Veri Seti Karşılaştırmalı Analiz Raporu

Tarih: 11 Temmuz 2024 **Konu:** Strateji 1/2 ve Strateji 4/6 için geliştirilen sentetik veri üretici script'lerinin karşılaştırmalı analizi ve "olağanüstü model" hedefi için değerlendirilmesi.

---

1. Genel Değerlendirme Özeti

Bu rapor, proje kapsamında geliştirilen iki farklı veri üretim metodolojisini incelemektedir:

1. **Strateji 1 & 2 (`tool_chaining_data_generator.py`, `disambiguation_data_generator.py`):** Pydantic şema doğrulaması, yüksek karmaşıklıkta senaryolar ve derinlemesine mühendislik yaklaşımını benimseyen script'ler.
2. **Strateji 4 & 6 (`enhanced_data_generator.py`):** Senaryo çeşitliliği için havuz (pool) yapısını kullanan, konsept odaklı script.

Analiz sonucunda, her iki yaklaşımın da değerli olduğu ancak **Strateji 1 & 2 için geliştirilen script'lerin "uzman seviye" ve "olağanüstü model" hedefleriyle daha uyumlu olduğu** tespit edilmiştir. İki setin birleştirilerek kullanılması, modelin hem temel hem de ileri seviye yetenekleri öğrenmesi için tavsiye edilmektedir.

---

2. `enhanced_data_generator.py` (Strateji 4 & 6) Script'inin Analizi

Bu script, Strateji 4 (Hata Yönetimi) ve Strateji 6 (Doğal Sohbet) başlıkları için veri üretmek amacıyla geliştirilmiştir.

✅ Artıları (Güçlü Yönleri)

- **Doğru Veri Yapısı:** Script, `advanced_training_scenarios.md`'de tanımlanan temel `donguler` yapısına sadık kalmıştır. Bu, üretilen verinin model tarafından sorunsuzca kullanılabilmesini sağlar.
- **Strateji Odaklılık:** Görevler olan Strateji 4 ve 6 için özel fonksiyonlar (`generate_eh_data_v2`, `generate_nc_data_v2`) içermektedir.
- **Çeşitlilik için Havuz Kullanımı:** `ERROR_HANDLING_SCENARIOS` ve `ASSISTANT_INTERIM_PHRASES` gibi "havuz" (pool) yapıları, üretilen diyaloglarda tekdüzeliği kırarak çeşitliliği artırmaktadır. Bu, sentetik veri üretiminde etkili bir tekniktir.
- **Çok Turlu Doğal Sohbet:** Strateji 6 için `multi_turn_complaint` gibi birden fazla tur içeren doğal sohbet senaryoları üretmesi, modelin daha akıcı diyalogları yönetebilmesi için değerlidir.

❌ Eksileri (Geliştirilmesi Gereken Yönler)

- **KRİTİK EKSİK - Pydantic Doğrulaması Yok:** Script, `rol: "arac"` adımındaki API yanıtlarını manuel olarak oluşturmaktadır. Bu durum, `telekom_api_schema.py`'deki ana sözleşme ile %100 uyumluluğu garanti etmez ve gelecekteki değişikliklerde hatalı veri üretme riski taşır.
- **Gerçekçilikten Uzak Hata Mesajları:** Hata senaryolarında, API'den dönen hata mesajı her zaman jenerik olarak `"message": "API Error"` şeklindedir. Modelin, kullanıcıya spesifik ve anlamlı hata açıklamaları yapabilmesi için gerçekçi hata mesajları kritik öneme sahiptir.

- **Düşük Karmaşıklık Seviyesi:** Senaryolar genellikle doğrusaldır. Kullanıcının aniden konuyu değiştirdiği veya asistanın bilişsel esneklik göstermesi gereken karmaşık diyalog akışları bulunmamaktadır.

Sonuç

`enhanced_data_generator.py`, Strateji 4 ve 6'nın temel konseptlerini doğru bir şekilde uygulayan, **iyi bir temel seviye script'tir**. Ancak Pydantic doğrulaması, senaryo derinliği ve gerçekçilik gibi "uzman seviye" mühendislik yaklaşımlarını içermediği için tek başına "olağanüstü bir model" hedefi için **yetersizdir**.

---

3. Karşılaştırmalı Tablo

Özellik	Strateji 1 & 2 Script'leri	Strateji 4 & 6 Script'i	Kazanan & Neden
Mühendislik Yaklaşımı	**Schema-Driven (Şema Odaklı):** Pydantic ile %100 uyumluluk garantisi.	**Concept-Driven (Konsept Odaklı):** Manuel JSON üretimi, uyum garantisi yok.	**Strateji 1 & 2.** Güvenilirlik ve kalite için endüstri standardı bir yaklaşımdır.
Veri Gerçekçiliği	**Çok Yüksek:** `Faker` kütüphanesi ile dinamik ve gerçekçi veriler, spesifik API yanıtları.	**Orta:** Genellikle sabit veriler ve jenerik hata mesajları.	**Strateji 1 & 2.** Gerçek dünya verisine çok daha yakındır.
Senaryo Karmaşıklığı	**Çok Yüksek:** Koşullu mantık, zincir içi hata kurtarma, duygu analizi, konu değişikliği.	**Orta:** Genellikle doğrusal ve öngörülebilir senaryolar.	**Strateji 1 & 2.** Modelin sadece görev yapmasını değil, **düşünmesini ve adapte olmasını** öğretir.
Strateji Derinliği	**Derin:** Stratejilerin en zorlu ve karmaşık versiyonlarını modeller.	**Yüzeysel:** Stratejilerin temel gereksinimlerini karşılar.	**Strateji 1 & 2.** Konseptlerin özünü daha derinden işler.

---

4. Veri Seti Uyumluluğu ve Birleştirme Stratejisi

Evet, üretilen tüm veri setleri yapısal olarak birbiriyle uyumludur.

Her iki taraf da `advanced_training_scenarios.md`'de belirtilen temel JSON yapısını kullandığı için, tüm üretilen `.json` dosyaları **tek bir eğitim setinde güvenle birleştirilebilir.**

Bu birleştirme, modelin hem temel konuları (Strateji 4/6) hem de ileri seviye mantık yürütme ve diyalog akışlarını (Strateji 1/2) öğrenmesi için **gereklidir ve tavsiye edilir**.

---

5. Sonuç ve Öneri

Strateji 1 ve 2 için geliştirilen script'ler, projenin "olağanüstü model" hedefine ulaşması için gereken mühendislik disiplinini ve senaryo derinliğini sergilemektedir. Bu yaklaşım, projenin kalite standardını belirlemelidir.

Aksiyon Önerisi: Projedeki tüm veri üretim script'lerinde standart bir kalite seviyesi yakalamak için, `enhanced_data_generator.py` script'ine de **Pydantic ile API yanıtı doğrulaması** mekanizmasının (`telekom_api_schema.py` kullanımı) eklenmesi şiddetle tavsiye edilir. Bu, tüm veri setinin güvenilirliğini ve tutarlılığını en üst düzeye çıkaracaktır.