YAPAY ZEKÂ MİMARİSİ & KURUMSAL STRATEJİ RAPORU

Sinek Etkisi: Olay Güdümlü Ajan Mimarileri ve Maliyet Yanılgısı

Hareket Eden Araçtaki Atalet Prensibinden Yola Çıkarak Kurumsal LLM Maliyetlerini Düşürmek

Mehmet Keskinkılıç · Kasım 2026 **Kategori:** Sistem Tasarımı & Finansal Optimizasyon

Yaklaşım: Deterministik Durum Makineleri + Hedefli Çıkarım

Özet: *Birçok işletme, yapay zekâ dönüşümünü ve otonom ajan entegrasyonunu "öngörülemez API maliyetleri" nedeniyle ertelemektedir. Bu çekincenin önemli bir nedeni teknolojik yetersizlik değil, iş akışındaki her adımda büyük modelleri tetikleyen mimari tasarımdır. Bu çalışma; kapalı bir araçta uçan sineğin fiziksel durumundan ilham alan "İvmeli Ajan Modeli" ile rutin işlemleri sıfır token maliyetinde tutarak yalnızca anomali anlarında LLM çalıştırmanın yol haritasını ortaya koyar. Rapordaki rakamlar ölçüm sonucu değil, örnek senaryo varsayımıdır.*

1. Fiziksel Metafor: Kapalı Kabindeki Sinek Neden Savrulmaz?

100 km/s hızla seyreden kapalı bir otomobilin içindeki sinek, havada sabit durmak veya aracın ön camından arka koltuğuna doğru uçmak için ekstra bir efor sarf etmez. Bunun iki temel fiziksel nedeni vardır: *aynı hızla taşınan hava kütlesi* ve *eylemsizlik (atalet)* prensibi. Sinek aracın zeminine değil, kabin içindeki durgun havaya göre uçar. Sinek ancak araç ani bir fren yaptığında, gaza basıp ivmelendiğinde veya pencere açılıp dışarıdan türbülans girdiğinde dengesini sağlamak için yüksek enerji harcar.

Geleneksel ajan tasarımları ise sanki sinek aracın dışında, 100 km/s'lik rüzgâra karşı uçmaya zorlanıyormuş gibi kurgulanır: Her rutin veri girişi, her durum kontrolü ve her basit onay adımı için büyük dil modelleri (LLM) baştan sona tetiklenir, bu da kontrolsüz bir token tüketimine yol açar.

2. Kurumsal Hata: Her Adıma LLM Bağlama Refleksi

Şirketlerin "Yapay zekâ dönüşümü bütçemizi aşar" tezinin arkasında çoğu zaman, her operasyonel olayı doğrudan pahalı bir frontier modele paslayan monolitik ajan tasarımları yatar. Oysa kurumsal süreçlerin (müşteri talepleri, fatura doğrulamaları, ERP güncellemeleri, bilet yönlendirmeleri) önemli bir bölümü öngörülebilir, deterministik kurallarla yürütülebilir. Bu oran süreçten sürece değişir ve ölçülmesi gerekir; bu raporun örnek senaryosu işlerin %90'ının modele gitmeden tamamlandığını varsayar.



© 2026 Mehmet Keskinkılıç · mehmetkeskinkilic.com

3. Mimarinin Üç Taşıyıcı Sütunu

A. Deterministik Bekçi Katmanı (Gating Layer)

Sisteme düşen her girdi ilk olarak sıfır veya mikro maliyetli kontrol mekanizmalarından geçer. Girdi; standart bir kalıba, önceden tanımlanmış bir API rotasına veya kural tabanlı doğrulamaya uyuyorsa işlem doğrudan kod seviyesinde tamamlanır. LLM API çağrısı yapılmaz, token harcanmaz.

B. Durum Makinesi Tabanlı İş Akışı (Finite State Machine - FSM)

Ajanın iş süreçleri LangGraph gibi durum odaklı çerçevelerle modellenir. Sistem bir durumdan diğerine geçerken karar kuralları kodla işletilir. Yalnızca beklenmeyen bir istisna (exception), yapılandırılmamış bir veri anomalisi veya çok seçenekli stratejik bir yol ayrımı oluştuğunda sistem "ivmelenme" alarmı verir ve LLM düğümüne dallanır.

C. Semantik Önbellekleme ve Değişim Tabanlı (Diff) Çıkarım

Tekrarlayan sorgular yerel hafif embedding modelleriyle yakalanarak doğrudan önbellekten sunulur. Bir LLM çağrısı zorunlu olduğunda dahi, şirketin tüm doküman seti her seferinde yeniden gönderilmez; Prompt Caching ve yalnızca değişen durum parametreleri (delta) aktarılarak harcama azaltılır.

4. Maliyet ve Performans Karşılaştırması (Örnek Senaryo)

Aşağıdaki tablo gerçek bir kurulumun ölçümü değildir. Günde 10.000 operasyonel olay işleyen orta ölçekli bir destek/işlem merkezi için hazırlanmış örnek bir senaryodur.

Metrik / Aşama	Monolitik LLM Ajanı	Hibrit İvmeli Ajan Modeli	Tasarruf & Etki
Rutin İşlem Akışı (%75)	Her çağrıda ~1.500 token	0 token (kural ve API düzeyinde yürütme)	%100 LLM token tasarrufu
Benzer / Tekrar Eden Girdi (%15)	Model baştan akıl yürütür	0 token (semantik önbellek eşleşmesi)	%100 LLM token tasarrufu
Anomali & Karar Durumu (%10)	Pahalı bağlamla çalışır	Küçük/optimize model + minimum fark bağlamı	%70 token optimizasyonu
Örnek Yanıt Süresi	1.800 ms – 3.500 ms	< 40 ms (rutin) / 900 ms (anomali)	Rutinde %97'den fazla, anomalide yaklaşık %50–75 düşüş
Aylık Tahmini Maliyet	$4.500 – $7.000 / ay	$350 – $600 / ay	**Yaklaşık %87 – %95 tahmini tasarruf**

Hesap varsayımı: Ayda yaklaşık 300.000 olay × ~1.500 token = ~450 milyon token; milyon token başına ortalama yaklaşık $10–16. Hibrit modelde %10'luk anomali payının token tutarı bu hesapla yaklaşık $135–215 eder; tablodaki aralık bunun üzerinde, temkinli bir tahmindir. Kural yazımı, altyapı, bakım ve önbellek işletim giderleri iki sütuna da dahil değildir.

5. Yönetim Kurulu Özeti ve Aksiyon Planı

> **Stratejik Sonuç:** Yapay zekâ dönüşümü "tüm süreçleri bir LLM'e devretmek" demek değildir. Başarılı dönüşüm; yazılım mühendisliğinin deterministik güvenilirliği ile LLM'lerin problem çözme kabiliyetini doğru eşiklerle ayırmaktır. Örnek senaryodaki gibi işlerin %90'ı token harcamadan yürütülüp yalnızca kalan %10'luk belirsizlik anlarında yapay zekâ devreye girdiğinde, maliyet bariyeri avantaja dönüşebilir.

Aksiyon planı

1. **Ölçün:** Mevcut LLM çağrılarını olay türüne göre hacim, token, gecikme ve hata oranıyla kaydedin. Bu raporun oranlarının yerine kendi verinizi koyun.
2. **Ayırın:** Yüksek hacimli ve sonucu doğrulanabilir rutin işleri kurallara ve durum makinelerine taşıyın.
3. **Önbelleği denetimli açın:** Semantik önbelleği benzerlik eşiğiyle çalıştırın, örneklem denetimi yapın ve kişisel ya da hassas verilerin yanıtlar arasında karışmasını engelleyin.
4. **Yükseltme kuralı koyun:** Anomalileri önce küçük modele gönderin. Güven düşükse büyük modele veya insana yükseltin.
5. **Sürekli izleyin:** Modelsiz tamamlanan iş oranını, hatalı kural ve önbellek yanıtlarını ve aylık toplam maliyeti, mühendislik giderleriyle birlikte takip edin.

Mehmet Keskinkılıç
Havacılık Operasyonları Uzmanı & Bağımsız iOS Geliştirici
© 2026 Mehmet Keskinkılıç. Kaynak göstererek paylaşılabilir. Güncel sürüm: mehmetkeskinkilic.com/tr/writing/