



## Sıfır

token maliyeti (yerel)

## localOnly

veri makineden çıkmaz

## Air-gap

internetsiz tam AI

### İÇİNDEKİLER

1. Yönetici Özeti
2. Problem: Bulut AI Bağımlılığı
3. Gömülü Yerel Runtime
4. Donanım-Farkında Model Yönetimi
5. Yerel RAG ve Türkçe Yetenekler
6. localOnly ve Air-Gap
7. Maliyet ve Egemenlik
8. Sonuç

## Yönetici Özeti

KVKK-hassas kurumlar (kamu, finans, sağlık, savunma) bulut AI kullanamaz: veri yurt dışına çıkamaz, denetimde “veriniz nerede işlendi?” sorusu yanıtsız kalır. 361, AI iş yükünü tamamen tesis içinde çalıştıran gömülü bir yerel runtime sunar — LLM, embedding, OCR ve TTS token maliyeti olmadan yereldedir.

## Problem: Bulut AI Bağımlılığı

Bulut AI üç yükü beraberinde getirir: **veri egemenliği** kaybı (istem ve yanıt kurum sınırını terk eder), **öngörülemez token maliyeti** ve **denetim boşluğu**. Yerel AI kurmak isteyen ekipler ise Ollama/LM Studio gibi ayrı ürünleri kurup güncellemek, model seçmek ve GPU uyumunu elle çözmek zorunda kalır.

## Gömülü Yerel Runtime

AI motoru platformun içine gömülüdür (llama.cpp tabanlı): harici ürün gerekmez, doğru binary donanıma göre otomatik iner, modeller platform tarafından yönetilir. Chat, tool-calling, streaming ve görsel anlayan model (VLM) dahildir.

## Donanım-Farkında Model Yönetimi

Platform makineyi tarar (GPU/VRAM/CPU/RAM), backend'i otomatik seçer (CUDA → Vulkan → CPU) ve Model Merkezi'nde her modelin donanıma sığıp sığmadığını bir “**fit rozeti**” ile gösterir. GGUF modeller HuggingFace'ten aranır, otomatik indirilir ve indirme anında smoke-test ile doğrulanır.



### Otomatik backend

CUDA → Vulkan → CPU; donanıma göre seçim.



### Fit rozeti

Model donanıma sığıyor mu — indirmeden önce görünür.



### Yerel RAG

multilingual-e5 embedding CPU'da, token maliyeti sıfır.



### Türkçe OCR/TTS

PP-OCR ile taranmış Türkçe belge; Piper TTS ile Türkçe ses.

## Yerel RAG ve Türkçe Yetenekler

Embedding (multilingual-e5) süreç içinde, CPU'da, token maliyeti sıfır çalışır; bilgi tabanı ve RAG tamamen yereldedir. Taranmış Türkçe belgeler yerel PP-OCR ile okunur, Piper TTS Türkçe sesle yanıt üretir, SDXL-Turbo ile yerel görüntü üretimi mevcuttur (ticari kullanım için model lisansı kurulumda netleştirilir).

## localOnly ve Air-Gap

*Veri makineden çıkamaz — sözleşmeye yazılabilir teknik bir garanti.*

— **localOnly** ilkesi

Tek anahtarla (**localOnly=true**) bulut-proxy modeller hem listeden kalkar hem kategorik reddedilir. Air-gap (tamamen internetsiz) kurulumda tüm bileşenler yerel mirror'dan beslenir; hibrit ihtiyaçta bulut modeller yalnızca merkezi kayıttan, anahtarlar domain'lere inmeden eklenir.

## Maliyet ve Egemenlik

| Boyut               | Bulut AI         | 361 Yerel AI                      |
|---------------------|------------------|-----------------------------------|
| Token maliyeti      | Kullanımla büyür | Sıfır (donanım yatırımıyla sabit) |
| Veri yerleşimi      | Sağlayıcı bulutu | Tesis içi — makineden çıkmaz      |
| Denetim yanıtı      | Belirsiz         | “Burada” — tek kelime             |
| İnternetsiz çalışma | Yok              | Air-gap destekli                  |

1

### Donanım prob

GPU/VRAM/CPU otomatik taranır

2

### Model indir

fit rozeti + smoke-test doğrulama

3

### localOnly

bulut modeller kategorik reddedilir

4

### Air-gap

yerel mirror’dan tam AI

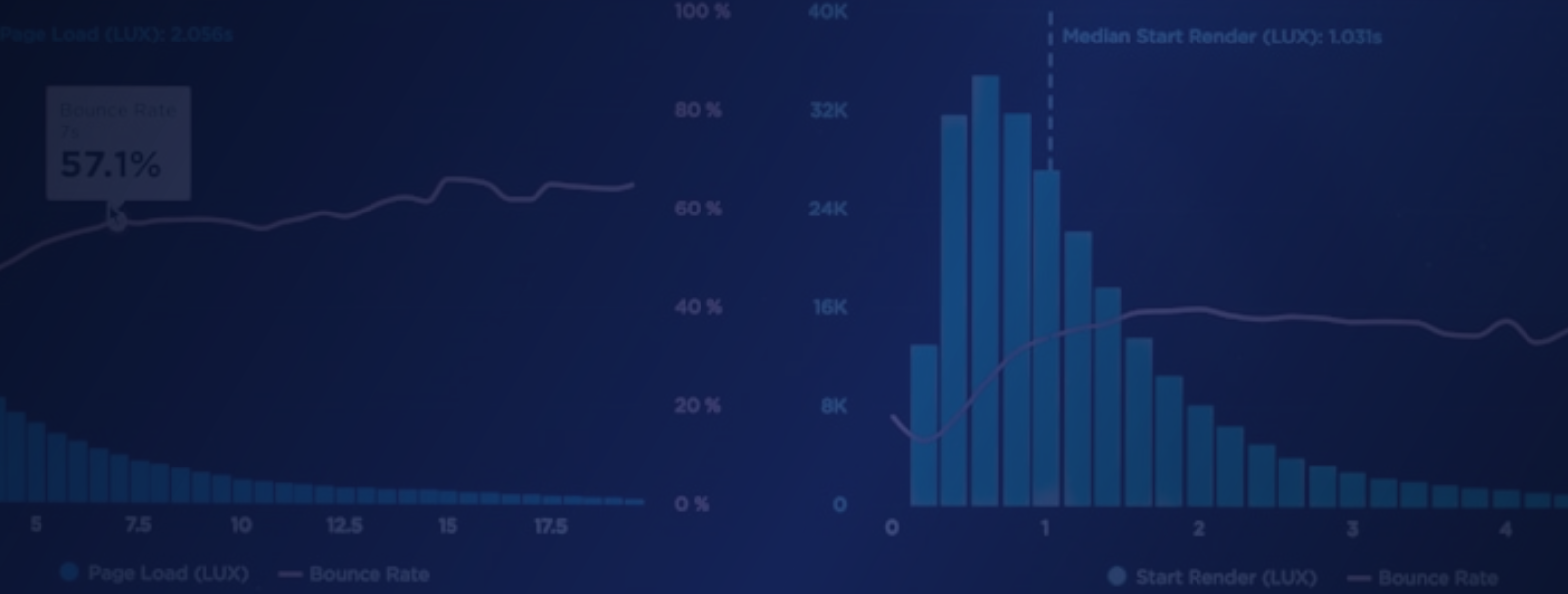
## Sonuç

Bulut AI’sız tam AI mümkündür: LLM + RAG + Türkçe OCR + Türkçe TTS yerelde, token ücreti olmadan çalıştığında, veri egemenliği bir kısıt değil, kurumun kontrolündeki bir yetenek hâline gelir.

SONRAKİ ADIM

# Yerel AI'ı kendi sunucunuzda POC'ta görün

2 haftalık ücretsiz POC — gerçek süreçlerinizde, sıfır risk. Kurulum yok, mevcut sisteminize dokunmadan; beğenmezseniz hiçbir şey değişmez.



| LOAD             | OPTIONS           | SESSIONS       | SESSIONS             | PVs Per Session |
|------------------|-------------------|----------------|----------------------|-----------------|
| Page Views (LUX) | Bounce Rate (LUX) | Sessions (LUX) | Session Length (LUX) | PVs Per Session |
| 2.7Mpvs          | 40.6%             | 479K           | 17min                | 2pv             |
| 500K 100%        |                   | 4 pvs          |                      |                 |

## Hemen başlayın

17 sağlayıcı · 52 tool · 8 kanal ailesi/44 adaptör · 198 modül



361.com.tr · info@361.com.tr