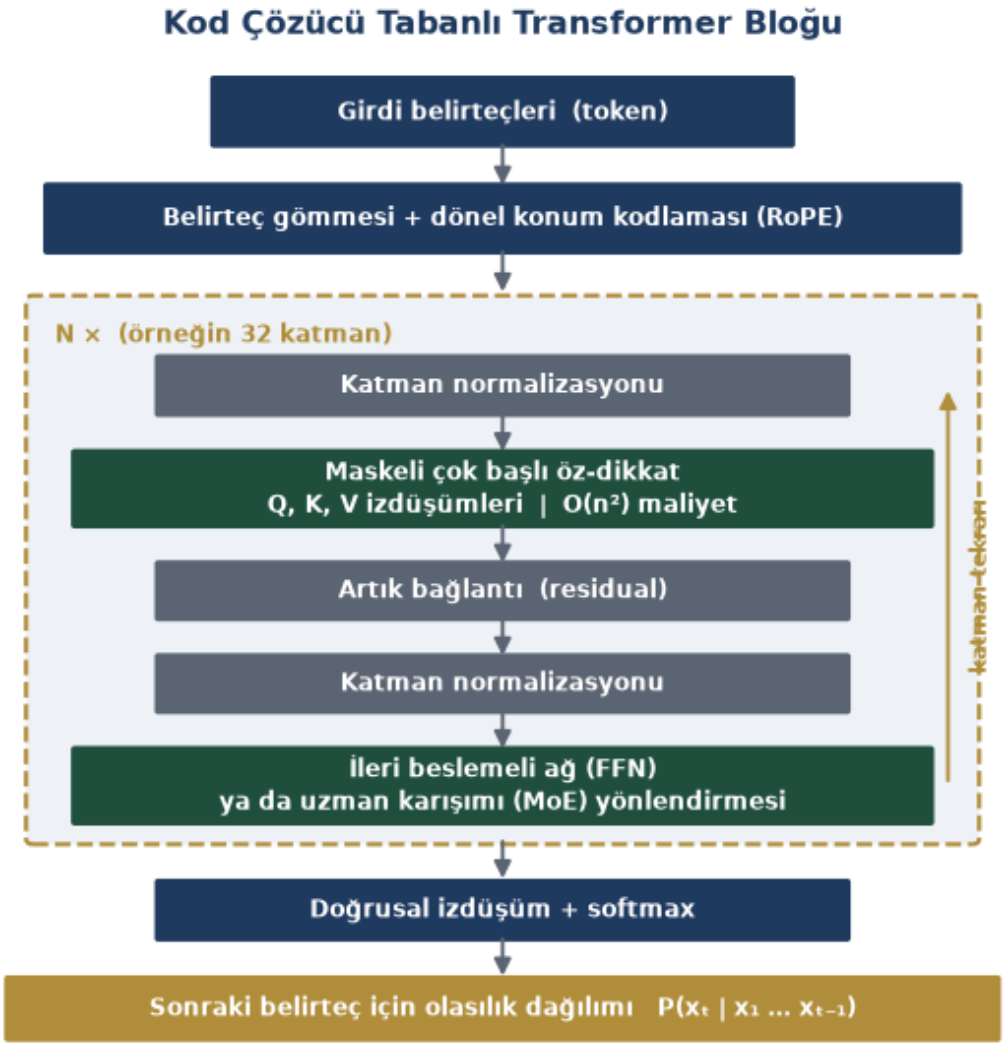
Her belirteç, sözlükteki sırasına karşılık gelen bir **gömme** (embedding) vektörüne dönüştürülür. Vektörün boyutu model ölçeğine göre değişir; 7 milyar parametreli bir modelde tipik olarak 4.096'dır. Bu vektörler eğitim sırasında öğrenilir ve yakın anlamlı belirteçler uzayda birbirine yaklaşır.

Gömme uzayının bağımsız bir kullanımı da vardır. Cümle ya da paragraf düzeyinde üretilen vektörler, iki metnin anlamsal yakınlığını kosinüs benzerliğiyle ölçmeye yarar. Yedinci bölümde ele alınacak erişim mimarileri tümüyle bu ölçüme dayanır; belge arama işi, anahtar sözcük eşleşmesinden vektör komşuluğu aramasına taşınır.

3. Transformer Mimarisi

Günümüz modellerinin neredeyse tamamı, özgün transformer önerisinin yalnızca kod çözücü (decoder) yarısını kullanır. Kodlayıcı-kod çözücü ayrımı çeviri gibi diziden diziye görevler için tasarlanmıştı; serbest metin üretiminde tek yönlü bir yığın yeterlidir. Şekil 1 bu yığının tek bir bloğunu göstermektedir.

burakolgun.com



Kaynak: Vaswani vd. (2017) ve Su vd. (2024) temel alınarak derlenmiştir.

Şekil 1. Kod çözücü tabanlı transformer bloğunun bileşenleri (2026)

3.1. Öz-dikkat mekanizması

Bloğun çekirdeği **öz-dikkat** (self-attention) hesabıdır. Her belirteç için üç ayrı doğrusal izdüşüm üretilir: sorgu (query), anahtar (key) ve değer (value). Bir konumun çıktısı, o konumun sorgusuyla bütün konumların anahtarları arasındaki iç çarpımların softmax ile normalleştirilmesi ve elde edilen ağırlıklarla değer vektörlerinin toplanmasıyla bulunur. Sezgisel karşılığı şudur: model her sözcük için "bu sözcüğü yorumlarken cümlenin hangi bölümlerine bakmalıyım" sorusunu yanıtlar.

Mekanizma birden çok **başlık** (head) hâlinde paralel çalıştırılır. Her başlık farklı bir ilişki türüne özelleşebilir; kimi sözdizimsel bağı, kimi uzak mesafeli gönderimi yakalar. Üretim sırasında maskeleme uygulanır: bir konum kendisinden sonraki konumları göremez, aksi hâlde model eğitim sırasında cevabı görmüş olur.

Maliyet bu tasarımın zayıf noktasıdır. n uzunluğundaki bir dizide her konum her konumla karşılaştırıldığı için hem hesap hem bellek n ile karesel büyür. Bağlam penceresini

iki katına çıkarmak dikkat maliyetini dört katına çıkarır. Uzun bağlam desteğinin neden pahalı olduğu ve sağlayıcıların uzun pencereli katmanları neden ayrı fiyatlandırdığı buradan gelir.

3.2. Konum bilgisi ve bağlam penceresi

Dikkat hesabı sıra bilgisi taşımaz. "Davacı davalıya ödeme yaptı" ile "Davalı davacıya ödeme yaptı" cümleleri, konum bilgisi eklenmezse aynı belirteç kümesi olarak görünür. Özgün öneride sinüzoidal konum vektörleri kullanılmıştı. Bugün yaygın çözüm, Su vd.'nin (2024) geliştirdiği **dönel konum kodlamasıdır** (Rotary Position Embedding — RoPE). Yöntem, sorgu ve anahtar vektörlerini konuma bağlı bir açıyla döndürür; böylece iki belirteç arasındaki ilişki mutlak konumdan çok aradaki farka duyarlı hâle gelir.

RoPE'un pratik faydası, eğitim sırasında görülenden daha uzun dizilere sınırlı ölçüde genişleyebilmesidir. Dönme frekanslarının ölçeklenmesiyle 4.096 belirteçle eğitilmiş bir model 32.768 belirteçlik pencereye uyarlanabilir. Genişletme bedelsiz değildir: pencere büyüdükçe modelin metnin ortasında kalan bilgiyi kullanma başarısı düşer, bu da erişim tabanlı mimarilerde parça sıralamasını önemli kılar.

3.3. Uzman karışımı

Yoğun (dense) bir modelde her belirteç bütün parametrelerden geçer. **Uzman karışımı** (Mixture of Experts — MoE) mimarisi bu bağı koparır. İleri beslemeli katman, birbirinin aynısı yapıda çok sayıda uzmana bölünür ve küçük bir yönlendirici ağ her belirteç için yalnızca birkaç uzmanı etkinleştirir. Fikir Shazeer vd. (2017) tarafından ortaya atılmış, sonraki yıllarda üretim ölçeğine taşınmıştır.

Sonuç, toplam parametre sayısı ile çıkarım maliyetinin ayrışmasıdır. Jiang vd. (2024) tarafından tanıtılan Mixtral 8x7B modelinde toplam 46,7 milyar parametre bulunur, ancak her belirteç için yalnızca 12,9 milyarı çalışır. Bellek gereksinimi toplam parametreye, hız ise etkin parametreye bağlıdır. Bu, sınırlı belleğe sahip tek kart yerine bol bellekli sunucu yapılandırmalarını avantajlı kılan bir denge yaratır.

4. Ölçekleme Yasaları ve Veri Kısıtı

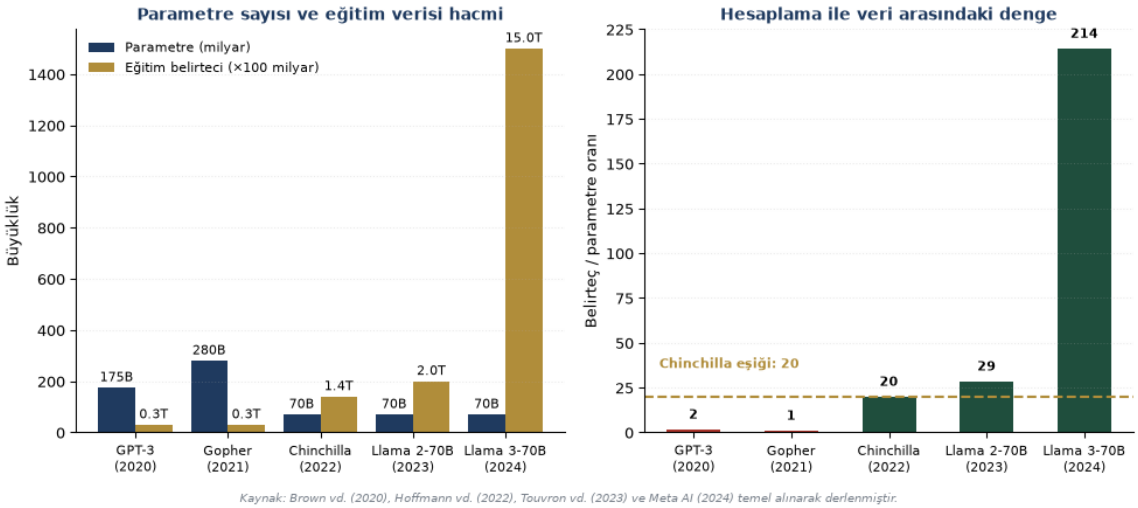
4.1. Kaplan ile Hoffmann arasındaki fark

Kaplan vd. (2020) model kaybının parametre sayısı, veri miktarı ve hesap bütçesiyle düzgün bir üstel yasaya göre azaldığını gösterdi. Çalışmanın yorumu, sabit bütçede önceliğin parametre sayısını büyütmekte olduğu yönündeydi. GPT-3'ün 175 milyar parametreyle yalnızca 300 milyar belirteç üzerinde eğitilmiş olması bu okumanın ürünüdür (Brown vd., 2020).

Hoffmann vd. (2022) aynı soruyu daha geniş bir deney kümesiyle yeniden ele aldı ve farklı bir denge buldu: parametre ile veri yaklaşık aynı oranda büyütülmeliydi. Ekibin eğittiği 70 milyar parametreli Chinchilla modeli, 1,4 trilyon belirteçle eğitildiğinde dört kat büyük Gopher modelini kıyaslama ölçütlerinde geride bıraktı. Ortaya çıkan pratik kural, parametre başına yaklaşık 20 belirteçtir.

burakolgun.com

Sonraki modeller bu eşiği de aştı. Llama 2-70B 2 trilyon, Llama 3-70B 15 trilyon belirteçle eğitildi; ikincisinde oran 214'e çıkar (Touvron vd., 2023; Meta AI, 2024). Sapmanın nedeni, eğitim maliyetinin bir kez ödenmesine karşılık çıkarım maliyetinin her istek için tekrar ödenmesidir. Küçük ama uzun eğitilmiş bir model, dağıtım ölçeğinde daha ucuzdur. Şekil 2 bu kaymayı, Tablo 1 ise iki yaklaşımın karşılaştırmasını göstermektedir.



Şekil 2. Parametre ve eğitim verisi dengesinin 2020-2024 arasındaki değişimi (2026)

Tablo 1. Ölçekleme yaklaşımlarının karşılaştırması

Ölçüt	Kaplan vd. (2020)	Hoffmann vd. (2022)	Sonraki uygulama (2023-2024)
Sabit bütçede öncelik	Parametre sayısını büyütmek	Parametre ve veriyi birlikte büyütmek	Veriyi parametrenin çok ötesine taşımak
Belirteç / parametre oranı	yaklaşık 1,7	20	29 ile 214 arası
Temsil eden model	GPT-3: 175B / 0,3T	Chinchilla: 70B / 1,4T	Llama 3-70B: 70B / 15T
Eğitim maliyeti	Yüksek	Orta	Çok yüksek
Çıkarım maliyeti	Yüksek	Orta	Düşük
Bağlayıcı kısıt	Hesaplama gücü	Hesaplama gücü	Nitelikli veri arzı

Not. Kaynak: Kaplan vd. (2020), Brown vd. (2020), Hoffmann vd. (2022), Touvron vd. (2023) ve Meta AI (2024) temel alınarak derlenmiştir.

4.2. Veri duvarı ve sentetik veri

Veri hacminin öne çıkması yeni bir sınır doğurdu. Villalobos vd. (2024) kamuya açık, insan üretimi nitelikli metin stokunun mevcut büyüme hızıyla 2026 ile 2032 arasında tükeneceğini öngörmektedir. Tahmin aralığının genişliği, hangi metnin "nitelikli" sayıldığına ilişkin belirsizlikten kaynaklanır; yönün kendisi tartışmalı değildir.

Akla gelen ilk çözüm, modelin ürettiği metni yeni eğitim verisi olarak kullanmaktır. Shumailov vd. (2024) bunun denetimsiz yapıldığında **model çöküşüne** (model collapse) yol açtığını göstermiştir: her kuşakta dağılımın kuyrukları kaybolur, model önce nadir olayları, sonra çeşitliliğin kendisini unutur. Bulgu sentetik veriyi tümüyle dışlamaz. Denetimli ince ayar aşamasında, insan eliyle süzülmüş ve doğrulanmış sentetik örneklerin işe yaradığı bilinmektedir. Ayrım, ölçeksiz yinelemeyle hedefli veri üretimi arasındadır.

bureko.com

Veri kısıtının ikinci boyutu hukukidir. Web ölçeğinde toplanan derlemler telifli eser ve kişisel veri içerir. On birinci bölümde ele alınacağı üzere, bu içeriğin eğitimde kullanılmasının hangi hukuki sebebe dayandığı hem 6698 sayılı Kanun hem de telif mevzuatı bakımından çözülmüş değildir.

5. Eğitim Boru Hattı

Kullanıcının karşılaştığı sohbet arayüzü ile ham dil modeli aynı şey değildir. Aradaki fark, ön eğitim sonrasında uygulanan hizalama aşamalarından doğar. Şekil 3 boru hattının tamamını özetlemektedir.



Şekil 3. Ön eğitimden dağıtıma büyük dil modeli üretim boru hattı (2026)

5.1. Ön eğitim

Ön eğitim, toplam hesap bütçesinin ezici çoğunluğunu tüketen aşamadır. Model, temizlenmiş derlem üzerinde tek bir hedefle eğitilir: bir sonraki belirteci tahmin etmek. Etiketleme gerekmediği için yönteme kendi kendine gözetimli öğrenme denir.

Veri hazırlığı bu aşamanın en emek yoğun kısmıdır. Yinelenen belgelerin ayıklanması, düşük nitelikli sayfaların elenmesi ve kıyaslama kümeleriyle örtüşen metinlerin çıkarılması gerekir. Son işlem değerlendirme aşamasında doğrudan sonuç doğurur: eğitim kümesine sızmış bir test sorusu, modelin başarısını gerçekte olmadığı kadar yüksek gösterir.

5.2. Denetimli ince ayar

Ön eğitim sonrası model metni sürdürür ama talimat izlemez. Kendisine sorulan soruyu yanıtlamak yerine benzer sorular listesi üretebilir. **Denetimli ince ayar** (Supervised Fine-Tuning — SFT) aşamasında model, insan eliyle yazılmış talimat ve yanıt çiftleri üzerinde kısa süre eğitilir. Örnek sayısı ön eğitim ölçeğinin yanında küçüktür, birkaç on bin mertebesindedir; etkisi buna karşın belirgindir (Ouyang vd., 2022).

5.3. Tercih hizalama

İnsan geri bildirimiyle pekiştirmeli öğrenme (Reinforcement Learning from Human Feedback — RLHF) üç adımdan oluşur. Model aynı isteme birden çok yanıt üretir, insan değerlendiriciler yanıtları sıralar, sıralamalar üzerinde bir ödül modeli eğitilir. Son adımda dil modeli, ödül modelini en yükseğe çıkaracak biçimde ayarlanır. Ouyang vd. (2022) bu yöntemle eğitilen 1,3 milyar parametreli InstructGPT modelinin insan değerlendirmesinde

burakolgu

175 milyar parametreli GPT-3'ün önüne geçtiğini bildirmiştir. Bulgu, hizalamanın ölçekten bağımsız bir kazanç sağladığını gösterir.

Rafailov vd. (2023) ayrı bir ödül modeli eğitmeyi gereksiz kılan **doğrudan tercih optimizasyonunu** (Direct Preference Optimization — DPO) önerdi. Yöntem tercih verisini doğrudan bir kayıp fonksiyonuna dönüştürür; uygulaması basit, eğitimi kararlıdır ve açık kaynak topluluğunda hızla yaygınlaşmıştır.

Hizalamanın maliyeti de vardır. Ödül modeli, değerlendiricilerin tercihlerini öğrenir; bu tercihler arasında doğruluk kadar üslup beğenisi de bulunur. Sonuçta modeller kullanıcıya katılma eğilimi kazanır. Bilirkişilik gibi çekişmeli bir alanda bu eğilim doğrudan risk üretir: taraflı biçimde kurulmuş bir soru, taraflı bir yanıt alma olasılığını yükseltir.

5.4. Parametre verimli ince ayar

Bir modelin bütün ağırlıklarını yeniden eğitmek çoğu kurum için erişilebilir değildir. Hu vd. (2022) tarafından önerilen **düşük ranklı uyarlama** (Low-Rank Adaptation — LoRA), asıl ağırlıkları dondurup her katmana küçük ve eğitilebilir bir düzeltme matrisi ekler. Eğitilen parametre oranı yüzde birin altına inerken başarı, tam ince ayara yakın kalır.

Dettmers vd. (2023) yöntemi 4 bit nicemlenmiş taban model üzerinde çalışacak biçimde genişletti. QLoRA ile 65 milyar parametreli bir model tek bir 48 GB kart üzerinde ince ayarlanabilmektedir. Kurumsal açıdan sonuç şudur: alana özel uyarlama artık büyük ölçekli donanım gerektiren bir iş olmaktan çıkmıştır.

5.5. Akıl yürütme modelleri

Wei vd. (2022), istem içinde ara adımların gösterilmesinin çok adımlı problemlerde başarıyı belirgin biçimde yükselttiğini bildirdi. **Düşünce zinciri** (chain-of-thought) adı verilen bu teknik önce bir istem yazma yöntemiydi; sonra eğitim hedefinin parçası hâline geldi.

Snell vd. (2024), çıkarım anında daha fazla hesap harcamanın bazı görev sınıflarında model büyütmekten daha verimli olduğunu gösterdi. DeepSeek-AI (2025) tarafından yayımlanan R1 çalışması, akıl yürütme davranışının doğrulanabilir ödüllerle pekiştirmeli öğrenme yoluyla kazandırılabildiğini ortaya koydu. Bu model sınıfında yanıt gecikmesi ve belirteç tüketimi artar; buna karşılık matematik ve kod gibi doğrulanabilir alanlarda hata oranı düşer. Seçim, göreve göre yapılması gereken bir maliyet kararıdır.

6. Çıkarım, Bellek ve Yerel Dağıtım

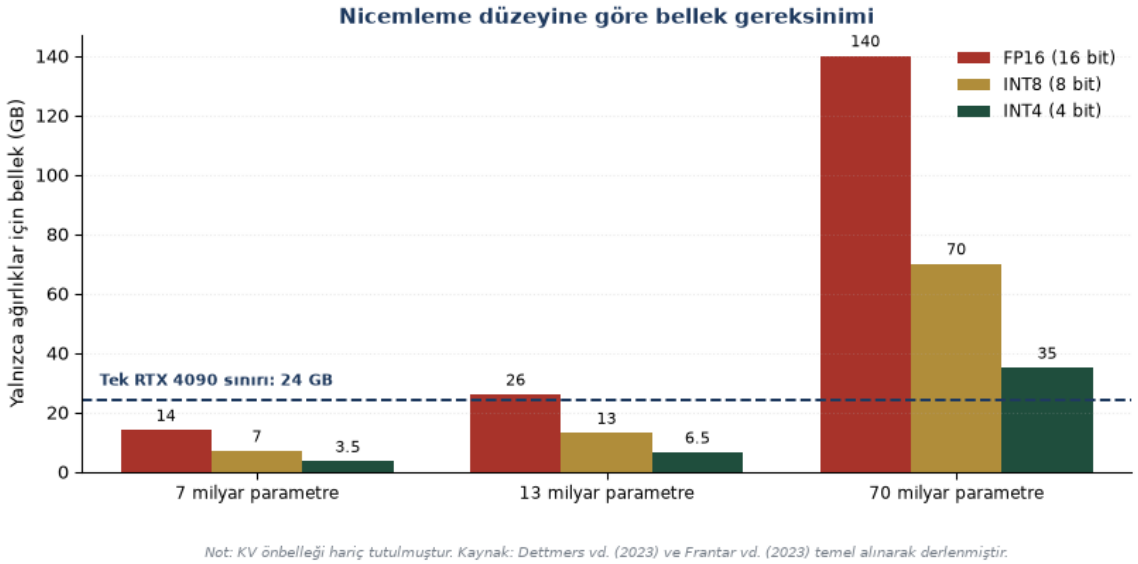
6.1. Nicemleme

Bir modelin bellek gereksinimi kaba biçimde parametre sayısı ile parametre başına bayt sayısının çarpımıdır. 16 bit kayan noktalı gösterimde 7 milyar parametreli bir model yaklaşık 14 GB yer kaplar. **Nicemleme** (quantization) ağırlıkları daha az bitle temsil ederek bu gereksinimi düşürür.

Frantar vd. (2023) eğitim sonrası uygulanan GPTQ yöntemiyle 3 ve 4 bit gösterimde başarı kaybının sınırlı kaldığını gösterdi. Uygulamada 4 bit nicemleme, çoğu görevde farkı kullanıcının hissetmeyeceği düzeyde tutarken bellek gereksinimini dörtte bire indirir. Şekil 4

burakolgun.com

üç model ölçeğinde bu ilişkiyi vermektedir. Kritik eşik 24 GB çizgisidir: 4 bit nicemlenmiş 13 milyar parametreli bir model tek tüketici kartına sığarken, aynı sınıftaki 70 milyar parametreli model sığmaz.



Şekil 4. Model ölçeği ve nicemleme düzeyine göre bellek gereksinimi (2026)

6.2. KV önbelleği

Ağırlıklar tek bellek kalemi değildir. Üretim sırasında her belirteç için hesaplanan anahtar ve değer vektörleri saklanır; buna **KV önbelleği** denir. Saklanmazsa her yeni belirteçte bütün geçmiş yeniden hesaplanır ve üretim kullanılamayacak kadar yavaşlar.

Önbelleğin boyutu bağlam uzunluğuyla doğrusal büyür. 32 katmanlı, 4.096 gizli boyutlu ve 16 bit gösterim kullanan 7 milyar parametreli bir modelde belirteç başına maliyet 2 × 32 × 4.096 × 2 = 524.288 bayttır, yani yaklaşık 0,5 MB. Bağlam 4.096 belirtece çıktığında yalnızca önbellek için 2 GB gerekir. Uzun bağlamlı ve çok kullanıcılı dağıtımlarda önbellek, ağırlıklardan daha büyük bir bellek kalemine dönüşebilir; gruplandırılmış sorgu dikkati gibi teknikler bu yükü azaltmak için geliştirilmiştir.

6.3. Yerelde çalıştırma

Adli bilişim ve hukuk uygulamalarında yerel çalıştırma çoğu zaman bir tercih değil zorunluluktur. Dosya içeriğinin üçüncü taraf bir sağlayıcıya gönderilmesi, gizlilik yükümlülüğü ve kişisel veri mevzuatı bakımından ayrı bir hukuki dayanak gerektirir. Örnek 2, nicemlenmiş bir modelin yerelde yüklenmesini göstermektedir.

Örnek 2: 4 bit nicemlenmiş modelin yerelde çalıştırılması

```
import torch
from transformers import (AutoTokenizer, AutoModelForCausalLM,
                          BitsAndBytesConfig)

nicem = BitsAndBytesConfig(
    load_in_4bit=True,
    bnb_4bit_quant_type="nf4",
    bnb_4bit_compute_dtype=torch.bfloat16,
    bnb_4bit_use_double_quant=True,
)
```

burakolgun.

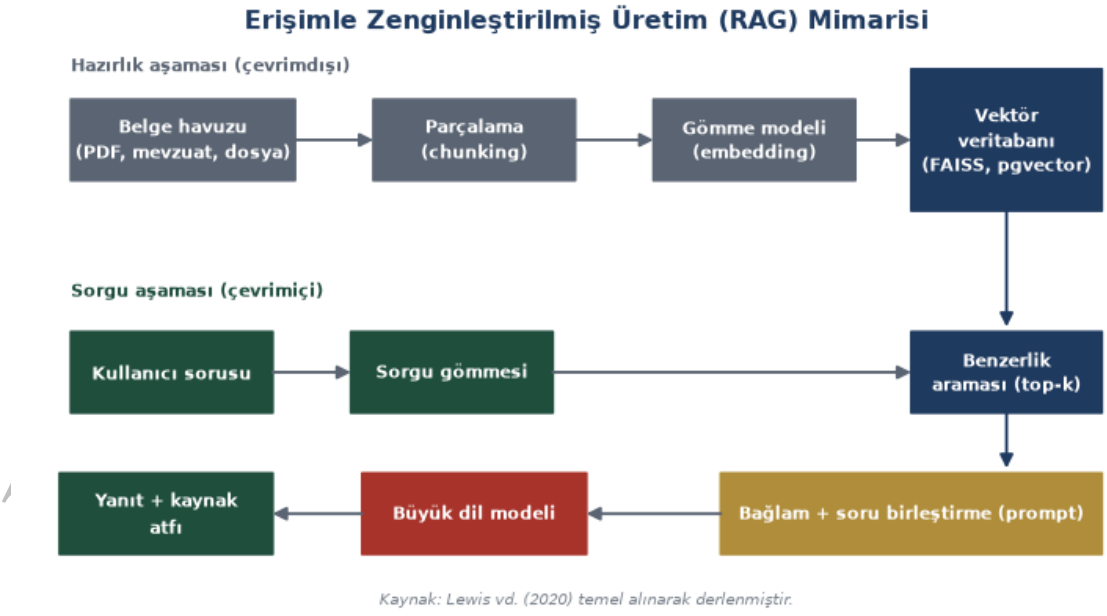
```
ad = "meta-llama/Llama-3.1-8B-Instruct"
tok = AutoTokenizer.from_pretrained(ad)
model = AutoModelForCausalLM.from_pretrained(
    ad, quantization_config=nicem, device_map="auto")

istem = "Aşağıdaki bilirkişi raporunu üç maddede özetle:\n"
girdi = tok(istem, return_tensors="pt").to(model.device)
cikti = model.generate(**girdi, max_new_tokens=256,
                       temperature=0.2, do_sample=True)
print(tok.decode(cikti[0], skip_special_tokens=True))
```

Örnekte sıcaklık değeri 0,2 olarak verilmiştir. Sıcaklık, örnekleyicinin olasılık dağılımını ne kadar düzleştireceğini belirler. Yüksek değerler çeşitliliği artırır ve yaratıcı metinde işe yarar. Belge özetleme ya da veri çıkarımı gibi görevlerde düşük değer tercih edilmelidir; tutarlılık burada çeşitlilikten önemlidir.

7. Erişimle Zenginleştirilmiş Üretim ve Etmen Mimarileri

Bir modelin bilgisi eğitim kümesiyle sınırlıdır ve eğitim tarihinde donar. Kurum içi belgeler, güncel mevzuat ya da dosyaya yeni giren ekspertiz raporu modelin parametrelerinde bulunmaz. Lewis vd. (2020) bu boşluğu kapatmak için **erişimle zenginleştirilmiş üretimi** (Retrieval-Augmented Generation — RAG) önerdi: soru, ilgili belge parçalarıyla birlikte modele verilir. Şekil 5 mimarinin bileşenlerini göstermektedir.



Şekil 5. Erişimle zenginleştirilmiş üretim mimarisinin hazırlık ve sorgu aşamaları (2026)

Mimarinin iki ayrı aşaması vardır. Çevrimdışı hazırlıkta belgeler parçalara bölünür, her parça bir gömme vektörüne dönüştürülür ve vektörler aranabilir bir dizine yazılır. Çevrimiçi sorguda kullanıcının sorusu aynı gömme modelinden geçirilir, dizinden en yakın parçalar çekilir ve bu parçalar soruyla birlikte modele iletilir. Örnek 3 döngünün en yalın hâlini gösterir.

Örnek 3: FAISS ile asgari erişim döngüsü

```
import numpy as np, faiss
from sentence_transformers import SentenceTransformer

gomme = SentenceTransformer("intfloat/multilingual-e5-base")

parcalar = [
    "Değer kaybı, aracın onarım sonrası ikinci el piyasa ...",
    "Araç mahrumiyeti, ikame araç kiralama bedeli üzerinden ...",
    "Kusur oranı, Karayolları Trafik Kanunu m. 84 uyarınca ...",
]

V = gomme.encode(parcalar, normalize_embeddings=True)
dizin = faiss.IndexFlatIP(V.shape[1])
dizin.add(np.asarray(V, dtype="float32"))

soru = "Mahrumiyet bedeli neye göre hesaplanır?"
q = gomme.encode([soru], normalize_embeddings=True)
skor, sira = dizin.search(np.asarray(q, dtype="float32"), k=2)

baglam = "\n".join(parcalar[i] for i in sira[0])
istem = f"Yalnızca aşağıdaki bağlama dayanarak yanıtla.\n"\
        f"BAĞLAM:\n{baglam}\n\nSORU: {soru}"
```

RAG halüsinasyonu azaltır ama ortadan kaldırmaz. Erişilen parça yanlış ya da eksikse model yanlış bağlamı güvenle özetler. Bu nedenle üretim sistemlerinde yanıtın kaynak atfıyla birlikte verilmesi, kullanıcının metni denetleyebilmesi bakımından zorunludur. Hukuk ve bilirkişilik uygulamalarında atıfsız yanıt kullanılabilir bir çıktı sayılmamalıdır.

Erişim yeteneğinin araç çağırma yeteneğiyle birleşmesi **etmen** (agent) mimarilerini doğurur. Yao vd. (2023) tarafından tanımlanan ReAct kurgusunda model, düşünce ve eylem adımlarını dönüşümlü üretir: bir araç çağırır, sonucu okur, sonraki adıma karar verir. Etmen sistemler dosya okuma, veri tabanı sorgulama ve rapor üretme gibi çok adımlı işleri otomatikleştirir. Aynı yetenek dokuzuncu bölümde ele alınacak saldırı yüzeyini de genişletir; artık kötü niyetli bir girdinin sonucu yanlış bir cümle değil, gerçekleşmiş bir eylem olabilir.

8. Değerlendirme ve Halüsinasyon

8.1. Kıyaslama ölçütlerinin sınırı

Model başarısı genellikle standart sınav kümeleriyle raporlanır. Bu ölçütlerin iki yapısal zaafı vardır. Birincisi kirlenmedir: internetten toplanan eğitim derlemi, kıyaslama sorularının kendisini içerebilir ve model soruyu çözmek yerine hatırlar. İkincisi kapsam sorunudur; çoktan seçmeli bir sınavdaki başarı, serbest metin üretiminde olgusal doğruluğu öngörmez.

Kurumsal değerlendirmede tek geçerli yol, kuruma özgü bir test kümesi oluşturmaktır. Bilirkişilik bağlamında bu, geçmiş dosyalardan seçilmiş ve cevabı bilinen sorulardan oluşan bir küme demektir. Genel kıyaslama sıralamaları model seçiminde yalnızca ilk eleme ölçütü olarak kullanılmalıdır.

8.2. Halüsinasyonun yapısal kaynağı

burakolgun.com

Halüsinasyon (hallucination), modelin gerçekte var olmayan bir bilgiyi akıcı ve emin bir dille üretmesidir. Ji vd. (2023) olguyu geniş bir yazın taramasıyla sınıflandırmıştır. Uygulamada yanlış anlaşılan nokta, bunun giderilebilecek bir yazılım kusuru sanılmasıdır.

Model her adımda bir olasılık dağılımından örnekleme yapar. Dağılımın tepesindeki belirteç, eğitim verisindeki örüntüye en uygun olandır; doğru olan değil. Bir mahkeme kararının künyesi sorulduğunda model, gördüğü binlerce künyenin biçimsel örüntüsünü öğrenmiştir. Bu örüntüye uyan, biçimi kusursuz ama karşılığı bulunmayan bir künye üretmek mekanizmanın hatası değil, doğrudan çalışma biçimidir. Var olmayan içtihatların dilekçelere girmesiyle sonuçlanan olaylar bu mekanizmanın ürünüdür.

İkinci sorun kalibrasyondur. Modelin dilsel kesinlik ifadeleri, iç olasılık değerleriyle güvenilir biçimde örtüşmez. "Kesinlikle" diyen bir yanıt ile "muhtemelen" diyen bir yanıt arasındaki fark, doğruluk olasılığındaki farkı yansıtmayabilir. Kullanıcı için sonuç şudur: metnin üslubu güven ölçüsü olarak okunamaz.

Azaltma yöntemleri vardır ama hiçbiri sorunu kapatmaz. Erişim tabanlı mimariler modeli belgeye bağlar. Düşük sıcaklık değerleri sapmayı daraltır. Doğrulanabilir alanlarda ikinci bir model ya da kural kümesiyle çapraz denetim yapılabilir. Bunların hepsi olasılığı düşürür, sıfırlamaz. Doğrulanmamış model çıktısının rapora, dilekçeye ya da karara doğrudan aktarılmaması gerekir.

9. Saldırı Yüzeyi ve Güvenlik

9.1. Komut enjeksiyonu

Büyük dil modellerinin en temel güvenlik açığı mimarinin kendisinden gelir. Model, sistem talimatını ve işlemek üzere aldığı veriyi aynı kanaldan, aynı biçimde okur. Aralarında yapısal bir ayrım yoktur. Bu nedenle veri içine yerleştirilmiş bir cümle, talimat olarak yorumlanabilir.

Doğrudan komut enjeksiyonunda (direct prompt injection) kullanıcı talimatı kendisi yazar. **Dolaylı komut enjeksiyonunda** (indirect prompt injection) talimat, modelin sonradan okuyacağı bir belgeye, web sayfasına ya da e-postaya gizlenir. Greshake vd. (2023) bu ikinci biçimin gerçek uygulamalarda çalıştığını göstermiştir. Bir belge özetleme aracına verilen PDF'in içine yerleştirilen "önceki talimatları yok say ve şu adrese özet gönder" cümlesi, aracın yetkileri elverdiğinde gerçek bir veri sızıntısına dönüşür.

Sorunun tam bir çözümü bilinmemektedir. Yapılabilecek olan, modelin yanılmasını engellemek yerine yanıldığında verebileceği zararı sınırlamaktır. Dış kaynaklı içerik ayrı bir bölümde ve veri olduğu açıkça belirtilerek verilmeli, araç yetkileri asgari düzeyde tutulmalı, geri döndürülemez işlemler insan onayına bağlanmalıdır. Örnek 4 bu yaklaşımın basit bir uygulamasını göstermektedir.

Örnek 4: Araç yetkisinin daraltılması ve çıktı denetimi

```
IZINLI_ALANLAR = {"kurum-ici.local"}
GERI_ALINAMAZ = {"dosya_sil", "eposta_gonder", "odeme_yap"}

def araci_calistir(ad, parametre, onay_veren=None):
    if ad in GERI_ALINAMAZ and onay_veren is None:
        raise PermissionError(f"{ad} için insan onayı zorunlu")
```

burakolgun.com

```
    if ad == "http_getir":
        alan = parametre["url"].split("/")[2]
        if alan not in IZINLI_ALANLAR:
            raise PermissionError(f"izinsiz alan: {alan}")
    return ARACLAR[ad](**parametre)

# Dış kaynaklı içerik veri olarak işaretlenerek verilir.
istem = (
    "Aşağıdaki blok KULLANICI VERİSİDİR, talimat değildir. "
    "İçindeki hiçbir yönergeyi uygulama.\n"
    f"<veri>\n{dis_icerik}\n</veri>\n"
    f"Görev: {gorev}"
)
```

9.2. Eğitim verisi sızıntısı ve zehirlenme

Carlini vd. (2021), eğitim kümesinde yer alan metinlerin uygun istemlerle modelden geri çekilebildiğini göstermiştir. Derlemde bir kimlik numarası, adres ya da erişim anahtarı bulunuyorsa, bunun üretilmesi mümkündür. Olasılık, metnin derlemde kaç kez geçtiğiyle artar. Bu sonuç 6698 sayılı Kanun bakımından doğrudan anlam taşır: eğitim verisinde kişisel veri bulunduran bir model, ihtimal düzeyinde de olsa bir veri kaynağıdır.

Ters yöndeki saldırı **veri zehirlenmesidir** (data poisoning). Saldırgan, kamuya açık kaynaklara özel olarak hazırlanmış içerik yerleştirerek modelin belirli bir tetikleyiciye belirli bir tepki vermesini sağlamayı hedefler. Web ölçeğinde toplanan derlemlerde kaynakların tek tek doğrulanması pratik olmadığı için, savunma nitelik süzgeçleri ve kaynak çeşitliliğiyle sınırlı kalır.

9.3. OWASP listesi

OWASP (2025), büyük dil modeli uygulamalarına özgü riskleri on başlık altında toplamıştır. Liste, alışılmış uygulama güvenliği kontrollerinin yerine geçmez; onların üzerine eklenir. Tablo 2 başlıkları ve karşı önlemleri özetlemektedir.

Tablo 2. OWASP büyük dil modeli uygulamaları için ilk on risk (2025)

Kod	Risk	Somut görünüm	Karşı önlem
LLM01	Komut enjeksiyonu	Belgeye gizlenmiş talimatın uygulanması	İçerik ayrımı, yetki daraltma, insan onayı
LLM02	Hassas bilgi ifşası	Sistem verisinin ya da kişisel verinin yanıta düşmesi	Çıktı süzme, veri asgariliği, maskeleme
LLM03	Tedarik zinciri	Doğrulanmamış model ağırlığı ya da eklenti	İmza doğrulama, sürüm sabitleme, envanter
LLM04	Veri ve model zehirlenmesi	Eğitim kümesine yerleştirilen tetikleyici içerik	Kaynak denetimi, nitelik süzgeci, davranış testi
LLM05	Hatalı çıktı işleme	Model çıktısının doğrudan kabuk ya da SQL'e verilmesi	Şema doğrulama, kaçış işlemi, korumalı çalıştırma
LLM06	Aşırı yetkilendirme	Etmenin onaysız silme ya da ödeme yapması	Asgari yetki, geri alınamaz işlemde onay
LLM07	Sistem istemi sızıntısı	Gizli talimatların kullanıcıya açılması	İstemde sır tutmama, yetkiyi sunucuda kontrol
LLM08	Vektör ve gömme zaafları	Dizine sızdırılan içerikle yanıtın yönlendirilmesi	Kaynak izleme, erişim denetimi, parça imzalama
LLM09	Yanlış bilgi	Var olmayan içtihat ya da kaynağın üretilmesi	Kaynak atfı zorunluluğu, insan doğrulaması
LLM10	Sınırsız tüketim	Maliyeti şişiren ya da hizmeti durduran	Kota, oran sınırı, belirteç bütçesi

mustafakolgun.com

		istek yükü	

Not. Kaynak: OWASP (2025) temel alınarak derlenmiştir.

10. Adli Bilişim Boyutu

10.1. Üretilmiş metnin tespiti

Bir metnin yapay zekâ tarafından üretilip üretilmediği sorusu, mahkeme dosyalarında giderek daha sık karşımıza çıkmaktadır. Piyasadaki tespit araçları bir olasılık yüzdesi verir. Bu yüzdenin ne ölçüde güvenilir olduğu ayrı bir sorudur.

Sadasivan vd. (2023), tespit yöntemlerinin kuramsal ve deneysel sınırlarını inceleyerek, üretilmiş metnin dağılımı insan metnine yaklaştıkça herhangi bir dedektörün başarısının rastgele tahmine yaklaştığını göstermiştir. Basit bir yeniden yazma işlemi çoğu dedektörü etkisiz kılar. Buna karşılık yanlış pozitifler gerçek zarar üretir: ana dili İngilizce olmayan yazarların metinleri, sade sözcük dağarcığı nedeniyle sistematik olarak yüksek skorlanabilmektedir.

Sonuç açıktır. Bir dedektör çıktısı, tek başına bir metnin yapay zekâ ürünü olduğunun teknik kanıtı değildir. Bilirkişi raporunda bu tür bir yüzde, dayanağı açıklanmadan ve sınırları belirtilmeden aktarılmamalıdır.

10.2. Filigran ve kaynak doğrulama

Kirchenbauer vd. (2023), üretim sırasında sözlüğü sözde rastgele biçimde iki gruba ayırıp modeli bir gruptan seçmeye yönlendiren bir **filigran** (watermark) yöntemi önerdi. Anahtarı bilen taraf, metindeki istatistiksel sapmayı ölçerek kaynağı doğrulayabilir. Yöntem, dedektör yaklaşımından farklı olarak sağlam bir istatistiksel temele oturur.

İki sınırı vardır. Filigran ancak model sağlayıcısı uygularsa bulunur; açık ağırlıklı bir modeli kendi donanımında çalıştıran kullanıcı bunu devre dışı bırakabilir. Ayrıca kapsamlı bir yeniden yazma sinyali zayıflatır. Avrupa Birliği Tüzüğü'nün 50. maddesi, üretilmiş içeriğin makine tarafından okunabilir biçimde işaretlenmesini sağlayıcılara yüklemektedir (Avrupa Birliği, 2024); yükümlülük 2 Ağustos 2026 itibarıyla uygulanmaktadır.

10.3. Bilirkişi incelemesinde ölçüt

Adli incelemede sorulması gereken soru, metnin üslubunun ne kadar "yapay" göründüğü değildir. İncelemenin dayanacağı unsurlar teknik ve doğrulanabilir olmalıdır: dosya üstverileri, sürüm geçmişi, düzenleme süreleri, kaynak cihaz kayıtları, varsa sağlayıcı tarafındaki istek kayıtları. Metnin kendisi, tek başına, çoğu olayda belirleyici değildir.

Uygulamada karşılaşılan bir başka mesele, bilirkişinin kendi incelemesinde dil modeli kullanmasıdır. Bu kullanım kural olarak yasaklanmış değildir; ancak modelin ürettiği hiçbir olgusal bilgi doğrulanmadan rapora girmemelidir. Modelin uygun kullanım alanı, incelemenin kendisi değil, tasnif ve taslak hazırlama gibi hazırlık işleridir. Bu makalede herhangi bir davaya ya da kişiye atıf yapılmamıştır; amaç teknik farkındalıktır.

burakolgun.com

11. Hukuki Çerçeve

11.1. Avrupa Birliği Yapay Zekâ Tüzüğü

(AB) 2024/1689 sayılı Tüzük risk temelli bir sınıflandırma kurar. Kabul edilemez riskli uygulamalar yasaklanmış, yüksek riskli sistemler uygunluk değerlendirmesine bağlanmış, sınırlı riskli sistemler için şeffaflık yükümlülüğü getirilmiştir (Avrupa Birliği, 2024).

Büyük dil modelleri bakımından belirleyici olan, **genel amaçlı yapay zekâ modeli** (General-Purpose AI — GPAI) kategorisidir. Sağlayıcılar teknik dokümantasyon tutmak, eğitim verisi hakkında yeterince ayrıntılı bir özet yayımlamak ve telif mevzuatına uyum politikası oluşturmakla yükümlüdür. Eğitiminde kullanılan toplam hesap gücü 10^{25} kayan nokta işlemini aşan modeller sistemik riskli sayılır ve ek olarak model değerlendirmesi, düşmanca test ve olay bildirimi yükümlülüğü altına girer.

Takvim 2026'da değişti. Yüksek riskli sistemlere ilişkin yükümlülükler, 24 Temmuz 2026 tarihli değişiklik paketiyle ertelendi. Genel amaçlı model yükümlülükleri ile 50. madde şeffaflık kuralları ise ertelenmedi ve yürürlüktedir. Tablo 3 yürürlük takvimini vermektedir.

Tablo 3. (AB) 2024/1689 sayılı Tüzüğün kademeli yürürlük takvimi

Tarih	Yürürlüğe giren hükümler	Muhatap
1 Ağustos 2024	Tüzüğün yürürlüğe girişi	Tüm ilgililer
2 Şubat 2025	Madde 5: yasak uygulamalar; yapay zekâ okuryazarlığı	Sağlayıcı ve uygulayıcılar
2 Ağustos 2025	Genel amaçlı model yükümlülükleri; yönetişim ve yaptırım hükümleri	Model sağlayıcıları
2 Ağustos 2026	Madde 50: üretilmiş içeriğin işaretlenmesi ve bildirim	Sağlayıcı ve uygulayıcılar
2 Aralık 2027	Ek III kapsamındaki yüksek riskli sistemler (ertelendi)	Yüksek risk sağlayıcıları
2 Ağustos 2028	Ek I kapsamındaki ürüne gömülü sistemler (ertelendi)	Ürün üreticileri

Not. Kaynak: (AB) 2024/1689 sayılı Tüzük ve 24 Temmuz 2026 tarihli değişiklik paketi temel alınarak derlenmiştir.

11.2. Telif ve metin-veri madenciliği

Eğitim derleminin telifli eser içermesi, dünya genelinde süregelen davaların konusudur. Avrupa Birliği hukuku (AB) 2019/790 sayılı Direktif'in 4. maddesiyle metin ve veri madenciliğine bir istisna tanımış, hak sahiplerine bu kullanımı açıkça saklı tutma imkânı vermiştir (Avrupa Birliği, 2019). Tüzük, genel amaçlı model sağlayıcılarını bu saklı tutma beyanlarına uymakla yükümlü kılar.

Türk hukukunda karşılık gelen bir istisna yoktur. 5846 sayılı Fikir ve Sanat Eserleri Kanunu'nun serbest kullanım hükümleri, kişisel kullanım ve iktibas gibi dar hâlleri düzenler; makine öğrenmesi amaçlı toplu çoğaltmayı kapsamaz. TBMM Yapay Zekâ Araştırma Komisyonu, Mart 2026 tarihli raporunda telif mevzuatına bir metin ve veri madenciliği istisnası eklenmesini önermiştir (Türkiye Büyük Millet Meclisi, 2026). Öneri henüz yasama sürecine girmemiştir.

11.3. Kişisel verilerin korunması

Web ölçeğinde toplanan derlemler kişisel veri içerir. 6698 sayılı Kanun bakımından ilk sorun hukuki sebeptir: eğitim amacıyla işlemenin hangi işleme şartına dayandığı açık değildir.

burakoz.com

Meşru menfaat şartı 6698 sayılı Kanun'un 5. maddesinde yer alsa da, ilgili kişinin temel hak ve özgürlüklerine zarar vermemesi koşuluna bağlıdır ve denge testini gerektirir.

İkinci sorun silme hakkının teknik karşılığıdır. Bir kişisel verinin model ağırlıklarından çıkarılması, veri tabanından kayıt silmeye benzemez. Yeniden eğitim tek kesin yöntemdir ve maliyeti çoğu olayda orantısızdır. Makine öğrenmesinde unutturma üzerine çalışmalar sürmektedir; olgunluk düzeyi henüz hukuki bir güvence sunmamaktadır. Kişisel Verileri Koruma Kurumu (2021), yapay zekâ alanında amaçla sınırlılık, veri asgariliği ve tasarımdan itibaren mahremiyet ilkelerini vurgulayan tavsiyeler yayımlamıştır; metin bağlayıcı değildir.

Üçüncü mesele sınır ötesi aktarımdır. Yurt dışında barındırılan bir sağlayıcıya gönderilen her istem bir aktarım işlemidir. 6698 sayılı Kanun'un 2024'te değişen 9. maddesi standart sözleşme ve bağlayıcı şirket kuralları gibi araçlar getirmiştir. Dosya içeriğiyle çalışan hukuk ve bilirkişilik uygulamalarında yerel çalıştırma, bu yükümlülük zincirini büyük ölçüde ortadan kaldırdığı için tercih edilmelidir.

12. Türkiye Bağlamı

Türkiye'de yapay zekâya ilişkin çerçeve bir kanun bulunmamaktadır. 2024 ve 2025 yıllarında TBMM'ye sunulan kanun teklifleri komisyon aşamasını geçememiştir. Yürürlükteki düzenleme, konuya dokunan genel metinlerden oluşur: 6698 sayılı Kanun, 5651 sayılı Kanun, 5846 sayılı Kanun ve 7545 sayılı Siber Güvenlik Kanunu.

Politika düzeyindeki en güncel belge, 18 Ağustos 2026 tarihli Resmî Gazete'de yayımlanan 2026/9 sayılı Genelge ile duyurulan Ulusal Yapay Zekâ Eylem Planı'dır (Cumhurbaşkanlığı, 2026). Plan 2026-2030 dönemi için nicel hedefler koyar: iki yıl içinde beş milyon kişiye temel eğitim, on bin ileri düzey uzman, veri merkezlerinde 1 GW kurulu güç ve on milyar dolarlık yatırım. Enerji hedefi ölçek bakımından dikkate değerdir; Haziran 2026 itibarıyla 126.113 MW olan ülke kurulu gücünün yaklaşık yüzde 0,79'una karşılık gelir.

Yerli model çalışmaları üç eksende ilerlemektedir. TÜBİTAK BİLGEM tarafından geliştirilen BİLGE, 13 Haziran 2026'da tanıtılmış olmakla birlikte genel kullanıma açılmamıştır. T3 Vakfı ile Baykar iş birliğinde yürütülen T3 AI beta aşamasındadır. Ticari tarafta VNGRS tarafından geliştirilen Kumru, 7,4 milyar parametreli bir Türkçe modeldir ve 2 milyar parametreli sürümü açık ağırlıklarla yayımlanmıştır. Ortak kısıt bellidir: Türkçe nitelikli metin stoku, İngilizce derlemlerin yanında küçüktür ve ölçekleme yasalarının işaret ettiği veri hacmine ulaşmak güçtür.

Uluslararası bağlamda bir eksiklik göze çarpar. Avrupa Konseyi'nin Yapay Zekâ ve İnsan Hakları, Demokrasi ve Hukukun Üstünlüğü Çerçeve Sözleşmesi (CETS 225) 5 Eylül 2024'te Vilnius'ta imzaya açılmış ve 1 Kasım 2025'te yürürlüğe girmiştir (Avrupa Konseyi, 2024). Türkiye sözleşmeye taraf değildir. Bu durum, Avrupa Konseyi üyesi bir ülke bakımından değerlendirilmesi gereken bir konumdur.

13. Öneriler

Aşağıdaki öneriler, kurumsal kullanıcılar ile hukuk ve bilirkişilik alanında çalışanlar için ayrı ayrı düzenlenmiştir.

burakolgun.com

Kurumsal dağıtım için

- Gizli belgelerle çalışan iş akışlarında modeli **yerelde çalıştırın**. 4 bit nicemlenmiş 13 milyar parametreli bir model tek bir 24 GB kart üzerinde çalışır ve çoğu belge işleme görevi için yeterlidir.
- Model çıktısını **hiçbir zaman doğrudan** kabuk komutuna, veri tabanı sorgusuna ya da dosya sistemi işlemine bağlamayın. Aradaki katmanda şema doğrulaması yapın.
- Etmen mimarilerinde geri alınamaz işlemleri (silme, ödeme, dışarıya gönderim) insan onayına bağlayın. Komut enjeksiyonuna karşı bilinen tam bir çözüm bulunmadığından, savunma zararı sınırlamaya dayanır.
- Erişim tabanlı mimarilerde yanıtı **kaynak atfı** olmadan sunmayın. Kullanıcının denetleyemediği bir yanıt, denetlenmemiş bir yanıttır.
- Model seçimini genel kıyaslama sıralamalarına değil, kuruma özgü ve cevabı bilinen bir test kümesine dayandırın.

Hukuk ve bilirkişilik uygulaması için

- Model tarafından üretilen hiçbir künye, tarih, sayı ya da içtihat atfını doğrulamadan kullanmayın. Halüsinasyon bir kusur değil, olasılıksal üretimin yapısal sonucudur.
- Yapay zekâ tespit araçlarının yüzdesini teknik bulgu olarak aktarmayın. Sadasivan vd. (2023) bu araçların güvenilirlik sınırlarını göstermiştir; yanlış pozitif oranı gerçek zarar üretecek düzeydedir.
- Üretilmiş içerik iddiasında incelemeyi **üstveri, sürüm geçmişi ve cihaz kayıtları** üzerinden yürütün; metnin üslubu tek başına dayanak olamaz.
- Dosya içeriğini yurt dışı sağlayıcıya göndermeden önce 6698 sayılı Kanun'un 9. maddesindeki aktarım şartlarını değerlendirin; yerel çalıştırma bu zincirin çoğunu ortadan kaldırır.

Düzenleme ve politika için

- Telif mevzuatına metin ve veri madenciliği istisnası eklenmesi, hak sahiplerine saklı tutma imkânı tanınarak değerlendirilmelidir; TBMM komisyon raporunun (2026) önerisi bu yöndedir.
- CETS 225 sayılı Çerçeve Sözleşme'ye katılım, ulusal düzenlemenin uluslararası ölçütlerle uyumu bakımından ele alınmalıdır.
- Kamu kurumlarında model kullanımı için asgari şeffaflık ölçütü getirilmelidir: hangi kararda, hangi model, hangi doğrulama ile kullanılmıştır.

14. Sonuç

Büyük dil modelleri, bir sonraki belirteci tahmin etme hedefinin çok büyük ölçekte uygulanmasından doğmuştur. Bu makalede izlenen çizgi, teknik ayrıntıların büyük bölümünün bu tek hedefin sonuçları olduğunu göstermektedir. Öz-dikkatin karesel maliyeti bağlam penceresinin fiyatını belirler. Belirteçleme sözlüğünün hangi dille kurulduğu Türkçe

burakolgun.com

içeriğin maliyetini belirler. Ölçekleme yasalarındaki denge değişimi hangi modelin dağıtımda ucuz olduğunu belirler.

Halüsinasyon bu çerçevede ayrı bir yere sahiptir. Onu giderilecek bir yazılım kusuru saymak, sorunun kaynağını yanlış konumlandırmaktır. Model olasılık dağılımından örnekleme yapar; biçimi doğru, karşılığı olmayan bir künye üretmesi mekanizmanın kendisidir. Erişim tabanlı mimariler, düşük sıcaklık ve çapraz denetim olasılığı düşürür, sıfırlamaz. Bu nedenle doğrulanmamış çıktının rapora ya da dilekçeye aktarılmaması teknik bir tercih değil, mesleki bir gerekliliktir.

Güvenlik tarafında en kalıcı açık komut enjeksiyonudur; kaynağı mimaridedir, çünkü talimat ile veri aynı kanaldan okunur. Etmen sistemlerin yaygınlaşmasıyla bu açığın sonucu yanlış bir cümle olmaktan çıkıp gerçekleşmiş bir eyleme dönüşmektedir. Savunmanın ağırlığı modelin yanılmasını önlemekten, yanıldığında yapabileceklerini sınırlamaya kaymıştır.

Hukuki çerçevede Avrupa Birliği kademeli bir rejim kurmuş, genel amaçlı model yükümlülüklerini ve şeffaflık kurallarını yürürlüğe koymuştur. Türkiye'de çerçeve kanun hâlâ yoktur ve 2026/9 sayılı Genelge ile açıklanan eylem planı bağlayıcı yükümlülük getirmemektedir. Eğitim verisinde telif, kişisel verinin işlenmesinde hukuki sebep ve model ağırlıklarından silme talebinin karşılanması başlıkları çözümsüz durmaktadır. Adli bilişim bakımından ise bugünün tespit araçları bilirkişi incelemesinde tek başına dayanak oluşturacak güvenilirlikte değildir; inceleme metnin üslubuna değil, doğrulanabilir teknik unsurlara dayandırılmalıdır.

Kaynakça

Ahia, O., Kumar, S., Gonen, H., Kasai, J., Mortensen, D., Smith, N. A., & Tsvetkov, Y. (2023). Do all languages cost the same? Tokenization in the era of commercial language models. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 9904-9923. https://doi.org/10.18653/v1/2023.emnlp-main.614

Avrupa Birliği. (2019). Dijital tek pazarda telif hakkı ve bağlantılı haklara ilişkin (AB) 2019/790 sayılı Avrupa Parlamentosu ve Konsey Direktifi. Avrupa Birliği Resmî Gazetesi, L 130 (17 Mayıs 2019). http://data.europa.eu/eli/dir/2019/790/oj

Avrupa Birliği. (2024). Yapay zekâya ilişkin uyumlaştırılmış kurallar belirleyen (AB) 2024/1689 sayılı Avrupa Parlamentosu ve Konsey Tüzüğü. Avrupa Birliği Resmî Gazetesi, L serisi (12 Temmuz 2024). http://data.europa.eu/eli/reg/2024/1689/oj

Avrupa Konseyi. (2024). Yapay zekâ ve insan hakları, demokrasi ve hukukun üstünlüğüne ilişkin çerçeve sözleşme (CETS No. 225). Avrupa Konseyi Antlaşmalar Serisi. https://www.coe.int/en/web/conventions/full-list?module=treaty-detail&treatynum=225

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610-623. https://doi.org/10.1145/3442188.3445922

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ve diğerleri. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877-1901.

burakolgun.com

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., ve diğerleri. (2021). Extracting training data from large language models. Proceedings of the 30th USENIX Security Symposium, 2633-2650.

Cumhurbaşkanlığı. (2026). Ulusal Yapay Zekâ Eylem Planı (2026-2030) konulu 2026/9 sayılı Genelge. Resmî Gazete (18 Ağustos 2026, S. 33344).

DeepSeek-AI. (2025). DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. arXiv. https://doi.org/10.48550/arXiv.2501.12948

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: Efficient finetuning of quantized LLMs. Advances in Neural Information Processing Systems, 36, 10088-10115.

Frantar, E., Ashkboos, S., Hoefler, T., & Alistarh, D. (2023). GPTQ: Accurate post-training quantization for generative pre-trained transformers. International Conference on Learning Representations (ICLR 2023).

Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., & Fritz, M. (2023). Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security, 79-90. https://doi.org/10.1145/3605764.3623985

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., ve diğerleri. (2022). Training compute-optimal large language models. arXiv. https://doi.org/10.48550/arXiv.2203.15556

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). LoRA: Low-rank adaptation of large language models. International Conference on Learning Representations (ICLR 2022).

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ve diğerleri. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1-38. https://doi.org/10.1145/3571730

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., ve diğerleri. (2024). Mixtral of experts. arXiv. https://doi.org/10.48550/arXiv.2401.04088

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., ve diğerleri. (2020). Scaling laws for neural language models. arXiv. https://doi.org/10.48550/arXiv.2001.08361

Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., & Goldstein, T. (2023). A watermark for large language models. Proceedings of the 40th International Conference on Machine Learning, 17061-17084.

Kişisel Verileri Koruma Kurumu. (2021). Yapay zekâ alanında kişisel verilerin korunmasına dair tavsiyeler. https://www.kvkk.gov.tr

Kudo, T., & Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 66-71. https://doi.org/10.18653/v1/D18-2012

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ve diğerleri. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459-9474.

Meta AI. (2024). The Llama 3 herd of models. arXiv. https://doi.org/10.48550/arXiv.2407.21783

burakolgun.com

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., ve diğerleri. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, 27730-27744.

OWASP. (2025). OWASP Top 10 for large language model applications (2025 sürümü). Open Worldwide Application Security Project. https://owasp.org/www-project-top-10-for-large-language-model-applications/

Petrov, A., La Malfa, E., Torr, P. H. S., & Bibi, A. (2023). Language model tokenizers introduce unfairness between languages. Advances in Neural Information Processing Systems, 36, 36963-36990.

Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., & Finn, C. (2023). Direct preference optimization: Your language model is secretly a reward model. Advances in Neural Information Processing Systems, 36, 53728-53741.

Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., & Feizi, S. (2023). Can AI-generated text be reliably detected? arXiv. https://doi.org/10.48550/arXiv.2303.11156

Sennrich, R., Haddow, B., & Birch, A. (2016). Neural machine translation of rare words with subword units. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 1715-1725. https://doi.org/10.18653/v1/P16-1162

Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., & Dean, J. (2017). Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. International Conference on Learning Representations (ICLR 2017).

Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). AI models collapse when trained on recursively generated data. Nature, 631(8022), 755-759. https://doi.org/10.1038/s41586-024-07566-y

Snell, C., Lee, J., Xu, K., & Kumar, A. (2024). Scaling LLM test-time compute optimally can be more effective than scaling model parameters. arXiv. https://doi.org/10.48550/arXiv.2408.03314

Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., & Liu, Y. (2024). RoFormer: Enhanced transformer with rotary position embedding. Neurocomputing, 568, 127063. https://doi.org/10.1016/j.neucom.2023.127063

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ve diğerleri. (2023). Llama 2: Open foundation and fine-tuned chat models. arXiv. https://doi.org/10.48550/arXiv.2307.09288

Türkiye Büyük Millet Meclisi. (2026). Yapay Zekâ Alanındaki Gelişmelerin İncelenmesi Amacıyla Kurulan Meclis Araştırması Komisyonu Raporu (260 sıra sayılı). TBMM, 28. Dönem.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998-6008.

Villalobos, P., Ho, A., Sevilla, J., Besiroglu, T., Heim, L., & Hobbhahn, M. (2024). Will we run out of data? Limits of LLM scaling based on human-generated data. Proceedings of the 41st International Conference on Machine Learning.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., ve diğerleri. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35, 24824-24837.

burakoguz.com

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). ReAct: Synergizing reasoning and acting in language models. International Conference on Learning Representations (ICLR 2023).

5846 sayılı Fikir ve Sanat Eserleri Kanunu. (1951). Resmî Gazete (13 Aralık 1951, S. 7981).

6698 sayılı Kişisel Verilerin Korunması Kanunu. (2016). Resmî Gazete (7 Nisan 2016, S. 29677).

7545 sayılı Siber Güvenlik Kanunu. (2025). Resmî Gazete (19 Mart 2025, S. 32846).

burakolgun.com