

Gümrük Mevzuatı Denetim ve Analiz Ajansı: Teknik Performans ve RAG Değerlendirme Raporu

Model: Llama 3.1 (8B - Ollama Local)

Özet

Bu rapor, 4458 sayılı Gümrük Kanunu dökümanlarını temel alan, RAG (Retrieval-Augmented Generation) mimarisiyle güçlendirilmiş otonom bir yapay zeka ajanının başarı metriklerini sunar. Yapılan 100 soruluk "Stress Test" ve "5 Hakemli Denetim" sonucunda sistemin mevzuata sadakat oranı %90.4 (4.52/5) olarak ölçülmüş, genel performans skoru ise 5 üzerinden 3.73 olarak belirlenmiştir. Rapor, sistemin zayıf noktalarını (eksiksizlik ve terim karmaşası) kök neden analiziyle incelemekte ve iyileştirme yol haritasını sunmaktadır.

1. Giriş ve Proje Amacı

Gümrük mevzuatı, teknik terim yoğunluğu ve sürekli güncellenen yapısı nedeniyle hata payının en düşük olması gereken hukuk alanlarından biridir. Bu projenin amacı:

- Gümrük müşavirleri ve uzmanları için hatasız bir karar destek sistemi oluşturmak.
- Mevzuat dökümanları içinde arama yapma süresini minimize etmek.
- Şirket verilerinin dış sunuculara (Cloud) çıkmasını engelleyerek, tamamen yerel (On-premise) bir yapıda güvenli analiz yapmaktır.

2. Sistem Mimarisi

2.1 Kullanılan Teknolojiler

- Beyin (LLM):** Llama 3.1 (Ollama üzerinden yerel entegrasyon).
- Hafıza (Vektör Veri Tabanı):** LlamaIndex kütüphanesi ile yapılandırılmış VectorStoreIndex.
- Embedding (Anlamsal Arama):** sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2 (Türkçe hukuk diline duyarlı).
- Veri Kaynağı:** 4458 Sayılı Gümrük Kanunu tam metni (PDF).

2.2 Prompt Mühendisliği ve Ajan Yapısı

Sistem, "ReAct" (Reasoning + Acting) döngüsünü kullanır. Bir soru geldiğinde ajan önce mevzuat dökümanını tarar (Search), ilgili maddeyi bulur (Observe), bulduğu maddeyi hukuk mantığıyla yorumlar (Think) ve nihai cevabı üretir.

3. Benchmark Metodolojisi: 5 Hakemli Hibrit Sistem

Sistemin doğruluğunu ölçmek için "Blind Judge" (Kör Hakem) problemini aşan **Hibrit Bir Değerlendirme Modeli** kurulmuştur. Her cevap, arka planda çalışan bir "Başmüfettiş LLM" tarafından şu 5 kriterde puanlanmıştır:

- Sadakat (Faithfulness):** Cevap sadece dökümanda yazan bilgiye mi dayanıyor?
- Sayısal Doğruluk:** Süreler ve Euro limitleri doğru mu?
- Atıf Doğruluğu:** Referans verilen Madde numarası içerikle uyumlu mu?
- Eksiksizlik:** İstisnalar ve şartlar belirtilmiş mi?
- Üslup:** Yanıt profesyonel gümrük terminolojisine uygun mu?

4. Analiz: Ajan Nasıl Düşünüyor? (Örnek Log)

Soru: "Yolcu beraberinde getirilen 430 Euro değerindeki bir eşya için muafiyet uygulanır mı?"

Thought: Kullanıcı yolcu muafiyeti limitini soruyor. 4458 sayılı kanun ve ilgili kararnameyi taramam gerekiyor.

Action: Query database for "yolcu muafiyeti 430 Euro".

Observation: 2009/15481 sayılı karar ve 167. madde; 18 yaş üstü için 430 Euro, 15 yaş altı için 150 Euro limitini belirtir.

Final Answer: Evet, 18 yaş üzerindeki yolcular için 430 Euro'ya kadar olan hediyeelik eşyalar vergiden muaftır. Ancak 15 yaş altındakiler için bu limit 150 Euro'dur.

5. Performans ve Kök Neden Analizi (Hata Kayıtları)

100 soruluk testte tespit edilen 2 kritik hata türü:

5.1 Karar Verme ve İtiraz Süreleri (Sayısal Karmaşa)

- Hata:** İtiraz süresine "30 gün" denilmesi (Doğrusu 15 gün).
- Kök Neden:** Kanunun 242. maddesinde geçen "İdare 30 gün içinde karar verir" cümlesindeki 30 rakamının, itiraz süresiyle karıştırılması.

5.2 Terimsel Kayma (Context Drift)

- Hata:** Bağlayıcı Tarife Bilgisi sorusuna "Gümrük Kıymeti" cevabı verilmesi.
- Kök Neden:** "Tarife" ve "Kıymet" kelimelerinin vektör uzayında birbirine çok yakın olması.

6. Rakip Analizi: Yerel Gümrük Ajansı vs. Standart LLM

Özellik	Gümrük Ajansı (Llama 3.1 RAG)	Standart ChatGPT / GPT-4
Mevzuat Hakimiyeti	Mükemmel: Güncel PDF'e doğrudan erişim.	Zayıf: Eğitim verisindeki eski bilgilerle konuşur.
Veri Gizliliği	Tam Güvenli: Veri lokal sunucudan çıkmaz.	Düşük: Sorgular bulut sunucularına iletilir.
Referans Verme	Madde numarası ve metin kesiti sunar.	Genelde genel geçer ifadeler kullanır.
Maliyet	Sıfır (Donanım maliyeti hariç).	Yüksek (Token bazlı abonelik).

Metrik	Yerel Model (Llama 3.1)	API Model (OpenAI)	Fark (Yerel - API)	Kazanan
Sadakat (Faithfulness)	4.52	3.58	+0.94	Yerel
Sayısal Doğruluk	4.30	3.72	+0.58	Yerel
Atıf Doğruluğu	3.34	2.63	+0.71	Yerel
Eksiksizlik	3.12	3.09	+0.03	Eşit
Üslup (Profesyonellik)	3.36	4.07	-0.71	API
GENEL SKOR	3.73	3.42	+0.31	Yerel
Hız (Süre/Sn)	1.20	2.27	-1.07 sn	Yerel

Örnek 1: Bilgi Sınırı ve "Bilmiyorum" Deme Eşiği

Soru 100: "Kripto madencilik cihazlarının ithalatında ÖTV ve KDV matrahı nasıl oluşturulur?"

- Yerel Model (Llama 3.1):** "Cihazın türüne göre %18, %8 veya %2 vergi oranları uygulanabilir... KDV matrahı maliyet bedelleridir... ÖTV matrahı Türkiye'de satışa göre hesaplanır." (**Puan: 4.0**)
- OpenAI API:** "Bu konuda spesifik bir bilgi verilmemiştir... Genellikle ilgili vergi mevzuatına bağlıdır... Gümrük kıymeti üzerinden hesaplanabilir." (**Puan: 3.2**)

Prompt/Ajan Yorumu:

- **Yerel Model:** Mevzuatın genel mantığını (matrah oluşumu) alıp sorudaki spesifik konuya (kripto) **adapte etmeye** çalışmış. Daha "üretken" bir tavır sergilemiş.
 - **OpenAI:** Çok daha **muhafazakar**. Dökümanda "kripto" kelimesini görmediği an savunmaya geçip genel geçer cümleler kurmuş.
 - **Ders:** Eğer sistemin uzman gibi yorum yapmasını istiyorsan Llama; sadece dökümanda yazanı (veya yazmayanı) söylesin istiyorsan OpenAI daha güvenli.
-

Örnek 2: Sayısal Veri ve Kaynak Karmaşası

Soru 20: "Gümrük Kanunu'na göre itiraz süresi kaç gündür?"

- **Yerel Model (Llama 3.1):** "Gümrük vergileri ve cezalarına karşı tebliğ tarihinden itibaren **onbeş gün** içinde itiraz edilebilir... İdare 30 günde karara bağlar." (**Puan: 2.4 - Yanlış süre vurgusu nedeniyle**)
- **OpenAI API:** "Sistem cevabında itiraz süresi **otuz gün** olarak belirtilmiştir." (**Puan: 1.2**)

Prompt/Ajan Yorumu:

- **Analiz:** Bu bir "Context Contamination" (Bağlam Kirlenmesi) örneğidir. Her iki model de 15 ve 30 rakamlarını aynı paragrafta görmüş.
 - **Fark:** Llama 15 günü doğru yakalamış ama 30 günü de cevaba ekleyerek kafayı karıştırmış. OpenAI ise direkt 30 günü "itiraz süresi" sanarak **halüsinasyon** görmüş.
-

Örnek 3: Tanımsal Derinlik vs. Yüzeysel Geçiş

Soru 3: "Bağlayıcı Tarife Bilgisi (BTB) nedir?"

- **Yerel Model (Llama 3.1):** BTB'yi yanlışlıkla "Gümrük kıymetine eklenen unsurlar" (lisans, telif vb.) olarak açıklamış. (**Puan: 3.0**)
- **OpenAI API:** "Eşyanın gümrük işlemlerine tabi tutulması için yapılan resmi bildirimdir..." diyerek aslında "Gümrük Beyanı" tanımını yapmış. (**Puan: 1.0**)

Prompt/Ajan Yorumu:

- **Analiz:** Her iki model de burada **Retrieval (Bilgi Getirme)** hatası kurbanı olmuş. Llama "Tarife" kelimesini "Kıymet" ile karıştırırken, OpenAI soruyu genel bir "Beyan" sorusu gibi algılamış.

- **İlginçlik:** OpenAI'nin üslubu o kadar profesyonel ki, yanlış bilgi verdiğini ilk bakışta anlamak zor (Hakem bu yüzden "Üslup 5" ama "Sadakat 1" vermiş).

7. Sonuç ve Gelecek Adımlar

Yapılan benchmark çalışmaları, sistemin temel gümrük sorularında oldukça başarılı olduğunu, ancak karmaşık sürelerin ayrıştırılmasında daha keskin filtreleme ihtiyacı duyduğunu kanıtlamıştır.

Yol Haritası:

1. **Hibrit Arama:** Sadece anlamsal değil, anahtar kelime bazlı (BM25) aramanın entegrasyonu.
2. **Multimodal Entegrasyon:** Gümrük beyannamelerinin (taranmış dökümanların) OCR ile analize dahil edilmesi.
3. **Prompt Sertleştirme:** "Sadece dökümandaki veriyi kullan" talimatının sistem promptunda en üst sıraya alınması.