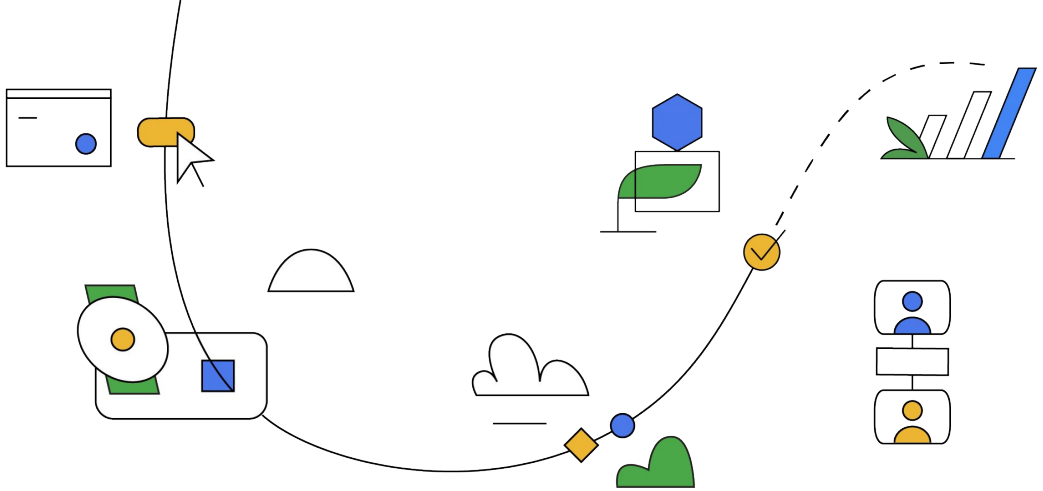
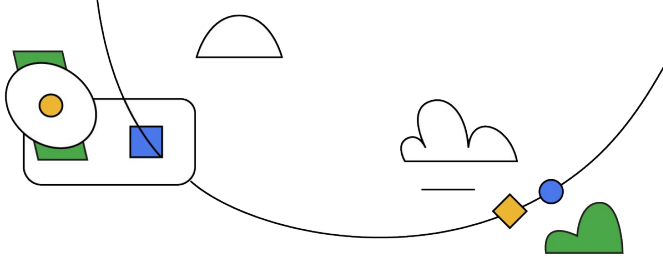


Halüsinasyon Makinesini Yönetmek: LLM & RAG

Google for Developers

Hoşgeldiniz !

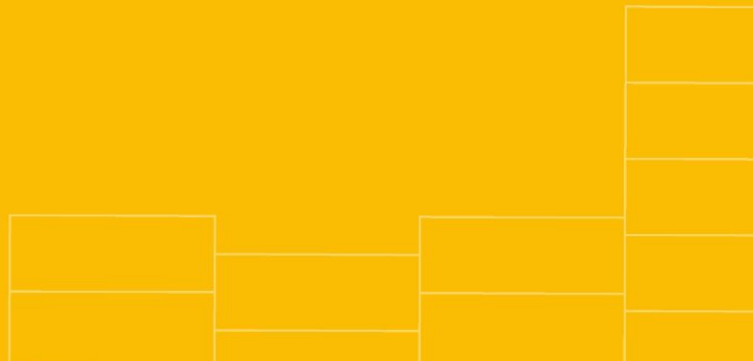
Onuralp SEZER
Armada Yazılım - Senior Software Engineer
Machine Learning Enthusiastic
TensorFlow Developer



Neler Hakkında Konuşacağız ?

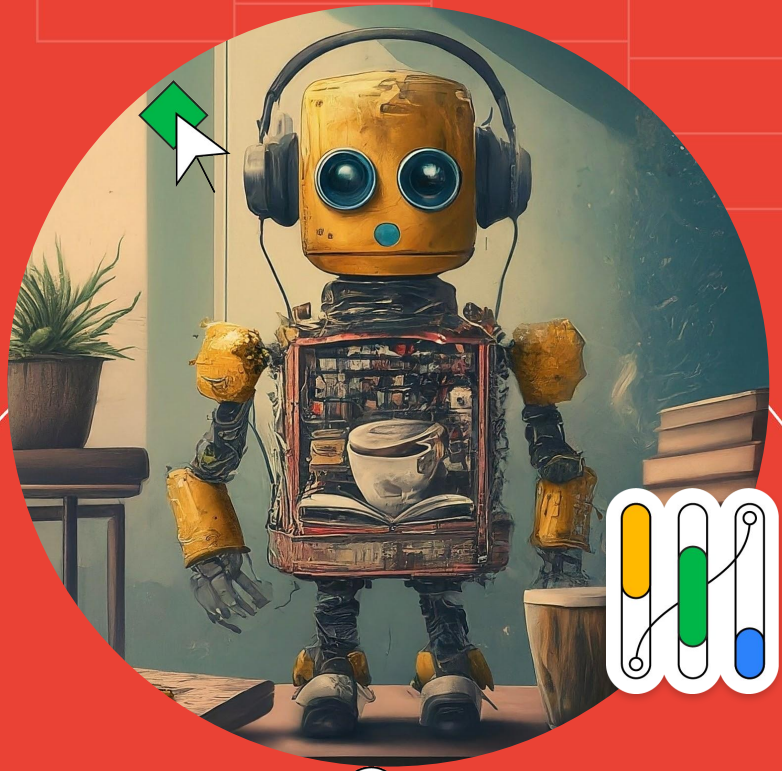
Large Language Model

Retrieval Augmented Generation



Google for Developers

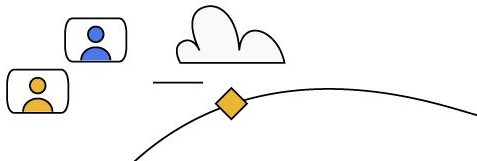
Large Language Model (LLM) Nedir ?



Google for Developers

Large Language Model ...

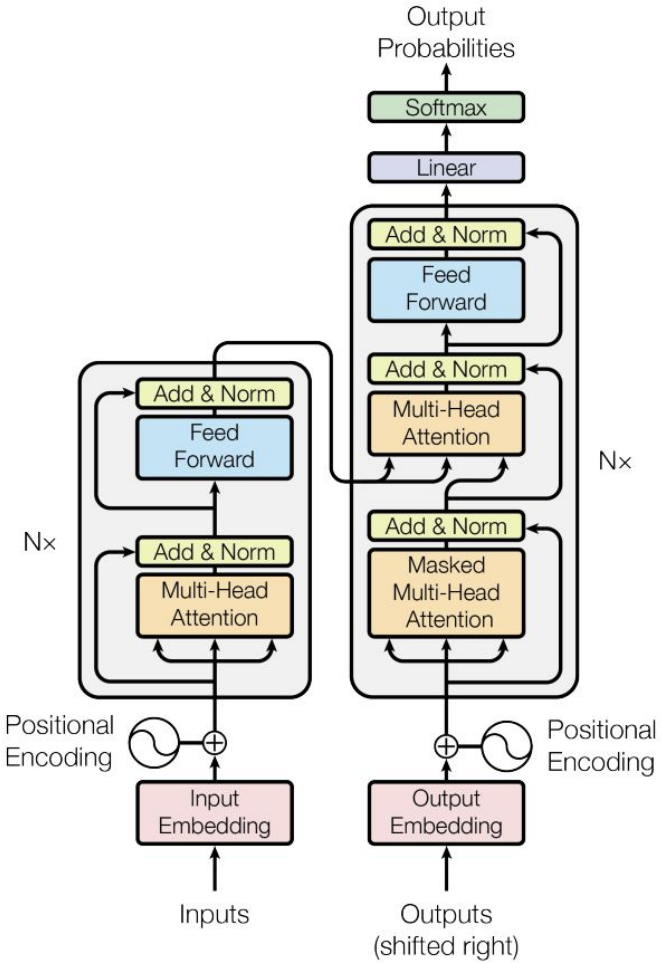
Büyük veri kümelerinden elde edilen bilgilere dayalı olarak metni ve diğer içerik biçimlerini tanıyabilen, özetleyebilir, tercüme edebilen, çıkarım yapabilen ve yeni içerikler oluşturabilen bir derin öğrenme algoritmasıdır.Large Language Models, Transformers modellerinin en başarılı uygulamaları arasındadır.



Transformers ?

Transformers, bu cümledeki kelimeler gibi ardışık verilerdeki ilişkileri izleyerek bağlamı ve dolayısıyla anlamı öğrenen bir neural network mimarisidir. Transformer modelleri, bir serideki uzak veri öğelerinin bile birbirini nasıl etkilediğini ve birbirine bağımlı olduğunu tespit etmek için dikkat veya kendine dikkat olarak adlandırılan gelişen bir dizi matematiksel tekniği uygular.

- https://arxiv.org/abs/1706.03762 Attention Is All You Need - Vaswani 2017 (Google)
- https://research.google/blog/transformer-a-novel-neural-network-architecture-for-language-understanding/



LLM Terimleri

LLM Terimleri

Top-K Sampling

Her adımda, yalnızca en olası K token (birim) dikkate alınarak, bu alt kümeden rastgele bir token seçilen bir çözümleme stratejisidir. **Bu, daha çeşitli ve tutarlı metinler üretmeye yardımcı olur.**

Temperature

LLM çıktının daha rastgele ve yaratıcı mı yoksa daha öngörülebilir mi olacağını belirleyen bir parametredir. Daha yüksek bir değer daha düşük olasılıkla, yani daha yaratıcı çıktılarla sonuçlanacaktır.

Token

LLM tarafından işlenen metnin en küçük birimi olup, **bu birim bir kelime, alt kelime veya karakter olabilir.** Tokenlar, modelin metni üretmek ve anlamak için kullandığı temel yapı taşlarıdır.

LLM Terimleri

Top-P

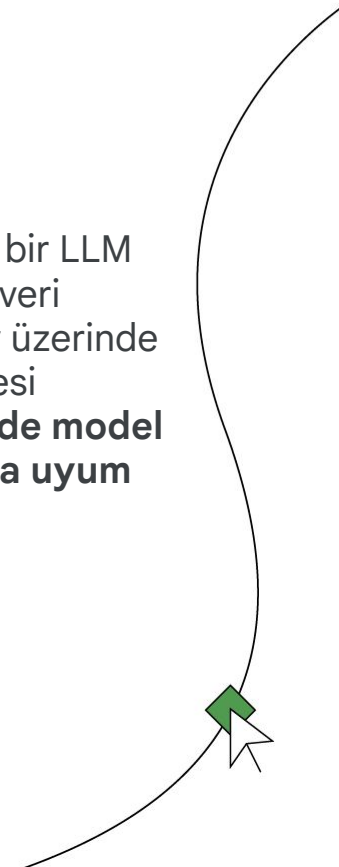
Modelin ne kadar deterministik olacağını kontrol etmenizi sağlar. **Kesin ve gerçeklere dayalı cevaplar arıyorsanız bu değeri düşük tutun. Daha çeşitli değerler arıyorsanız yüksek bir değer girebiliriz.**

Prompt

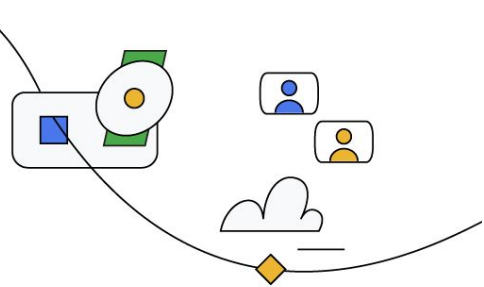
Metin üretmesi için verilen başlangıç metnidir. İsteğin kalitesi ve özgüllüğü modelin çıktısını önemli ölçüde etkileyebilir.

Fine-Tuning

Önceden eğitilmiş bir LLM modelini belirli bir veri kümesi veya görev üzerinde daha fazla eğitilmesi sürecidir. **Bu sayede model özel uygulamalara uyum sağlar.**



LLM Terimleri



Context Size

Bir LLM'nin bir seferde dikkate alabileceği maksimum token sayısıdır. Bu modelin yanıt üretmek veya anlamak için kullanabileceği metin miktarını sınırlar.

Quantization

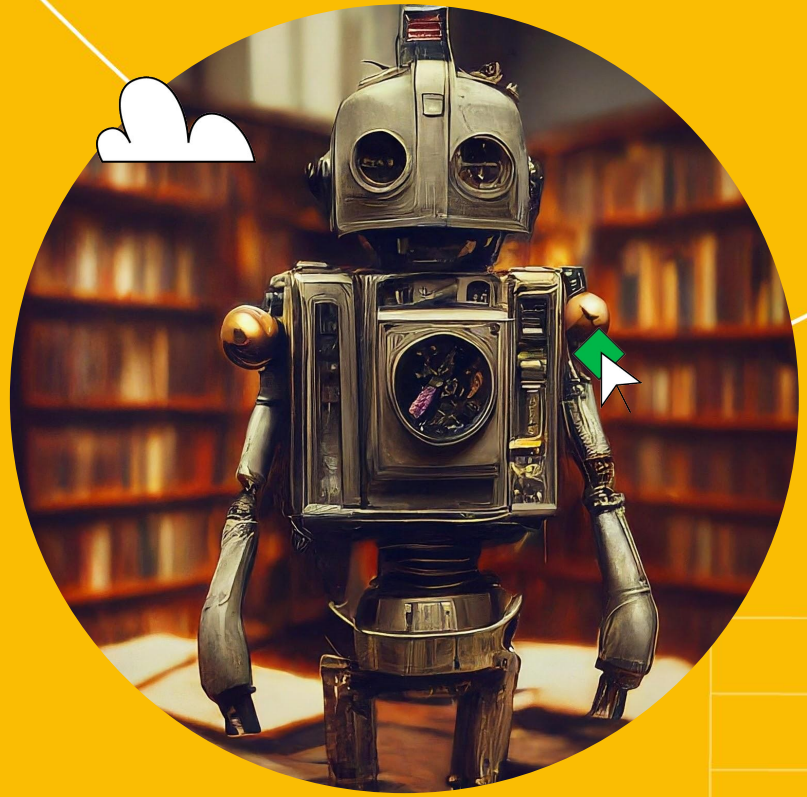
Bir modelin bellekteki boyutunu azaltmak için ağırlıkları ve aktivasyonları daha az sayıda bit ile temsil etme işlemidir. Bu, modelin depolama ve işlemci kaynakları açısından daha verimli olmasını sağlar.

Pre-Training

Bir LLM'nin genel dil kalıplarını ve bilgilerini öğrenmesi için büyük bir metin verisi üzerinde gerçekleştirilen **ilk eğitim aşamasıdır.** Bu aşama genellikle belirli görevler için ince ayar ile takip edilir.

Demo Zamanı ! :)

Retrieval Augmented Generation (RAG)



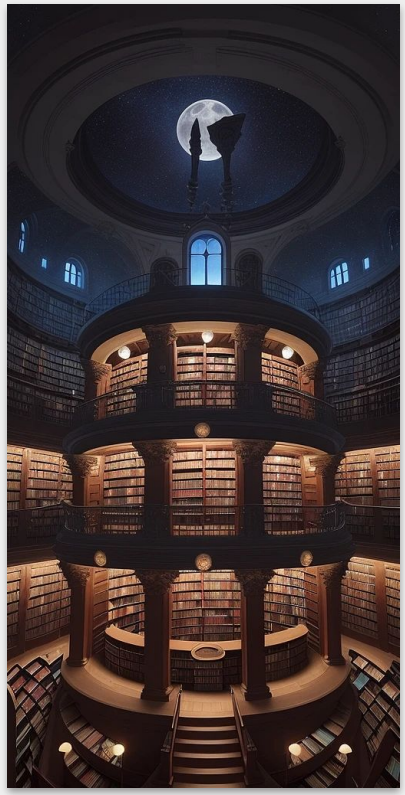
Google for Developers

RAG

RAG, Büyük Dil Modellerinin (LLM'ler) **güvenilirliğini ve doğruluğunu** geliştirmek için tasarlanmış bir yapay zeka (AI) tekniğidir. LLM'ler, metin üretme, dilleri çevirme ve soruları yanıtlama gibi birçok görevde oldukça başarılı olsa da, **yanlış veya yanıltıcı bilgiler** üretme eğiliminde de olabilirler.

RAG, bu sorunu şu şekilde çözer:

1. **Retrieval :** LLM, önce kullanıcının sorgusuyla **ilgili bilgileri** bir bilgi tabanından **arar**.
2. **Augmentation:** Ardından, bulunan bilgiler **sorgu ile zenginleştirilir**.
3. **Generation :** Son olarak, LLM, **zenginleştirilmiş sorguyu** kullanarak **doğru ve bilgilendirici bir yanıt** üretir.



RAG Avantajları

Daha Doğru Bilgi

LLM'lerin **yanlış veya yanıltıcı bilgiler üretme olasılığını azaltır.**

Artırılmış Kapsam

LLM'ler, **gerçek dünya bilgilerine erişerek** daha kapsamlı ve bilgilendirici yanıtlar üretebilir.

Geliştirilmiş Güvenilirlik

LLM'lerin daha güvenilir olmasını sağlayarak **önemli kararlar verirken kullanılmalarını daha güvenli hale getirir.**

RAG as Service

RAG as Service

LLM kullanımını **API ile yaptığımız servis üzerinden RAG kullanımıdır.**

Kullanım Kolaylığı

Sunucu kurulumuna ihtiyaç olmayan direk servis üzerinde hazır şekilde kullanabileceğimiz servis şeklinde sunulur.

Sınırlamalar

Her ne kadar RAG as Service kullanımı kolay olsada **yüklenebilen veri limiti, işlem sınırı gibi unsurlar yüksek gereksinim gerektiren kullanımlarda yetersiz kalabilir.** Özel ihtiyaçlara cevap vermeyebilir.

RAG on Premises

RAG on Premises (Prem)

RAG servisini **kendi sunucumuzda** veya localimizde kullanıma verilen addır.

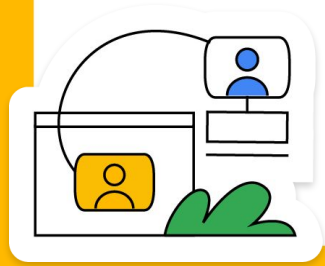
Sunucu bilgisi

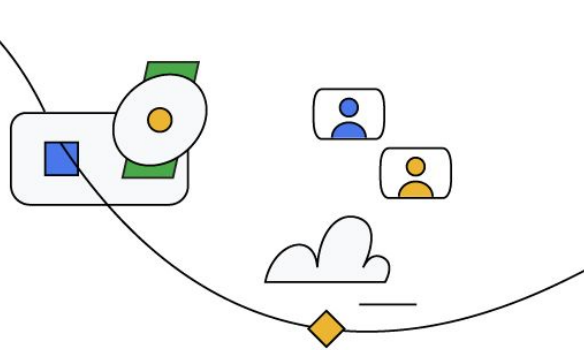
Sunucu kurulumuna ihtiyaç olan ve on Pren LLM veya LLM servisi ile **gerekli network ayarları yapılarak kullanabilmektedir.**

Avantajlar

Yüksek miktarda bilgi yükleyebilme avantajı, **daha büyük ve özel ihtiyaçlara cevap verebilme avantajının olmasıdır.**
Aynı zamanda gizlilik gerektiren projelerde servise göre daha iyi bir s[illegible]mdir.

Demo Zamanı ! :)





Little Little into the Middle Series

End of Chapter Two

Onuralp SEZER