# **LLM Kullanarak Sentetik Metin Veri Üretimi için Hazır Python Kütüphaneleri ve GitHub Depoları**

**1. Giriş**

Sentetik veri, gerçek dünya verilerinin özelliklerini taklit etmek üzere yapay olarak oluşturulan bilgilerdir ve çeşitli alanlarda giderek artan bir öneme sahiptir. Özellikle, gerçek dünya verilerinin kıt, hassas veya elde edilmesi maliyetli olduğu durumlarda, Büyük Dil Modellerini (LLM'ler) eğitmek ve değerlendirmek için sentetik veri kullanımı önemli avantajlar sunmaktadır.1 LLM'lerin çeşitli Doğal Dil İşleme (NLP) görevlerindeki başarısı, sağlam ve çeşitli veri kümelerine olan ihtiyacı artırmıştır. Sentetik veri, bu ihtiyacı karşılamak için umut verici bir çözüm sunmaktadır.

LLM'lerin sentetik metin verisi üretimindeki özel uygulaması, geleneksel yöntemlere (veri artırma veya manuel oluşturma gibi) kıyasla zenginlik, nüans ve bağlamsal uygunluk açısından önemli avantajlar sunar.3 LLM'ler, dilin yapısı ve bağlamı hakkındaki doğal anlayışları sayesinde, kural tabanlı veya daha basit üretken modellerin oluşturabileceğinden çok daha karmaşık sentetik metinler üretebilirler. Bu raporun amacı, LLM tabanlı sentetik metin veri üretimini kolaylaştıran, kullanıma hazır Python kütüphanelerine ve GitHub depolarına ayrıntılı bir genel bakış sunmaktır. Rapor, bu araçların temel amaçlarını, sundukları ana özellikleri, ilgili web bağlantılarını ve sentetik veri üretimi sürecini nasıl kolaylaştırdıklarına dair örnek kullanımlarını içerecektir.

Bu raporun yapısı şu şekildedir: İlk olarak, LLM tabanlı sentetik metin veri üretimi için temel Python kütüphaneleri incelenecektir. Ardından, bu alandaki ilgili GitHub depoları ele alınacaktır. Raporun ilerleyen bölümlerinde, LLM sentetik veri üretiminde dikkate alınması gereken önemli teknikler ve hususlar tartışılacaktır. Son olarak, elde edilen bulgular özetlenecek ve bu alandaki gelecekteki yönelimlere dair çıkarımlar sunulacaktır.

**2. LLM Tabanlı Sentetik Metin Veri Üretimi için Python Kütüphaneleri**

* **2.1. Hugging Face Transformers** 6  
  Hugging Face Transformers kütüphanesi, önceden eğitilmiş LLM'ler ve ilgili araçlar için merkezi bir platform olarak hizmet vermektedir. Kütüphane, metin üretimi de dahil olmak üzere çeşitli NLP görevleri için yaygın olarak kullanılmaktadır. Transformers, BERT, T5, Falcon, LLaMA gibi binlerce önceden eğitilmiş açık kaynaklı modeli kullanıma sunmaktadır ve Hugging Face'in genişleyen LLM ekosisteminin amiral gemisi kütüphanesidir.16  
  Kütüphanenin temel özelliklerinden biri, çeşitli modellerden metin üretmek için kullanılan generate() yöntemidir.8 Bu yöntem, GPT, Mistral ve Llama serisi gibi popüler modellerle uyumludur. generate() yönteminin temel kullanımı, bir başlangıç metni (prompt) sağlamayı ve modelin bu metni devam ettirmesine veya tamamlamasına izin vermeyi içerir.7 Örneğin, aşağıdaki gibi basit bir kodla metin üretilebilir:  
  Python  
  from transformers import pipeline  
    
  generator = pipeline('text-generation', model='gpt2')  
  result = generator("Merhaba dünya, devam et...")  
  print(result)  
    
  Transformers kütüphanesi, sentetik veri üretimi için çeşitli kod çözme stratejileri sunar.8 Açgözlü arama (greedy search), her adımda en yüksek olasılığa sahip olan kelimeyi seçerken, ışın arama (beam search) birden fazla olası kelime dizisini takip ederek daha tutarlı sonuçlar elde etmeyi hedefler. Örnekleme (sampling) yöntemleri ise, olasılık dağılımından rastgele kelimeler seçerek daha çeşitli çıktılar üretir. Bu stratejiler, sentetik verinin istenen özelliklerine göre ayarlanabilir.  
  GenerationConfig sınıfındaki parametreler aracılığıyla metin üretimi özelleştirilebilir.8 max\_new\_tokens parametresi, üretilecek maksimum yeni token sayısını belirlerken, num\_beams ışın arama sırasında takip edilecek hipotez sayısını kontrol eder. do\_sample parametresi etkinleştirildiğinde, olasılık dağılımından örnekleme yapılır ve top\_k ile top\_p parametreleri, örnekleme sırasında dikkate alınacak olası kelimelerin sayısını ve olasılık kütlesini sınırlar.  
  Transformers, metin sınıflandırma veya diyalog gibi belirli görevler için sentetik veri üretmek amacıyla da kullanılabilir.6 LLM'lere uygun promptlar veya talimatlar sağlanarak, istenen formatta ve içerikte sentetik veri oluşturulabilir. Transformers kütüphanesinin esnekliği ve çok çeşitli önceden eğitilmiş modellerin mevcudiyeti, farklı ihtiyaçlara yönelik sentetik metin oluşturmak için onu çok yönlü bir araç haline getirir. Hugging Face Hub entegrasyonu sayesinde, çok sayıda önceden eğitilmiş modele ve veri kümesine kolayca erişilebilir.3 Bu, belirli alanlara veya stillere uygun sentetik veri üretmek için özellikle faydalıdır.  
  Örneğin, ürün açıklamaları oluşturmak için basit bir örnek verilebilir:  
  Python  
  from transformers import pipeline  
    
  generator = pipeline('text-generation', model='gpt2')  
  prompt = "Anahtar kelimeler: yüksek çözünürlük, 15 inç ekran, hafif. Ürün açıklaması:"  
  result = generator(prompt, max\_length=50, num\_return\_sequences=3)  
  for res in result:  
   print(res['generated\_text'])  
    
  Bu örnekte, verilen anahtar kelimelerden yola çıkarak kısa ürün açıklamaları üretilmektedir.  
  Resmi dokümantasyon bağlantısı: <https://huggingface.co/docs/transformers/>
* **2.2. OpenAI Python Kütüphanesi** 17  
  OpenAI Python kütüphanesi, metin üretimi ve diğer NLP görevleri için OpenAI'nin güçlü dil modellerine (GPT-3.5, GPT-4 vb.) bir arayüz sağlar. Bu kütüphane, OpenAI API'si ile etkileşimi kolaylaştırarak, geliştiricilerin uygulamalarına son teknoloji ürünü yapay zeka modellerini entegre etmelerine olanak tanır.19  
  Kütüphanenin temel işlevleri arasında, promptlara dayalı olarak metin üretmek için kullanılan openai.ChatCompletion.create() ve openai.Completion.create() yöntemleri bulunur.19 openai.ChatCompletion.create() yöntemi, özellikle sohbet modelleri için tasarlanmıştır ve model, messages (kullanıcı ve sistem rol mesajlarının bir listesi), temperature (rastgelelik seviyesi), max\_tokens (maksimum çıktı uzunluğu) gibi önemli parametreleri içerir.19 openai.Completion.create() yöntemi ise, daha genel metin tamamlama görevleri için kullanılır ve benzer parametrelere sahiptir, ancak messages yerine doğrudan bir prompt parametresi alır.  
  Örneğin, bir hikaye oluşturmak için aşağıdaki gibi bir kod kullanılabilir:  
  Python  
  from openai import OpenAI  
  import os  
    
  client = OpenAI(api\_key=os.environ.get("OPENAI\_API\_KEY"))  
    
  response = client.chat.completions.create(  
   model="gpt-3.5-turbo",  
   messages=[  
   {"role": "user", "content": "Arkadaşlık hakkında 3 cümlelik bir hikaye oluştur."},  
   ],  
   temperature=0.7,  
   max\_tokens=100  
  )  
  print(response.choices.message.content)  
    
  OpenAI API'si, yapılandırılmış sentetik veri (örneğin, CSV formatında) oluşturmak için de kullanılabilir.3 LLM'lere CSV formatını, şemayı ve sütunlar arasındaki ilişkileri belirten yapılandırılmış promptlar sağlanarak, anlamlı ve tutarlı sentetik tablolar oluşturulabilir. OpenAI Cookbook, bu tür uygulamalar için pratik örnekler sunmaktadır.3  
  OpenAI API'si ayrıca, web araması gibi araçları kullanma ve fonksiyon çağırma gibi gelişmiş özellikler sunar.20 Bu özellikler, üretilen sentetik verinin gerçekçiliğini ve kullanışlılığını artırabilir. Örneğin, bir LLM'ye güncel bir olay hakkında sentetik bir haber makalesi yazması istendiğinde, web arama aracı kullanılarak en son bilgilere erişilebilir.  
  Resmi dokümantasyon bağlantısı: <https://platform.openai.com/docs/api-reference>
* **2.3. LangChain** 16  
  LangChain, LLM'ler tarafından desteklenen uygulamalar oluşturmak için bir çerçeve olarak tanımlanabilir ve karmaşık sentetik veri üretimi iş akışları oluşturmak için kullanılabilir.16 LangChain, birden fazla promptun zincirlenmesine, harici veri kaynaklarının kullanılmasına ve daha karmaşık sentetik veri üretimi süreçleri oluşturmak için farklı bileşenlerin entegre edilmesine olanak tanır.29  
  Tipik bir LangChain tabanlı sentetik veri üretimi süreci, belge parçalama (document chunking), bağlam oluşturma (context generation), sorgu oluşturma (query generation) ve veri evrimi (data evolution) gibi adımları içerir.29 Belge parçalama, büyük belgeleri daha küçük, anlamlı parçalara ayırarak bağlamın korunmasına yardımcı olur. Bağlam oluşturma aşamasında, ilgili bilgileri bulmak için rastgele bir veri parçası seçilir ve benzerlik algoritmaları kullanılarak ilişkili parçalar gruplandırılır. Sorgu oluşturma adımında, oluşturulan bağlam kullanılarak bir LLM aracılığıyla bir dizi soru veya ifade üretilir. Veri evrimi ise, üretilen sorguların karmaşıklığını ve gerçekçiliğini artırmak için çeşitli şablonlar uygulanarak gerçekleştirilir.  
  LangChain iş akışlarında bilgi grafikleri veya benzerlik algoritmaları entegre edilerek daha gerçekçi ve çeşitli sentetik veri kümeleri oluşturma potansiyeli bulunmaktadır.29 LangChain'in modülerliği ve entegrasyon yetenekleri, basit LLM promptlamanın ötesine geçen, yüksek düzeyde özelleştirilmiş sentetik veri üretimi süreçlerinin oluşturulmasına olanak tanır.  
  Örneğin, belirli bir belgeye dayalı olarak sentetik soru-cevap çiftleri oluşturmak için yüksek düzeyli bir LangChain iş akışı şu şekilde olabilir: İlk olarak, belge parçalara ayrılır. Ardından, her parça için ilgili bağlamlar oluşturulur. Bu bağlamlar kullanılarak, LLM'den belge içeriğiyle ilgili sorular ve cevaplar üretmesi istenir. Son olarak, üretilen soru-cevap çiftleri veri evrimi teknikleriyle zenginleştirilebilir.  
  Resmi dokümantasyon bağlantısı: <https://www.langchain.com/>
* **2.4. Diğer İlgili Kütüphaneler** 16  
  Belirli sentetik veri üretimi senaryolarında ilgili olabilecek diğer Python kütüphaneleri şunlardır:
  + **ydata-synthetic**: GAN tabanlı modeller (CTGAN, TimeGAN) dahil olmak üzere çeşitli sentetik veri üretimi teknikleri sunar.31
  + **SDV (Synthetic Data Vault)**: Klasik makine öğrenimi ve derin öğrenme modellerini içeren çeşitli sentezleyiciler sağlar ve tek ve çok tablolu verileri destekler.31
  + **Mostly AI's Synthetic Data SDK**: Yüksek doğruluklu, gizlilik açısından güvenli sentetik veri üretmeye odaklanan açık kaynaklı bir araç kitidir.32
  + **NLTK (Natural Language Toolkit)** ve **SpaCy**: Öncelikli olarak sentetik veri üretimi için olmasa da, sentetik veri iş akışlarında metin verilerinin ön ve son işlenmesi için faydalı olabilirler.16
  + **SentenceTransformers**: Cümle gömülmeleri oluşturmak için kullanılabilir, bu da sentetik verinin semantik benzerliğini ve çeşitliliğini analiz etmek için kullanılabilir.16

Sentetik veri üretimi için çeşitli Python kütüphaneleri mevcuttur ve her birinin kendine özgü güçlü yönleri ve odak noktaları vardır. Kütüphane seçimi, veri türü (tablolu, zaman serisi, metin), istenen kontrol seviyesi ve performans değerlendirmeleri gibi görevin özel gereksinimlerine bağlıdır. LLM'ler metin üretimi için güçlü olsa da, diğer kütüphaneler farklı veri modaliteleri veya gizlilik koruma gibi belirli kalite özelliklerine yönelik özel teknikler sunar. Bu kütüphanelerin LLM tabanlı yaklaşımlarla entegre edilmesi, daha kapsamlı sentetik veri çözümlerine yol açabilir.

**3. LLM Tabanlı Sentetik Metin Veri Üretimi için GitHub Depoları**

* **3.1. Awesome-LLM-Synthetic-Data** ((<https://github.com/wasiahmad/Awesome-LLM-Synthetic-Data>)) 33  
  Bu depo, LLM'lerin sentetik verileri, LLM'ler tarafından sentetik veri üretimi ve LLM'ler için sentetik veriler hakkında makalelerin, araçların ve blogların derlenmiş bir listesidir.35 Depo, anketler, yöntemler, uygulama alanları (matematiksel muhakeme, kod üretimi, metin-SQL, hizalama vb.), veri kümeleri, araçlar ve bloglar gibi çeşitli kategorileri kapsamaktadır. Depoda listelenen, özellikle sentetik metin veri üretimiyle ilgili olan belirli teknikler ve uygulama alanları (örneğin, yüksek kaliteli/karmaşıklıkta talimat üretimi, hizalama) dikkat çekicidir. Bu depo, LLM tabanlı sentetik veri üretimi alanındaki en son araştırmaları, metodolojileri ve araçları takip etmek için değerli bir kaynak olarak hizmet vermektedir. Deponun kapsamlı yapısı, kullanıcıların ilgili akademik makaleleri, pratik araçları ve bilgilendirici blog gönderilerini hızlı bir şekilde keşfetmelerini sağlayarak, sentetik veri üretimi tekniklerini keşfetmelerini ve uygulamalarını kolaylaştırır.
* **3.2. LLM-Synthetic-Data** ((<https://github.com/pengr/LLM-Synthetic-Data>)) 33  
  Bu depo, LLM odaklı sentetik verilere odaklanan GitHub bağlantıları ve blog gönderileri dahil olmak üzere bir başka kaynak koleksiyonudur.33 Depo, LLM'lerin sentetik veri üretimi bağlamını ve zorluklarını anlamak için ilgili olan ön eğitim, talimat ayarlama, model çöküşü ve uzun bağlam gibi konuları içermektedir. Depoda bahsedilen, sentetik metin veri üretimi hakkında pratik rehberlik veya içgörü sağlayan belirli blog gönderileri veya kaynaklar bulunmaktadır. Önceki depoya benzer şekilde, bu koleksiyon da LLM tabanlı sentetik veri üretimi hakkında kaynak arayan araştırmacılar ve uygulayıcılar için değerli bir başlangıç noktası sunmaktadır. İlgili bağlantıları ve makaleleri bir araya getirerek, depo kullanıcıların bilgi arama çabalarını azaltır ve alanındaki temel kaynaklara merkezi bir erişim noktası sağlar.
* **3.3. Zhen-Tan-dmml/LLM4Annotation** ((<https://github.com/Zhen-Tan-dmml/LLM4Annotation>)) 36  
  Bu depo, LLM'leri açıklama (annotation) için kullanmaya odaklandığı için potansiyel olarak ilgili olabilir ve bu, LLM'leri veri oluşturmak için kullanma açısından sentetik veri üretimiyle benzerlikler taşır.36 Deponun belirli yönleri (örneğin, teknikler, betikler), sentetik metin veri üretimi için uyarlanabilir veya bilgilendirici olabilir. Birincil odak noktası açıklama olsa da, bu depoda LLM'leri veri oluşturma için promptlama ve yönlendirme ile ilgili teknikler ve içgörüler, sentetik veri üretimi görevlerine aktarılabilir. LLM'lerden belirli veri türlerini elde etme zorlukları, hem açıklama hem de sentetik veri üretimi için ortaktır. İlgili görevlere odaklanan depoları keşfetmek, değerli stratejiler ve yaklaşımlar ortaya çıkarabilir.
* **3.4. Kiln-AI/Kiln** (<https://github.com/Kiln-AI/Kiln>) 37  
  Bu depo, LLM'leri potansiyel olarak kullanan etkileşimli bir sentetik veri üretimi özelliği içeren bir araç olarak tanımlanmaktadır.37 Depoda bahsedilen, sentetik metin veri üretimiyle ilgili belirli işlevler veya özellikler bulunmaktadır. Kiln gibi, açıkça sentetik veri üretimi yetenekleri sunan araçlar, genel amaçlı kütüphaneler kullanarak sentetik veri üretimi iş akışlarını sıfırdan oluşturmak yerine, LLM'leri kullanarak sentetik veri kümeleri oluşturmak için daha kullanıcı dostu bir arayüz ve önceden oluşturulmuş işlevler sağlayabilir.
* **3.5. SciPhi-AI/synthesizer** ((<https://github.com/SciPhi-AI/synthesizer>)) 37  
  Bu depo, sentetik veriyi mevcut verilere dayandırmak için AgentSearch ile entegre edilebilen bir sentezleyici aracı olarak tanımlanmaktadır.37 Mevcut verilere dayandırma yeteneği, daha gerçekçi ve bağlamsal olarak ilgili sentetik metin üretmek için faydalı olabilir. Sentetik veri üretimini gerçek dünya bilgileriyle dayandırma yeteneği, LLM tabanlı sentetik veri oluşturmada temel bir zorluk olan, üretilen metnin kalitesini ve doğruluğunu önemli ölçüde artırabilir. LLM'ler bazen yanlış veya anlamsız bilgiler üretebilir (halüsinasyonlar). Bu tür sentezleyiciler, onlara mevcut veri kümelerinden ilgili bağlam sağlayarak bu sorunu hafifletmeye yardımcı olabilir.
* **3.6. Open R1** (<https://github.com/grpo727/open_r1>) 38  
  Bu depo, DeepSeek-R1'i yeniden üretmeye odaklanmakta ve Distilabel tekniğini kullanarak sentetik veri üretimi için betikler içermektedir.38 Bu tür araştırma odaklı depolardan gelişmiş sentetik veri üretimi tekniklerini öğrenme potansiyeli bulunmaktadır. En son modelleri ve teknikleri yeniden üretmeyi amaçlayan depoları incelemek, LLM'leri kullanarak sentetik veri üretimi için en son yaklaşımlar hakkında bilgi sağlayabilir. Araştırmacılar genellikle model performansını artırmak için yeni sentetik veri üretimi yöntemleri denemektedir. Onların uygulamalarını keşfetmek, değerli öğrenme fırsatları ve belirli sentetik veri zorlukları için potansiyel çözümler sunabilir.
* **3.7. mlabonne/llm-datasets** (<https://github.com/mlabonne/llm-datasets>) 39  
  Bu depo, bazıları sentetik olarak oluşturulmuş olabilecek bir LLM veri kümeleri koleksiyonu olarak tanımlanmaktadır.39 Özellikle sentetik olarak işaretlenmiş veya LLM tabanlı yöntemler kullanılarak oluşturulmuş ve örnek veya kıyaslama olarak hizmet edebilecek veri kümeleri bulunmaktadır. Mevcut sentetik veri kümelerini keşfetmek, oluşturulan veri türleri, kullanılan yöntemler ve potansiyel uygulamalar hakkında bilgi sağlayabilir. Herkese açık sentetik veri kümelerini analiz etmek, sentetik veri üretimi alanının mevcut durumunu anlamaya ve kendi görevleri için ilgili olabilecek veri kümelerini belirlemeye yardımcı olabilir.

**4. LLM Sentetik Veri Üretiminde Dikkat Edilmesi Gerekenler ve Teknikler**

* **4.1. Prompt Mühendisliği** 1  
  İyi tasarlanmış promptlar, LLM'leri istenen sentetik metni üretmeye yönlendirmede çok önemli bir rol oynar.1 Bağlam sağlama, formatı belirtme, örnekler verme (birkaç atış öğrenimi) ve persona tabanlı promptlar kullanma gibi teknikler bu süreçte önemlidir. Promptun kalitesi ve özgüllüğü, üretilen sentetik verinin kalitesini ve alaka düzeyini önemli ölçüde etkiler. Etkili prompt mühendisliği, istenen sonuçları elde etmek için gereklidir. LLM'ler, promptta sağlanan talimatlara ve bağlama göre yanıt verir. Görevi, istenen formatı ve belirli kısıtlamaları açıkça tanımlayan promptlar oluşturmak, faydalı sentetik veri üretmek için hayati öneme sahiptir.
* **4.2. Veri Kalitesi ve Çeşitliliği** 1  
  LLM tarafından üretilen sentetik verinin kalitesini, doğruluğunu ve çeşitliliğini sağlama zorlukları ele alınmalıdır. Halüsinasyonlar, stilistik gerçekçilik eksikliği ve doğal önyargılar gibi potansiyel sorunlar göz önünde bulundurulmalıdır.1 Veri kalitesini iyileştirmek için filtreleme, çıktıları ağırlıklandırma, bilgiye erişimli üretim ve yinelemeli kendini iyileştirme gibi teknikler tartışılmalıdır.1 Farklı promptlama teknikleri kullanma, üretim parametrelerini değiştirme ve kümeleme tabanlı yaklaşımlar kullanma gibi çeşitliliği artırma stratejileri açıklanmalıdır.3 Sentetik verinin etkinliği için veri kalitesi ve çeşitliliği arasında bir denge kurmak önemlidir. Aşırı homojen veya yanlış sentetik veri, model eğitimi veya değerlendirmesi için istenen faydaları sağlamayabilir. LLM'ler çeşitli metinler üretebilse de, üretilen verinin yüksek kalitede olması ve gerçek dünya verilerinin özelliklerini doğru bir şekilde yansıtması önemlidir. Çıktıyı filtreleme, iyileştirme ve çeşitlendirme teknikleri gereklidir.
* **4.3. Önyargıları Ele Alma** 1  
  Önyargılı gerçek dünya verileri üzerinde eğitilmiş LLM'ler tarafından üretilen sentetik verilerde önyargıların artma riski vurgulanmalı ve bu önyargıları azaltma yöntemleri tartışılmalıdır.1 Sentetik veri, LLM'nin eğitim verilerinde bulunan önyargıları istemeden devralabilir ve hatta artırabilir. Adalet ve zararlı sonuçları önlemek için dikkatli değerlendirme ve azaltma stratejileri gereklidir. LLM'ler eğitildikleri verilerden öğrenirler. Bu veriler cinsiyet, ırk veya diğer hassas özelliklerle ilgili önyargılar içeriyorsa, model tarafından üretilen sentetik veri de bu önyargıları yansıtabilir. Bu sorunun farkında olmak ve sentetik veriyi önyargılardan arındırmak için teknikler kullanmak önemlidir.
* **4.4. Sentetik Verinin Değerlendirilmesi** 4  
  Sentetik verinin faydasını ve doğruluğunu değerlendirmenin önemi kısaca değinilmeli, kalitesini ve amaçlanan görev için uygunluğunu değerlendirmek için kullanılan metrikler ve yöntemler belirtilmelidir.4 Model eğitimi veya değerlendirmesi gibi kritik görevler için kullanmadan önce sentetik verinin kalitesini ve kullanışlılığını değerlendirmek esastır. Sentetik verinin gerekli standartları karşıladığından emin olmak için uygun metrikler ve değerlendirme yöntemleri kullanılmalıdır. Her sentetik veri eşit kalitede değildir. Gerçek dünya verilerine benzerliğini, aşağı akış model performansına etkisini ve istenen özellikleri temsil etme yeteneğini değerlendirmek, sentetik veri üretimi sürecinde çok önemli adımlardır.

**5. Sonuç**

Bu rapor, LLM tabanlı sentetik metin veri üretimi için değerli kaynaklar sağlayan temel Python kütüphanelerini (Hugging Face Transformers, OpenAI Python Kütüphanesi, LangChain ve diğerleri) ve GitHub depolarını özetlemektedir. Prompt mühendisliği gibi tekniklerin ve sentetik veri üretimi sürecinde veri kalitesi, çeşitliliği ve önyargı konularındaki hususların önemi vurgulanmıştır. LLM tabanlı sentetik metin veri üretiminin, veri kıtlığını gidermek, maliyetleri düşürmek ve NLP modellerinin performansını ve dayanıklılığını artırmak için taşıdığı potansiyel yeniden teyit edilmiştir. Sonuç olarak, bu alanın yapay zeka alanındaki sürekli ilerlemeler ve artan önemiyle birlikte geleceğe yönelik umut verici bir görünüm sunmaktadır.

**Tablo 1: LLM Tabanlı Sentetik Metin Veri Üretimi için Temel Python Kütüphanelerinin Özeti**

| **Kütüphane Adı** | **Açıklama** | **Sentetik Veri Üretimi için Temel Özellikler** | **Bağlantı** |
| --- | --- | --- | --- |
| Hugging Face Transformers | Önceden eğitilmiş LLM'ler ve NLP görevleri için araçlar sunan merkezi kütüphane. | generate() metodu, çeşitli kod çözme stratejileri, GenerationConfig ile özelleştirme. | <https://huggingface.co/docs/transformers/> |
| OpenAI Python Kütüphanesi | OpenAI'nin dil modellerine (GPT-3.5, GPT-4 vb.) arayüz sağlar. | openai.ChatCompletion.create() ve openai.Completion.create() metotları, çeşitli parametrelerle kontrol. | <https://platform.openai.com/docs/api-reference> |
| LangChain | LLM'ler tarafından desteklenen uygulamalar oluşturmak için çerçeve. | Prompt zincirleme, harici veri kaynakları entegrasyonu, belge parçalama, bağlam ve sorgu oluşturma. | <https://www.langchain.com/> |
| ydata-synthetic | GAN tabanlı modeller dahil çeşitli sentetik veri üretimi teknikleri sunar. | CTGAN, TimeGAN, GMM tabanlı modeller. | <https://ydata.ai/solutions/synthetic-data> |
| SDV (Synthetic Data Vault) | Klasik ML ve derin öğrenme modellerini içeren çeşitli sentezleyiciler sağlar. | Tek ve çok tablolu veri desteği, anonimleştirme kontrolü, kısıtlama uygulama. | <https://sdv.dev/> |

**Tablo 2: LLM Tabanlı Sentetik Metin Veri Üretimi için Temel GitHub Depolarının Özeti**

| **Depo Adı** | **Açıklama** | **Temel Özellikler/Önemli Noktalar** | **Bağlantı** |
| --- | --- | --- | --- |
| Awesome-LLM-Synthetic-Data | LLM'lerin sentetik verileri hakkında derlenmiş makaleler, araçlar ve bloglar listesi. | Anketler, yöntemler, uygulama alanları, veri kümeleri, araçlar ve bloglar. | (https://github.com/wasiahmad/Awesome-LLM-Synthetic-Data) |
| LLM-Synthetic-Data | LLM odaklı sentetik verilere odaklanan kaynak koleksiyonu. | GitHub bağlantıları, blog gönderileri, ön eğitim, talimat ayarlama konuları. | (https://github.com/pengr/LLM-Synthetic-Data) |
| Zhen-Tan-dmml/LLM4Annotation | LLM'leri açıklama için kullanmaya odaklanır. | LLM'leri veri oluşturma için promptlama ve yönlendirme teknikleri. | (https://github.com/Zhen-Tan-dmml/LLM4Annotation) |
| Kiln-AI/Kiln | Etkileşimli bir sentetik veri üretimi özelliği içeren bir araç. | Kullanıcı dostu arayüz, sentetik veri üretimi için önceden oluşturulmuş işlevler. | <https://github.com/Kiln-AI/Kiln> |
| SciPhi-AI/synthesizer | Sentetik veriyi mevcut verilere dayandırmak için AgentSearch ile entegre edilebilir. | Sentetik veri üretimini gerçek dünya bilgileriyle dayandırma yeteneği. | (https://github.com/SciPhi-AI/synthesizer) |
| Open R1 | DeepSeek-R1'i yeniden üretmeye odaklanır. | Sentetik veri üretimi için Distilabel tekniğini kullanan betikler. | <https://github.com/grpo727/open_r1> |
| mlabonne/llm-datasets | Bazıları sentetik olarak oluşturulmuş LLM veri kümeleri koleksiyonu. | Sentetik olarak işaretlenmiş veya LLM tabanlı yöntemlerle oluşturulmuş veri kümeleri. | <https://github.com/mlabonne/llm-datasets> |

#### Alıntılanan çalışmalar

1. [2503.14023] Synthetic Data Generation Using Large Language Models: Advances in Text and Code - arXiv, erişim tarihi Nisan 10, 2025, <https://arxiv.org/abs/2503.14023>
2. Large language models generating synthetic clinical datasets: a feasibility and comparative analysis with real-world perioperative data - Frontiers, erişim tarihi Nisan 10, 2025, <https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2025.1533508/full>
3. Synthetic data generation (Part 1) | OpenAI Cookbook, erişim tarihi Nisan 10, 2025, <https://cookbook.openai.com/examples/sdg1>
4. LLM-Driven Synthetic Data Generation, Curation & Evaluation | by ..., erişim tarihi Nisan 10, 2025, <https://cobusgreyling.medium.com/llm-driven-synthetic-data-generation-curation-evaluation-33731e33b525>
5. Synthetic Data Generation Using Large Language Models: Advances in Text and Code, erişim tarihi Nisan 10, 2025, <https://arxiv.org/html/2503.14023v1>
6. Synthetic Data Generator Simplifies Dataset Creation with Large Language Models - InfoQ, erişim tarihi Nisan 10, 2025, <https://www.infoq.com/news/2025/01/synthetic-data-generator/>
7. What is Text Generation? - Hugging Face, erişim tarihi Nisan 10, 2025, <https://huggingface.co/tasks/text-generation>
8. Generation strategies - Hugging Face, erişim tarihi Nisan 10, 2025, <https://huggingface.co/docs/transformers/v4.27.0/generation_strategies>
9. Text generation - Hugging Face, erişim tarihi Nisan 10, 2025, <https://huggingface.co/docs/transformers/llm_tutorial>
10. Generating text word by word - Transformers - Hugging Face Forums, erişim tarihi Nisan 10, 2025, <https://discuss.huggingface.co/t/generating-text-word-by-word/29312>
11. Generation - Hugging Face, erişim tarihi Nisan 10, 2025, <https://huggingface.co/docs/transformers/main_classes/text_generation>
12. Text Generation with Hugging Face Transformers: A Beginner's Guide | by Ganesh Lokare, erişim tarihi Nisan 10, 2025, <https://medium.com/@lokaregns/text-generation-with-hugging-face-transformers-a-beginners-guide-6b0b4b957379>
13. How to generate text: using different decoding methods for language generation with Transformers - Hugging Face, erişim tarihi Nisan 10, 2025, <https://huggingface.co/blog/how-to-generate>
14. How to generate text using GPT2 model with Huggingface transformers? - Stack Overflow, erişim tarihi Nisan 10, 2025, <https://stackoverflow.com/questions/76663419/how-to-generate-text-using-gpt2-model-with-huggingface-transformers>
15. Generating text while model is still training - Beginners - Hugging Face Forums, erişim tarihi Nisan 10, 2025, <https://discuss.huggingface.co/t/generating-text-while-model-is-still-training/57482>
16. 10 Must-Know Python Libraries for LLMs in 2025 - MachineLearningMastery.com, erişim tarihi Nisan 10, 2025, <https://machinelearningmastery.com/10-must-know-python-libraries-for-llms-in-2025/>
17. Top 50 Python Libraries to Know in 2025 - Analytics Vidhya, erişim tarihi Nisan 10, 2025, <https://www.analyticsvidhya.com/blog/2024/12/python-libraries/>
18. 8 Top Open-Source LLMs for 2024 and Their Uses - DataCamp, erişim tarihi Nisan 10, 2025, <https://www.datacamp.com/blog/top-open-source-llms>
19. How can I use OpenAI for text generation? - Milvus, erişim tarihi Nisan 10, 2025, <https://milvus.io/ai-quick-reference/how-can-i-use-openai-for-text-generation>
20. Developer quickstart - OpenAI API, erişim tarihi Nisan 10, 2025, <https://platform.openai.com/docs/quickstart>
21. Advanced usage - OpenAI API, erişim tarihi Nisan 10, 2025, <https://platform.openai.com/docs/advanced-usage>
22. How to Generate Text with OpenAI, GPT-3, and Python - Matt on ML.NET - Accessible AI, erişim tarihi Nisan 10, 2025, <https://accessibleai.dev/post/generating_text_with_gpt_and_python/>
23. Building a Tool To Generate Text With OpenAI's GPT-4 Model - DZone, erişim tarihi Nisan 10, 2025, <https://dzone.com/articles/building-a-tool-to-generate-text-with-gpt4>
24. data-winners/generation-api-openai/openai-text-generation-examples-in-python.ipynb at main - GitHub, erişim tarihi Nisan 10, 2025, <https://github.com/FrontAnalyticsInc/data-winners/blob/main/generation-api-openai/openai-text-generation-examples-in-python.ipynb>
25. Text Generation: I'm struggling with Output Consistency - API - OpenAI Developer Forum, erişim tarihi Nisan 10, 2025, <https://community.openai.com/t/text-generation-im-struggling-with-output-consistency/539537>
26. GPT-4 API Guide With Examples (All Scripts Included), erişim tarihi Nisan 10, 2025, <https://hasanaboulhasan.medium.com/gpt-4-api-guide-with-examples-all-scripts-included-afa1b6a5c3d9>
27. OpenAI Chat Completion API - OpenAI Platform, erişim tarihi Nisan 10, 2025, <https://platform.openai.com/docs/guides/text-generation>
28. Text-Generation App with OpenAI Python API & Streamlit - YouTube, erişim tarihi Nisan 10, 2025, <https://www.youtube.com/watch?v=WPlSpjiv4tk>
29. Using LLMs for Synthetic Data Generation: The Definitive Guide ..., erişim tarihi Nisan 10, 2025, <https://www.confident-ai.com/blog/the-definitive-guide-to-synthetic-data-generation-using-llms>
30. Generating Synthetic Dataset Using LLM - DeepLearning.AI, erişim tarihi Nisan 10, 2025, <https://community.deeplearning.ai/t/generating-synthetic-dataset-using-llm/737049>
31. The Top 5 Python Packages to Generate Realistic Synthetic Data - YData, erişim tarihi Nisan 10, 2025, <https://ydata.ai/resources/top-5-packages-python-synthetic-data>
32. The Synthetic Data SDK - An open-source python toolkit for high-fidelity privacy-safe ... - Mostly AI, erişim tarihi Nisan 10, 2025, <https://mostly.ai/blog/the-synthetic-data-sdk-an-open-source-python-toolkit-for-high-fidelity-privacy-safe-synthetic-data>
33. Real-time updated, fine-grained reading list on LLM-synthetic-data. - GitHub, erişim tarihi Nisan 10, 2025, <https://github.com/pengr/LLM-Synthetic-Data>
34. README.md - pengr/LLM-Synthetic-Data - GitHub, erişim tarihi Nisan 10, 2025, <https://github.com/pengr/LLM-Synthetic-Data/blob/main/README.md>
35. wasiahmad/Awesome-LLM-Synthetic-Data - GitHub, erişim tarihi Nisan 10, 2025, <https://github.com/wasiahmad/Awesome-LLM-Synthetic-Data>
36. Zhen-Tan-dmml/LLM4Annotation - GitHub, erişim tarihi Nisan 10, 2025, <https://github.com/Zhen-Tan-dmml/LLM4Annotation>
37. In 2024, what is the best tool/framework for creating synthetic data that can then be used to fine-tune with? - Reddit, erişim tarihi Nisan 10, 2025, <https://www.reddit.com/r/LocalLLaMA/comments/194m01m/in_2024_what_is_the_best_toolframework_for/>
38. Top GitHub Projects of Jan 2025 - OpenCV, erişim tarihi Nisan 10, 2025, <https://opencv.org/blog/top-github-projects-jan-2025/>
39. mlabonne/llm-datasets: Curated list of datasets and tools for post-training. - GitHub, erişim tarihi Nisan 10, 2025, <https://github.com/mlabonne/llm-datasets>
40. Synthetic Data Generation with Large Language Models for Text Classification: Potential and Limitations | OpenReview, erişim tarihi Nisan 10, 2025, [https://openreview.net/forum?id=MmBjKmHIND&referrer=%5Bthe%20profile%20of%20Zhuoyan%20Li%5D(%2Fprofile%3Fid%3D~Zhuoyan\_Li2)](https://openreview.net/forum?id=MmBjKmHIND&referrer=%5Bthe+profile+of+Zhuoyan+Li%5D(/profile?id%3D~Zhuoyan_Li2))
41. Synthetic Data Generation with Large Language Models for Text Classification: Potential and Limitations - ACL Anthology, erişim tarihi Nisan 10, 2025, <https://aclanthology.org/2023.emnlp-main.647.pdf>
42. Few-shot LLM Synthetic Data with Distribution Matching - arXiv, erişim tarihi Nisan 10, 2025, <https://arxiv.org/html/2502.08661v1>
43. Data Generation Using Large Language Models for Text Classification: An Empirical Case Study - arXiv, erişim tarihi Nisan 10, 2025, <https://arxiv.org/html/2407.12813v2>
44. [2310.07849] Synthetic Data Generation with Large Language Models for Text Classification: Potential and Limitations - arXiv, erişim tarihi Nisan 10, 2025, <https://arxiv.org/abs/2310.07849>
45. On LLMs-Driven Synthetic Data Generation, Curation, and Evaluation: A Survey - arXiv, erişim tarihi Nisan 10, 2025, <https://arxiv.org/html/2406.15126v1>