HAFTALIK YAPAY ZEKA ÖZETİ

#2

Bu haftanın en önemli 6 gelişmesi

Açık ağırlıklı yeni kodlama modelleri, Google'ın en güçlü modeli, bilgisayarında çalışan yapay zeka ve tek GPU'da dakikalık video üreten dünya modeli. Hepsi tek yerde, sade ve doğrulanmış şekilde.

01 DeepSeek V4.1-Flash

02 Locally Uncensored

03 Gemini 4 Argon

04 Reflection AI Beam

05 PentAGI

06 NVIDIA SANA-WM

Ön koşullar ve erişim

Bu haftaki gelişmelerin bir kısmı ücretsiz ve açık, bir kısmı ise ücretli plan ya da güçlü donanım istiyor. Denemeye başlamadan önce hangisinin sana uygun olduğuna bu tablodan bak.

Gelişme	Ne gerekiyor?	Nerede çalışır?
DeepSeek V4.1-Flash	Ağırlıklar ücretsiz. Barındırılan API kullanım başına ücretli.	API üzerinden her yerde. Kendi sunucunda çalıştırmak çok güçlü donanım ister.
Locally Uncensored	Ücretsiz, hesap gerekmez.	Windows ve Linux masaüstü. macOS masaüstü sürümü yok.
Gemini 4 Argon	Uygulamada Google AI Pro veya Ultra aboneliği. Geliştiriciler için ücretli API.	Gemini uygulaması, web ve Google Arama'daki AI Mode. Ücretsiz kullanıcılara açık değil.
Reflection AI Beam	Ağırlıklar ücretsiz. API şu an beta aşamasında.	API üzerinden. Kendi sunucunda çalıştırmak çoklu GPU gerektirir.
PentAGI	Ücretsiz ve açık kaynak. Kendi seçtiğin model sağlayıcısının API anahtarı ya da Ollama.	Docker kurulu bir bilgisayar veya sunucu.
NVIDIA SANA-WM	Ücretsiz ve açık kaynak.	Güçlü bir NVIDIA GPU. Geliştiriciye yönelik, tıkla-çalıştır bir uygulama değil.

Fiyatlar ve plan kapsamları sık değişiyor. Ücretli bir şey kullanmadan önce güncel durumu kendi hesabında kontrol et.

Terimler

- **Açık ağırlık (open weights):** Modelin kendisi indirilebilir. İstersen kendi sunucunda çalıştırır, üzerine ince ayar yaparsın.
- **Mixture-of-Experts (MoE):** Model çok büyük ama her kelimede sadece küçük bir kısmı çalışıyor. Bu yüzden büyük modelin kalitesini daha düşük maliyetle almak mümkün oluyor.
- **Token:** Modelin metni böldüğü küçük parçalar. Fiyatlar ve bağlam sınırları token üzerinden hesaplanır.
- **Terminal-Bench / SWE-Bench:** Modellerin terminalde ve gerçek yazılım projelerinde ne kadar iyi iş çıkardığını ölçen testler.

DeepSeek V4.1-Flash açık ağırlıklarla geldi

552 milyar parametreli, 1 milyon token bağlam penceresine sahip ve görselleri de anlayan yeni DeepSeek modeli. Ağırlıkları herkes indirebiliyor.

Ne oldu?

DeepSeek, V4.1-Flash'ı Eylül 2026'da duyurdu. Şirkete göre model, önceki amiral gemisi V4-Pro'yu performans, maliyet ve hız tarafında geride bırakıyor. Hatta DeepSeek, V4-Pro'ya gelen istekleri otomatik olarak V4.1-Flash'a yönlendirmeye başladı. Yeni bir V4.1-Pro için henüz tarih yok.

Öne çıkanlar

- **Küçük aktif boyut:** Model girdiyi okurken yaklaşık 8 milyar, cevap yazarken yaklaşık 16 milyar parametre kullanıyor. Toplam boyuta göre çok küçük, bu da onu ucuz ve hızlı yapıyor.
- **Uzun bağlam:** 1 milyon tokene kadar girdi alabiliyor. Büyük bir kod tabanını ya da uzun dokümanları tek seferde verebilirsin.
- **Çok modlu:** Görselleri de anlıyor. DeepSeek, karmaşık görselleri okumada en iyi kapalı modellerin gerisinde kaldığını kendisi de kabul ediyor.
- **Lisans:** MIT. Ticari projede kullanabilir, değiştirebilir, kendi sunucunda çalıştırabilirsin.

DeepSeek'in paylaştığı test sonuçları

Aşağıdaki rakamlar DeepSeek'in kendi tablosundan. Bağımsız testlerde sonuçlar farklı çıkabilir.

Test	V4.1-Flash	Kimi K3	GLM-5.3	Opus 5	GPT-5.6 Sol
Terminal-Bench 3.0	30.0	17.7	28.3	**43.3**	34.4
DeepSWE v1.1	**74.2**	67.5	66.9	74.0	73.0
CyberGym	**88.1**	80.0	84.5	84.5	84.5
Automation-Bench	**54.8**	46.7	48.8	50.3	45.8

Kısacası: yazılım mühendisliği, siber güvenlik ve otomasyon testlerinde kapalı modellerle başa baş ya da önde, Terminal-Bench 3.0'da ise Opus 5 ve GPT-5.6 Sol'un gerisinde. Uygun fiyatlı, uzun bağlamlı bir model arayan geliştiriciler için API üzerinden denemeye değer.

Dene: openrouter.ai/deepseek/deepseek-v4.1-flash

Locally Uncensored: yerel yapay zeka tek uygulamada

Sohbet, görsel üretimi, video üretimi ve kodlama ajanını tek bir masaüstü uygulamasında toplayan, ücretsiz ve açık kaynak bir proje. Her şey kendi bilgisayarında çalışıyor.

Neden önemli?

Yerel model çalıştırmak normalde terminal, Docker ve ayar dosyalarıyla uğraşmak demek. Bu uygulama kurulum sihirbazıyla bunların çoğunu senin yerine hallediyor ve verin bilgisayarından çıkmıyor.

Dört sekme, aynı modeller

- **Chat:** Yerel modellerle sohbet, görsel yükleme, kendi dokümanlarınla sohbet ve sesli kullanım.
- **Create:** Görsel ve video üretimi. Arka planda ComfyUI kullanıyor ve onu da uygulama kuruyor.
- **Code:** Kodlama ajanı. Bir değişikliği uygulamadan önce farkı (diff) gösteriyor, sonra testleri çalıştırıyor.
- **Agent:** Web araması, dosya erişimi ve komut çalıştırma. Her araç çağrısında senden izin istiyor.

Nasıl kurulur?

- **1. İndir:** Projenin GitHub sayfasındaki Releases bölümünden işletim sistemine uygun kurulum dosyasını indir.
- **2. Kur ve aç:** İlk açılışta sihirbaz, bilgisayarında zaten bir yerel AI motoru (Ollama, LM Studio gibi) varsa onu buluyor. Yoksa kendi motorunu kuruyor ve ComfyUI'ı hazırlıyor.
- **3. Model seç:** Model yöneticisi, hangi modellerin senin donanımına sığdığını işaretliyor. Başlangıç için işaretli küçük bir modelle dene.

Donanım ihtiyacı

Kullanım	Yaklaşık ihtiyaç
Sohbet	3B model: 8 GB RAM, ekran kartı gerekmez. 8B model: 6 GB'lık ekran kartı.
Görsel üretimi	6–8 GB VRAM ile başlar, bazı modeller 16–24 GB ister.
Video üretimi	Hafif modeller 6–8 GB, Wan ve HunyuanVideo gibi modeller 12 GB ve üstü VRAM ister.

> **Platform:** Windows 10/11 ve Linux destekleniyor, macOS masaüstü sürümü yok. Lisans: AGPL-3.0.

Dene: github.com/PurpleDoubleD/locally-uncensored

Google, Gemini 4 Argon'u herkese açtı

Google'ın şu ana kadarki en yetenekli modeli 13 Ekim'de geniş erişime açıldı. Uzun kodlama, ajan görevleri, hukuk ve finans işleri için tasarlandı.

Nerelerde kullanabilirsin?

- **Gemini uygulaması ve AI Mode:** Google AI Pro ve Ultra abonelerine açık. Ücretsiz kullanıcılar şimdilik erişemiyor.
- **Antigravity:** Google'ın ajan tabanlı kodlama ortamında, AI Ultra aboneliğiyle.
- **Gemini API:** Geliştiriciler için ücretli API. 30 Kasım'a kadar tanıtım fiyatı uygulanıyor, sonrasında fiyat yükseliyor. Güncel fiyatı Google'ın fiyat sayfasından kontrol et.

Teknik olarak ne farklı?

- **1 milyon token bağlam:** Girdi, düşünme ve çıktı aynı 1 milyonluk bütçeyi paylaşıyor.
- **Çok uzun çıktı:** Normal bir cevap 65 bin tokene kadar çıkabiliyor. "Long decode" denen modla üretim parça parça devam ederek bağlam sınırına kadar uzayabiliyor.
- **Görev modu:** Sohbet penceresindeki bir geçişle işi Gemini'ye devredebiliyorsun. Sen bilgisayar başında değilken çalışmaya devam ediyor.

Google'ın paylaştığı sonuçlar

Aşağıdaki skorlar Google'ın kendi açıkladığı rakamlar. Google'ın tablosuna göre Argon, GPT-6 Astra'yı karşılaştırılan testlerin çoğunda geçiyor. FrontierSWE v2 gibi bazı testlerde ise Astra hâlâ önde.

Test	Ne ölçüyor?	Skor
DeepSWE v1.1	Uzun soluklu yazılım mühendisliği görevleri	77.9%
Zapier AutomationBench	İş akışı otomasyonu	51.3%
LVBench	Uzun videoları anlama	91.7%
CWE-bench v1	Güvenlik açıklarını düzeltme	68%

Kimin işine yarar?

- **Zaten Google AI Pro/Ultra kullananlar:** Ek bir şey yapmana gerek yok. Gemini uygulamasında model seçicisinden Argon'u seç.
- **Uzun doküman ve kod ile çalışanlar:** Bir projenin tamamını ya da yüzlerce sayfalık bir dokümanı tek seferde verip üzerinde çalışmak için uygun.

> Özellikler hesaplara kademeli açılabilir. Model seçicide Argon'u göremiyorsan planını ve uygulama sürümünü kontrol et.

Dene: gemini.google.com

Reflection AI, Beam'in ağırlıklarını yayınladı

501 milyar parametreli kodlama ve ajan modeli Beam, Apache 2.0 lisansıyla açık ağırlıklı olarak geldi. Batı'dan çıkan en iddialı açık modellerden biri olarak konumlanıyor.

Neden konuşuluyor?

Son dönemde açık ağırlıklı model yarışına büyük ölçüde Çinli laboratuvarlar (DeepSeek, Kimi, GLM) yön veriyordu. NVIDIA destekli Reflection AI, Beam ile bu alanda ABD merkezli güçlü bir alternatif sunmayı hedefliyor. Beam, şirketin ilk açık ağırlıklı modeli ve bir serinin ilki.

Öne çıkanlar

- **Mixture-of-Experts:** 501 milyar parametrenin her tokende yalnızca 23 milyarı aktif.
- **Kodlama skorları:** Reflection'a göre SWE-Bench Verified'da 80.9, Terminal-Bench 2.1'de 80.1 puan alıyor. Terminal-Bench'te Kimi K3, GLM-5.3 ve DeepSeek V4.1-Flash hâlâ önde.
- **Verimlilik:** Reflection, Beam'in benzer skorlara rakiplerine göre çok daha az token ve işlem gücüyle ulaştığını söylüyor. Bu, aynı işi daha ucuza yaptırmak anlamına geliyor.
- **Bağlam:** Model 1 milyon tokene kadar bağlam için eğitildi. Beta API'de istek başına sınır daha düşük tutuluyor.

Nasıl denersin?

- **API ile:** platform.reflection.ai üzerinden beta API'ye kaydol. API, OpenAI uyumlu Chat Completions formatını destekliyor. Yani mevcut OpenAI kodunda sadece adresi ve model adını değiştirerek deneyebilirsin.
- **Kendi sunucunda:** Ağırlıklar Hugging Face'te. 8-bit veya 4-bit sıkıştırılmış hâlleri sekiz GPU'lu bir sunucuya sığıyor. Kişisel bilgisayar için uygun değil.

Kimin işine yarar?

Kurumsal tarafta lisans ve menşe hassasiyeti olan, ama açık ağırlıklı ve güçlü bir kodlama modeli arayan ekipler için önemli bir seçenek. Bireysel kullanıcı için en kolay yol API üzerinden denemek.

> API şu an beta aşamasında. Ücretlendirme ve erişim koşulları değişebilir, kullanmadan önce platform sayfasında güncel durumu kontrol et.

Dene: platform.reflection.ai

PentAGI: kendi sistemine sızma testi yapan AI ajanları

Yetkili olduğun sistemlerde sızma testi (pentest) yapan, kendi sunucunda çalışan, açık kaynak bir yapay zeka ajan sistemi.

> **Önemli:** Sızma testini yalnızca sahibi olduğun ya da yazılı izin aldığın sistemlerde yap. İzinsiz test, Türkiye dahil birçok ülkede suç teşkil eder.

Nasıl çalışıyor?

Ona bir hedef ve amaç veriyorsun. Sistem işi planlıyor, araçları çalıştırıyor ve bulduklarını kaydediyor. Bunu tek bir model değil, görevleri paylaşan bir ajan ekibi yapıyor:

- **Orkestratör:** Hedefi alıyor, işi parçalara bölüyor ve diğer ajanlara dağıtıyor.
- **Araştırmacı:** Bilinen güvenlik açıklarını ve ilgili kaynakları araştırıyor.
- **Geliştirici:** Saldırı stratejisini planlıyor.
- **Uygulayıcı:** Komutları izole konteynerlerde çalıştırıyor.

Öne çıkanlar

- **İzole çalışma:** Tüm işlemler Docker konteynerlerinde yürüyor. Varsayılan imaj Kali Linux ve içinde nmap, Metasploit, sqlmap gibi 20'den fazla güvenlik aracı hazır geliyor.
- **Hafıza:** Bulgular vektör veritabanında saklanıyor. Sonraki testlerde önceki bulgular kullanılabiliyor.
- **Modelini sen seç:** OpenAI, Anthropic, Google Gemini, AWS Bedrock, DeepSeek, OpenRouter ve yerel modeller için Ollama destekleniyor. Verinin dışarı çıkmasını istemiyorsan Ollama ile tamamen yerel çalıştırabilirsin.
- **Entegrasyon:** REST ve GraphQL API'leri var. CI/CD hattına bağlayıp düzenli test yaptırabilirsin.

Ön koşullar

Gereksinim	Detay
Kurulum yöntemi	Docker Compose. Etkileşimli kurulum aracı veya manuel yapılandırma.
Minimum donanım	2 vCPU, 4 GB RAM, 20 GB disk.
Model erişimi	Seçtiğin sağlayıcının API anahtarı veya yerel Ollama.
Lisans	MIT. GitHub'da 30 binin üzerinde yıldıza sahip.

Proje, gerçek kullanımda çalışan konteynerlerin ayrı bir sunucuda tutulmasını öneriyor. Böylece test sırasında çalışan güvenilmeyen kod, ana sistemine dokunmuyor.

Dene: pentagi.com

NVIDIA, SANA-WM dünya modelini açık kaynak yaptı

Bir görsel, bir metin ve bir kamera rotası veriyorsun. Model, kameranın o rotayı izlediği bir dakikalık 720p video üretiyor. Hem de tek GPU'da.

Dünya modeli ne demek?

Normal video modelleri bir sahneyi "çeker". Dünya modelleri ise sahnenin içinde dolaşmanı sağlar: kamerayı nereye götürürsen sahne ona göre oluşur. Bu yüzden oyun, robotik ve simülasyon için önemli bir yapı taşı olarak görülüyor.

Öne çıkanlar

- **Küçük ama uzun:** 2.6 milyar parametreli. Doğrudan bir dakikalık video üretmek için eğitildi.
- **Hassas kamera kontrolü:** Kamera altı eksende (ileri-geri, sağa-sola, yukarı-aşağı ve dönüşler) hareket edebiliyor. Model, verdiğin rotayı bir dakika boyunca takip ediyor.
- **Tek GPU'da hızlı:** NVIDIA'nın makalesine göre damıtılmış (distilled) sürüm, tek bir RTX 5090'da 60 saniyelik 720p bir klibi yaklaşık 34 saniyede üretebiliyor.
- **Verimli:** NVIDIA, benzer kapalı sistemlere göre 36 kata kadar daha yüksek üretim hızı bildiriyor.

Nasıl denersin?

SANA-WM, bir uygulama değil, geliştiricilere yönelik bir model. Hugging Face sayfasındaki örnek kodla çalıştırılıyor. Girdi olarak bir başlangıç görseli, bir metin açıklaması ve kamera hareketi veriyorsun. Kamera hareketini klavye tarzı komutlarla (WASD gibi) ya da kamera matrisleriyle tanımlayabiliyorsun.

Ön koşullar

Gereksinim	Detay
Donanım	Güçlü bir NVIDIA GPU. Tam iş akışı (iyileştirici model ve metin kodlayıcı dahil) onlarca GB'lık model dosyası indiriyor.
Bilgi	Python ve komut satırı kullanmaya alışkın olmak gerekiyor.
Lisans	Apache 2.0. Birlikte kullanılan LTX-2 bileşenleri kendi lisanslarına tabi.
Çıktı	Sabit 720p çözünürlük. Girdi görseli bu boyuta göre kırpılıyor.

Kimin işine yarar?

- **Oyun ve 3D tarafında çalışanlar:** Bir konsept görselden gezilebilir sahne denemeleri yapmak için.
- **Robotik ve simülasyon:** Büyük bir GPU kümesine ihtiyaç duymadan eğitim ve test ortamı üretmek için.

Dene: huggingface.co/Efficient-Large-Model/SANA-WM_bidirectional

Bu haftanın büyük resmi

Altı gelişmeyi yan yana koyunca üç net eğilim görünüyor.

1. Açık modeller kapalı modellere yetişiyor

DeepSeek V4.1-Flash ve Beam, bazı kodlama ve ajan testlerinde kapalı amiral gemisi modellerle başa baş skorlar açıklıyor. Fark artık "açık model mi kapalı model mi" değil, hangi görevde hangisinin daha uygun ve ucuz olduğu. Yine de bu skorların çoğu şirketlerin kendi testleri, bağımsız sonuçları beklemek mantıklı.

2. Büyük model, küçük maliyet

Bu haftaki büyük modellerin neredeyse hepsi Mixture-of-Experts mimarisi kullanıyor. Toplamda yüzlerce milyar parametre var ama her adımda sadece küçük bir kısmı çalışıyor. Sonuç: daha iyi kalite, daha düşük fiyat. Bağlam pencerelerinde de 1 milyon token yeni standart olmaya başladı.

3. Yapay zeka senin bilgisayarına iniyor

Locally Uncensored ve PentAGI gibi araçlar, güçlü yapay zeka iş akışlarını kendi bilgisayarında veya sunucunda çalıştırmayı kolaylaştırıyor. SANA-WM ise daha önce büyük GPU kümeleri isteyen video dünya modellerini tek GPU'ya taşıyor. Gizlilik ve maliyet kontrolü isteyenler için seçenekler hızla artıyor.