Pharma+ Projesi: Faz 1 - LLM Entegrasyon ve Veri Çıkarım Raporu

Hazırlayan: Abdulcelil Elmas

Konu: OCR Prospektüs Çıktılarının Yapılandırılmış JSON Formatına Dönüştürülmesi ve Pharma+ Projesinde kullanılıcak LLM modelinin Seçilmesi

1. Sistemin Amacı ve Veri Mimarisi

Pharma+ projesinin veri hattında (data pipeline), kullanıcıların çektiği prospektüs fotoğraflarından elde edilen karmaşık OCR metinlerinin temizlenerek veritabanına işlenmesi hedeflenmiştir. Veritabanı (DB) ekibiyle yapılan ER Diyagramı analizleri sonucunda, uygulamanın omurgasını oluşturacak `Drugs` tablosunun yapısı referans alınarak sistemin Veri Sözleşmesi güncellenmiştir. ML ekibi olarak, ham metni bu karmaşık tablo yapısına hatasız dönüştürecek bir LLM mimarisi tasarlanmıştır.

2. Metodoloji (Prompt Engineering)

LLM entegrasyonu sürecinde standart özetleme ("Summarize") algoritmaları yerine, **Stanford Üniversitesi'nden Dr. Andrew Ng'nin "Prompt Engineering" metodolojileri** referans alınarak, veri madenciliği kalitesini artırmak amacıyla aşağıdaki ileri düzey teknikler kullanılmıştır:

- **Extracting & Transforming (Çıkarım ve Dönüştürme):** Modele düz özetleme yaptırılmamış; OCR kaynaklı harf hatalarını düzeltmesi (Proofreading) ve metni doğrudan DB ekibinin talep ettiği iç içe (nested) JSON formatına çevirmesi sağlanmıştır.
- **Condition Checking (Halüsinasyon Önleyici):** Prospektüste açıkça yer almayan bilgiler için modelin kendi veritabanından uydurma yapması yasaklanmış, bulunamayan String değerler için `"Belirtilmemiş"`, listeler için boş dizi `[]` döndürmesi şart koşulmuştur.
- **Multi-Task Inferring (Çoklu Çıkarım):** Karmaşık uyarı metinleri tek bir döngüde `TargetCondition` (Riskli Durumlar), `Food_interactions` (Gıda Etkileşimi) ve `Drug_interactions` (İlaç Etkileşimi) olarak üç farklı dalda başarıyla kategorize edilmiştir.

3. Stres Testi ve Benchmark Sonuçları

Mimarinin dayanıklılığını ölçmek adına OpenRouter üzerinden Gemini ve GPT-4o modelleri 5 farklı uç durum senaryosuyla 10 kez test edilmiştir:

1. **Harf Hatalı Metin:** OCR hataları başarıyla filtrelenmiş ve temiz Türkçe çıktı alınmıştır.
2. **Eksik Veri Testi:** Metinde olmayan bilgiler için modeller uydurma (halüsinasyon) yapmamıştır.
3. **İç İçe Uyarı Testi:** Karmaşık gıda ve ilaç etkileşimleri hatasız ayrıştırılmıştır.
4. **İlgisiz Veri (Market Fişi):** Kullanıcının yanlış fotoğraf çekme senaryosu test edilmiş; sistem çökmek yerine tüm şemaya `"Belirtilmemiş"` basarak güvenliği sağlamıştır.
5. **Kısa Tıbbi Jargon:** "Prl 500mg tb." gibi kısaltmalar başarıyla genişletilmiştir.

Sonuç: Google üzerinden sağlanan ücretsiz geliştirici kotaları sayesinde API maliyeti sıfırlandığı için; testlerde %100 format bütünlüğünü koruyan, halüsinasyon yapmayan ve Apriori algoritmasına uygun kategorik kelime seçimiyle mantıksal çıkarım (inferring) kapasitesi en yüksek olan **Gemini Pro** modeli, sistemin ana motoru olarak seçilmiştir.

4. Nihai JSON Şeması (Backend Entegrasyonuna Hazır)

Aşağıdaki yapı, LLM API'mizden Backend endpoint'ine dönecek olan standart formattır. DB ekibi `Drugs` tablosunu bu şemaya göre eşleştirebilir:

```
{
  "Name": "string",
  "Manufacturer": "string",
  "Dosage_form": "string",
  "Recommended_Dosage": "string",
  "Age_of_use": "string",
```

```
  "Storage": "string",
  "Side_Effect": ["string", "string"],
  "Warning": {
    "TargetCondition": ["string"],
    "Food_interactions": ["string"],
    "Drug_interactions": ["string"]
  }
}
```