HAFTALIK YAPAY ZEKA ÖZETİ

#5

Bu haftanın en önemli 6 gelişmesi

Laptopunda internetsiz çalışan 20 GB'lık bir açık model, kayıttaki arka plan sesini silen ücretsiz bir araç, 4,5 milyon parametreli bir metinden sese modeli, videoyu sahne sahne çözen bir araç, araştırma makalesi yazan bir agent ekibi ve OpenAI'ın iki yeni modeli. Hepsi tek yerde, sade ve doğrulanmış şekilde.

6	Paper Orchestra	5	Voice Isolator
4	GPT-6.2 Astra ve Sol	3	tiny-voca
2	rea-video	1	Inkling 2 Mini

Ön koşullar ve erişim

Bu haftaki altı gelişmeden beşi ücretsiz ve açık kaynak. Asıl fark, hangi donanımı istediği ve arkasında ücretli bir servis olup olmadığı. Denemeye başlamadan önce sana uygun olanı bu tablodan seç.

Gelişme	Ne gerekiyor?	Nerede çalışır?
Paper Orchestra	Kod ücretsiz (Apache 2.0). Agent'ları çalıştıran model (Gemini ya da OpenAI) API üzerinden ücretli.	Python 3.11 ortamı, API anahtarları ve hedef konferansın LaTeX şablonu.
Voice Isolator	Ücretsiz. Kod ve ağırlıklar Apache 2.0.	Python 3.10+, PyTorch 2.1+. NVIDIA GPU, Apple silicon ya da işlemci.
GPT-6.2 Astra ve Sol	Ücretli. ChatGPT Plus, Pro, Enterprise ve Edu planlarında, API'de token başına.	ChatGPT Work, Codex ve OpenAI API.
tiny-voca	Ücretsiz. Kod ve ağırlıklar Apache 2.0, veri seti CC BY 4.0.	Sıradan işlemci ya da tarayıcı. PyTorch veya ONNX Runtime.
rea-video	Ücretsiz (Apache 2.0). İstersen bağlayacağın video modeli (Gemini, Doubao) ayrı ücretli.	Claude Code, Codex ya da komut çalıştırabilen bir agent. ffmpeg ve Python 3.10+.
Inkling 2 Mini	Ücretsiz, açık ağırlık, Apache 2.0.	32 GB ve üstü bellekli Apple silicon Mac, DGX Spark ya da 32 GB+ VRAM'li GPU.

Ücretsiz araçlar da ücretli servislere bağlanabiliyor. Bir modeli API üzerinden kullanacaksan güncel fiyatı kendi hesabında kontrol et.

Terimler

- **AI agent (ajan):** Sadece cevap vermeyen, araç kullanıp adım adım iş yapan yapay zeka.
- **Açık ağırlık:** Modelin kendisi indirilebilir. Kendi bilgisayarında çalıştırabilir, üzerine eğitim yapabilirsin.
- **Parametre:** Modelin eğitimde öğrendiği sayılar. Kabaca modelin büyüklüğünü gösterir.
- **MoE (uzman karışımı):** Her token'da modelin sadece bir kısmının çalıştığı mimari. Toplam parametre büyük, çalışan kısım küçük.
- **Looped (döngülü) transformer:** Bazı katmanları aynı ağırlıklarla birden fazla kez çalıştıran model. Daha az ağırlıkla daha fazla hesap yapar.
- **Gerçek zamandan 70 kat hızlı:** 1 saniyelik sesi yaklaşık 1/70 saniyede üretmek.
- **SWE-bench Verified:** Modelin gerçek GitHub hatalarını düzeltip düzeltemediğini ölçen kodlama testi.

6 numara

Paper Orchestra: fikirden kaynaklı araştırma makalesine

Google Cloud AI Research ekibinin açık kaynak çerçevesi. Fikir notlarını ve deney kayıtlarını veriyorsun, beş uzman agent'lık bir ekip konferans şablonuna uygun bir LaTeX makale çıkarıyor.

Nasıl çalışıyor?

Ham malzemeni (fikir ve deney kayıtları) bir klasöre koyuyorsun ve hedef konferansın LaTeX şablonunu seçiyorsun. Gerisini bir orkestra gibi sırayla çalışan agent'lar yapıyor:

- **Taslak:** Makalenin iskeletini çıkarıyor.
- **Literatür taraması:** İlgili makaleleri arıyor, öncelik sırasına koyuyor ve metne atıflarıyla yerleştiriyor.
- **Bölüm yazımı:** Bölümleri tek tek yazıyor.
- **Düzeltme:** Yazılan içeriği gözden geçirip iyileştiriyor.
- **Grafik:** Deney verisinden kodla grafik üretiyor, kavram şemalarını da ekleyebiliyor.

Nasıl kurulur?

```
git clone https://github.com/google-research/paper-orchestra.git
cd paper-orchestra
conda create -n paper_orchestra python=3.11
conda activate paper_orchestra
pip install -r requirements.txt
```

Sonra proje kökündeki `.env` dosyasına model anahtarını (`GEMINI_API_KEY` ya da `OPENAI_API_KEY`) ve `SEMANTIC_SCHOLAR_API_KEY` yaz, makaleyi başlat:

```
python paper_writing_cli.py \
  --raw_materials_dir ./malzemeler/ \
  --latex_template_dir templates/cvpr2025 \
  --output_dir ./cikti/ --use_plotting true
```

Ön koşullar

Gereksinim	Detay
Ücret	Kod ücretsiz, Apache 2.0 lisanslı. Agent'ları çalıştıran model API üzerinden ücretli.
Yazılım	Conda ile Python 3.11 ortamı ve konferansın LaTeX şablonu (repoda örnekler var).

> Araç deneyi senin yerine yapmıyor; kendi fikir notların ve deney kayıtların gerekiyor. README'ye göre Google'ın resmi olarak desteklediği bir ürün değil. Literatür taraması atıf ekliyor ama makaleyi göndermeden önce her atıfı ve sonucu kendin kontrol et.

Dene: github.com/google-research/paper-orchestra · arxiv.org/abs/2604.05018

5 numara

Voice Isolator: kayıttaki arka plan sesini sil, sesini koru

Inworld'ün 18 Kasım'da açık kaynak yaptığı ses ayırma modeli. Arka plandaki konuşmaları, müziği ve gürültüyü siliyor; sesini tınısı, nefesleri ve ses seviyesiyle olduğu gibi bırakıyor.

Ne işe yarıyor?

Gürültülü bir kaydı veriyorsun, iki dosya geri alıyorsun: temizlenmiş ses ve silinen arka plan. Inworld'e göre model gülme, nefes ve iç çekme gibi sesleri de koruyacak şekilde eğitildi. Videoları yayınlamadan önce temizlemek, ses veri seti hazırlamak ya da canlı görüşmede gürültüyü kesmek için kullanılabiliyor.

İki model

Model	Ne için?	Boyut
voice-isolator-v1	Kayıtlı ses, en iyi kalite. Her an 8 saniyelik parçanın tamamını görüyor.	64M parametre
voice-isolator-v1-realtime	Canlı ses ve yayın. 20 ms ileriye bakarak akış hâlinde çalışıyor.	58M parametre

Nasıl kurulur?

```
pip install inworld-voice-isolation
voice-isolate kayit.wav kayit_temiz.wav
```

Model ilk kullanımda Hugging Face'ten iniyor. Temiz ses `kayit_temiz.wav`, silinen arka plan `kayit_temiz_background.wav` olarak yanına yazılıyor; çıktılar 48 kHz. Canlı mod için `--mode streaming` ekliyorsun.

Ön koşullar

Gereksinim	Detay
Ücret	Ücretsiz. Kod ve ağırlıklar Apache 2.0, ticari kullanım serbest.
Yazılım	Python 3.10+ ve PyTorch 2.1+.
Donanım	NVIDIA GPU varsa onu, Apple silicon Mac'te Metal'i (PyTorch 2.6+), yoksa işlemciyi kullanıyor. Küçük GPU'lar henüz test edilmedi.

> Konuşmacı seçici değil: birkaç kişiden birini ayırmıyor. Kötü mikrofonun kaybettiği detayı da geri getirmiyor; Inworld bunu ikinci sürüme planlıyor. Paylaşılan kalite skorları Inworld'ün kendi, dağıtılmayan test setinden.

Dene: github.com/inworld-ai/voice-isolator · huggingface.co/inworld/voice-isolator

4 numara

GPT-6.2 Astra ve Sol: OpenAI'ın yeni modelleri Codex'te

OpenAI iki yeni model yayınladı: en zor işler için Astra, günlük iş için daha verimli Sol. İkisi de ChatGPT Work'te, Codex'te ve API'de.

Ne oldu?

- **Erişim:** ChatGPT Plus, Pro, Enterprise ve Edu kullanıcıları iki modeli de ChatGPT Work ve Codex'te kullanabiliyor.
- **Sol neden önemli?:** OpenAI'a göre Cognition'ın FrontierCode 2.0 testinde Sol, Astra'ya 0,5 puan kadar yaklaşıyor ve maliyeti yaklaşık beşte biri (ikisi de en yüksek ayarda).
- **Daha az token:** OpenAI'ın GDPval v2 testinde Astra, GPT-6 Astra'dan %22 daha az çıktı token'ı kullanıyor.
- **Ultrafast:** Pro 500 abonelerinde iki modelde, Pro 200'de önce Sol'da açık. Enterprise ve Edu'da yönetici açınca geliyor.

API fiyatları (1 milyon token başına)

Model	Girdi	Önbellekli girdi	Çıktı
gpt-6.2-astra	8 dolar	0,40 dolar	40 dolar
gpt-6.2-sol	2 dolar	0,10 dolar	10 dolar

Kısacası: Astra'nın girdi ve çıktı fiyatı GPT-6 Astra'dan %20 düşük; Sol, GPT-6.1 Sol ile aynı fiyatta.

Nasıl denersin?

Ücretli bir ChatGPT planın varsa Codex'te modeli Sol ya da Astra olarak seçmen yeterli. API'de model adları `gpt-6.2-astra` ve `gpt-6.2-sol`. Kodlama işlerinde önce Sol'u dene; OpenAI'ın kendi rakamlarına göre fark küçük, maliyet farkı büyük.

> Bütün test sonuçları OpenAI'ın kendi açıklaması. Bağımsız bir karşılaştırma henüz yok.

Dene: openai.com/index/introducing-gpt-6-2-astra-sol

3 numara

tiny-voca: 4,5 milyon parametreli metinden sese modeli

Sıfırdan eğitilmiş, iki sesli küçük bir İngilizce seslendirme modeli. Sıradan bir işlemcide ya da doğrudan tarayıcıda çalışıyor.

Öne çıkanlar

- **Minicik:** Tam 4.537.291 parametre, ses üreticisiyle birlikte 18 MB.
- **Hızlı:** Geliştiriciye göre dört işlemci iş parçacığında gerçek zamandan yaklaşık 70 kat hızlı.
- **İki ses:** F2 (kadın) ve M2 (erkek), Amerikan İngilizcesi, 24 kHz.
- **Her yerde:** PyTorch, ONNX Runtime ve tarayıcıda ONNX Runtime Web. Kütüphanesiz düz JavaScript denemesi de var.
- **Tamamen açık:** Kod, ağırlıklar, eğitim tarifi ve veri seti. Bütün proje tek bir kiralık RTX 4090'da 9,56 dolarlık GPU süresine mal oldu.

Nasıl kurulur?

Hiçbir şey kurmadan denemek için tarayıcı demosunu aç: tiny-voca.n4ze3m.com. Kendi bilgisayarında çalıştırmak için (Windows'ta Git Bash ile):

```
git clone https://github.com/n4ze3m/tiny-voca.git
cd tiny-voca
pip install -e ".[own]"
hf download n4ze3m/tiny-voca --local-dir weights
export DIZIONE_DATA=weights/phonemizer
python -m tiny_voca.infer --checkpoint weights/tiny-voca.pt \
  --voice F2 --text "Hello! This is tiny voca." --output hello.wav
```

`--speed` ile konuşma hızını, `--pitch-shift` ile ses perdesini ayarlayabiliyorsun. Uzun metni cümle cümle okuyup birleştiriyor.

Ön koşullar

Gereksinim	Detay
Ücret	Ücretsiz. Kod ve ağırlıklar Apache 2.0, veri seti CC BY 4.0.
Donanım	Sıradan bir işlemci yeterli, GPU gerekmiyor.
Türkçe	Model İngilizce. Topluluğun yaptığı Türkçe sürümler README'de listeleniyor (zerre-tts, clarvia-tts-tiny).

> Sınırları: yalnız İngilizce, iki sabit ses, ses klonlama yok, henüz akış (streaming) yok. Yabancı isimleri ve alışılmadık yazımları sık yanlış okuyor.

Dene: github.com/n4ze3m/tiny-voca · huggingface.co/n4ze3m/tiny-voca

2 numara

rea-video: herhangi bir videoyu sahne sahne çöz

Bir videoyu veriyorsun; her sahnenin gerçekten çekilmiş mi, yapay zekayla mı üretilmiş, yoksa kodla mı yapılmış olduğunu söylüyor ve yeniden yapman için bir plan çıkarıyor.

Ne işe yarıyor?

rea-video videoyu sahnelere bölüyor, sesi konuşma, müzik ve efekt olarak ayırıyor. Her sahne için kanıt topluyor (kare düzeni, sensör gürültüsü, dosyadaki yapay zeka etiketleri, eller ve yazılar) ve karar veriyor: çekim, AI video, avatar, 3D/animasyon, kodla yapılmış şablon ya da ekran kaydı.

- **Yeniden yapım planı:** Çekim sahneleri için kamera ve ışık planı, AI sahneleri için sahne başına prompt, kod sahneleri için Remotion yaklaşımı.
- **Altyazı ve kurgu:** Altyazıların zamanlarını SRT/ASS olarak çıkarıyor; kesimlerin yerini, hızını ve müziğe oturup oturmadığını gösteriyor.
- **Ses:** İnsan sesi mi AI sesi mi, müziğin BPM'i ve her efektin nereye düştüğü.
- **İki çıktı:** Senin okuman için bir rapor (HTML/PDF) ve agent'ın katman katman okuduğu VIDEO.md.

Nasıl kurulur?

Claude Code ya da Codex'e şunu yazman yeterli; agent aracı kurup analizi kendisi yapıyor:

```
Install https://github.com/JuneYaooo/rea-video, then break down this video:
/path/to/video.mp4
```

```
# ya da Claude Code eklentisi olarak:
claude plugin marketplace add JuneYaooo/rea-video
claude plugin install rea-video@rea-video
```

Ön koşullar

Gereksinim	Detay
Ücret	Ücretsiz, Apache 2.0 lisanslı. Yalnız araçlarla çalışan mod internetsiz ve ücretsiz.
Yazılım	ffmpeg ve Python 3.10+. Claude Code, Codex ya da komut çalıştırabilen bir agent.
İsteğe bağlı	Video anlayan bir model (Gemini ya da Doubao) bağlarsan kararlar daha isabetli oluyor ve her sahneye prompt ekleniyor; ücretini o servis alıyor.

> Platformlardan indirilen videolar yeniden sıkıştırıldığı için kanıtların bir kısmı siliniyor: geliştiricinin kendi testinde yalnız araçlarla isabet bu videolarda %88,6'ya düşüyor, görsel kontrolle %99'un üstüne çıkıyor. Projenin hiçbir kripto token'ı yok.

Dene: github.com/JuneYaooo/rea-video

1 numara

Inkling 2 Mini: laptopta internetsiz çalışan açık model

Mira Murati'nin kurduğu Thinking Machines Lab, 20 Kasım'da Inkling 2 Mini'yi Apache 2.0 lisanslı açık ağırlıklarla yayınladı. Yaklaşık 20 GB'lık tek bir dosya; kendi bilgisayarında, internete bağlanmadan çalıştırabiliyorsun.

Ne oldu?

Inkling 2 Mini, şirketin yayınladığı ilk looped transformer. Toplam 26,4 milyar parametresi var, her token'da 3,4 milyarı çalışıyor. Ortadaki 12 katman aynı ağırlıklarla üç kez çalıştığı için token başına etkin çalışan parametre 5,4 milyara çıkıyor. Metin ve görsel okuyor, metin yazıyor; bağlam penceresi 1 milyon token.

Thinking Machines'in paylaştığı test sonuçları

Test	Inkling 2 Mini	Qwen3.8-27B	Gemma 4 31B
SWE-bench Verified (yalnız bash)	**79,2**	77,8	46,1
SWE-bench Pro v2 (Zor)	**77,9**	73,8	14,2
BrowseComp	**79,0**	61,6	37,0
Artificial Analysis Index v5.0	**20,1**	**20,1**	4,4

Kısacası: şirketin kendi testlerinde kodlamada aynı boydaki açık modellerin önünde. Artificial Analysis'in bağımsız endeksinde Qwen3.8-27B ile berabere.

Nasıl denersin?

32 GB ya da daha fazla belleği olan bir Apple silicon Mac'te en kolay yol Ollama:

```
ollama run inkling-2-mini
```

vLLM, SGLang, llama.cpp ve Unsloth'un GGUF sürümleri de destekleniyor.

Ön koşullar

Gereksinim	Detay
Ücret	Ücretsiz, açık ağırlık, Apache 2.0. Ticari kullanım serbest.
Donanım	32 GB ve üstü bellekli Apple silicon (MacBook, Mac mini, Mac Studio), bir NVIDIA DGX Spark ya da 32 GB+ VRAM'li GPU.

> Şirketin kendi uyarıları: bilgi sorularında büyük modellerden zayıf ve yanlışı emin bir dille söyleyebiliyor. Agent olarak kullanırken dosyaları onay almadan silebiliyor; çıktı token'ı ve tur sınırı koy. İngilizce dışındaki dillerde, yani Türkçede de kalite düşebilir.

Dene: huggingface.co/thinkingmachines/Inkling-2-Mini · thinkingmachines.ai/news/inkling-2-mini

Bu haftanın büyük resmi

Altı gelişmeyi yan yana koyunca üç net eğilim görünüyor.

1. Küçük açık modeller yerelde gerçek iş yapıyor

Inkling 2 Mini 20 GB'la bir Mac'te çalışıyor ve kendi testlerinde kodlamada kendi boyundakileri geçiyor. tiny-voca 18 MB'la işlemcide ses üretiyor, Voice Isolator GPU olmadan da çalışıyor. Bulut aboneliği olmadan yapılabilecek işler hızla artıyor.

2. Agent'lar uzmanlık işlerini parçalara bölüyor

Paper Orchestra makale yazımını beş uzman agent'a, rea-video bir videoyu tek tek sahnelere bölüyor. Kurulum bile agent'a devrediliyor: rea-video ve Voice Isolator, agent'ın okuyup kendini kuracağı talimatlarla geliyor.

3. Kapalı modellerde yarış verimlilikte

OpenAI bu sefer en güçlü modelin yanında maliyeti beşte bir olan Sol'u öne çıkarıyor ve Astra'nın daha az token harcadığını vurguluyor. Seçim artık "en iyisi hangisi" değil, "bu iş için yeterli olan en ucuzu hangisi".