

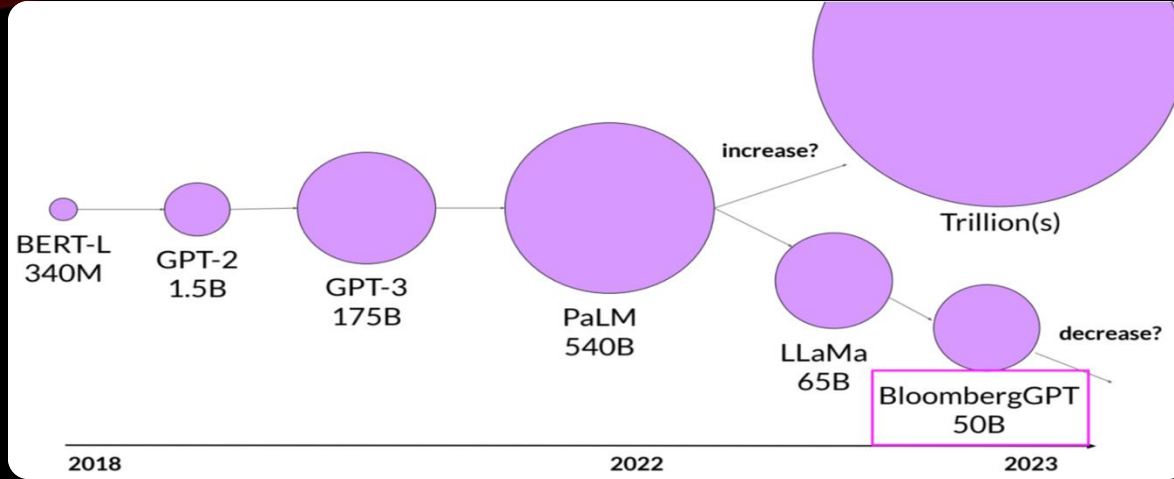
# Farklı Model Mimarileri ve Kullanım Alanları

Eğitmen  
Kubilay Tuna

# Ders İeriđi

1. Popler LLM'ler ve Teknik Detaylar
2. LLM'lerin Yetenekleri ve Sınırlamaları
3. LLM'lerin Kullanım Alanları ve rnek Uygulamalar
4. Diđer nemli Dil Modeli Mimarileri
5. LLM'lerin Zorlukları ve Limitleri
6. Geleceđe Ynelik Perspektifler

# Bert, GPT ve LLama: Temel Modeller



Bu modeller, LLM'lerin evriminde önemli dönüm noktalarıdır.

**BERT:** Dil anlama üzerine optimize edilmiş, iki yönlü Transformer modelidir ve özellikle dil anlama görevlerinde güçlüdür.

**GPT:** Dil üretiminde güçlü, tek yönlü bir Transformer modelidir ve cümle üretimi ve yaratıcı dil görevlerinde çok başarılıdır.

**LLama:** Meta tarafından geliştirilen, yüksek verimlilik ve çeşitli dil görevlerinde başarılı performans sunan, daha az parametreyle verimli sonuçlar sağlayan modeldir.

# BERT AİLESİ: BİDIRECTIONAL DEVRİM

## **BERT (Bidirectional Encoder Representations from Transformers)**, 2018 yılında Google AI

tarafından tanıtılan ve doğal dil işlemede devrim yaratan bir dil modelidir. En büyük yeniliği, dilin her iki yönünü (öncesi ve sonrası) dikkate alarak bağlamı daha derinlemesine anlayabilmesidir.

BERT yalnızca encoder bloklarını kullanır (GPT yalnızca decoder kullanır).

**Her encoder bloğu;** multi-head, self-attention, residual connections, layer normalization ve feed-forward layer'dan oluşur.

Geleneksel modeller ya soldan sağa ya da sağdan sola bakarken, BERT, maskelenmiş token dışında tüm çevre bağlamı görebilir (true bidirectionality). Bu sayede, kelimelerin çok daha anlamlı bağlam temsilleri öğrenilir.

Girdi token'larının sırasını göstermek için sabit veya öğrenilebilir pozisyon embedding'leri kullanılır.

BERT'in bir çok varyansı ve optimize edilmiş hali vardır.

- **RoBERTa (Facebook AI):** Performansı birçok NLP görevinde BERT'i geçti
- **DistilBERT (HuggingFace):** Bilgi sıkıştırma (knowledge distillation) ile daha küçük ve hızlı
- **ALBERT (Google):** Daha az parametre ile daha büyük modellerin eğitilmesini sağladı
- **SpanBERT:** Maskeleye tek token yerine span bazlı
- **TinyBERT:** Mobil cihazlar için ultra kompakt versiyon
- **ClinicalBERT, BioBERT:** Medikal NLP için özelleştirilmiş BERT modelleri

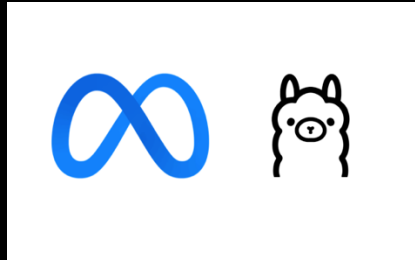
# GPT AİLESİ: OTOREGRESİF TRANSFORMER'LARIN EVRİMİ

Model	Parametre Sayısı	Yayın Yılı	Maks. Giriş Uzunluğu	Öne Çıkan Özellikler
GPT	117M	2018	512	İlk nesil, küçük ölçekli
GPT-2	1.5B	2019	1024	Büyük web verisi ile eğitildi
GPT-3	175B	2020	2048	Few-shot öğrenme kabiliyeti
GPT-4	? (500B+ tahmini)	2023	32k+ (bazı sürümleri)	Multimodal, daha stabil & tutarlı

- **Transformer Decoder tabanlı yapı:** Sadece decoder blokları kullanılır.
- **Causal (maskeli) attention:** Gelecekteki token'lara bakmadan geçmişe göre tahmin yapar.
- LayerNorm + GELU + Residual Connections
- **Position Embeddings:** Sabit (sinüs-kosinüs) veya öğrenilebilir.

# LLAMA AİLESİ: AÇIK KAYNAK BÜYÜK DİL MODELLERİ DEVRİMİ

Model	Parametre Sayısı	Bağlam Penceresi	Eğitildiği Token
LLaMA-2-7B	7B	4096	2 Trilyon
LLaMA-2-13B	13B	4096	2 Trilyon
LLaMA-2-70B	70B	4096	2 Trilyon



Model	Parametre Sayısı	Eğitim Verisi	Token Sayısı
LLaMA-7B	7 milyar	1.4T token	2048
LLaMA-13B	13 milyar	1.4T token	2048
LLaMA-30B	30 milyar	1.4T token	2048
LLaMA-65B	65 milyar	1.4T token	2048

**LLaMA (Large Language Model Meta AI)**, Meta (Facebook AI) tarafından geliştirilen, **verimli, yüksek performanslı ve açık kaynak** bir büyük dil modeli ailesidir. GPT-3 gibi dev modellere rakip olacak şekilde ancak daha az veri ve daha düşük compute ile benzer performans göstermeyi hedefler.

İlk sürüm 2023'te tanıtılmıştır ve açık kaynak topluluğu üzerinde çok büyük bir etki yaratmıştır.

- Decoder-only transformer (GPT gibi)
- Causal attention (sadece geçmişe bakar)
- Pre-normalization: LayerNorm bloklardan önce uygulanır
- Rotary Positional Embedding (RoPE): Sabit sin-cos yerine daha verimli pozisyonel gömme

# DiĞER ÖNEMLİ DİL MODELİ MİMARİLERİ

## XLNet: Permütasyon Bazlı Dil Modellemesi

- BERT'in "masking" yaklaşımının sınırlamaları (görmediği kelimeleri tahmin etmesi).
- Autoregressive + Bidirectional yapı birleşimi.
- Girdideki token sırasını farklı permütasyonlarla modelleyerek bağlamı daha esnek şekilde öğrenir.

## T5: Text-To-Text Transfer Transformer

- Full Transformer (Encoder + Decoder)
- Öğrenilebilir pozisyon embeddin
- Çok görevli eğitim (multitask learning)
- Transfer learning için evrensel yapı
- Benchmark lideri (SuperGLUE, SQuAD)

## Kod Odaklı Modeller

### Codex (OpenAI)

- GPT-3 türevi, yazılım üretimi üzerine eğitildi
- GitHub Copilot'un temel modelidir
- 12'den fazla programlama dilini destekler

### CodeBERT (Microsoft + HuggingFace)

- BERT mimarisi, çift dilli eğitim (code + natural language)
- Görevler: Code search, classification, clone detection

### StarCoder (BigCode Projesi)

- Open-source kod modeli, ~15B parametre
- Geniş bağlam penceresi (8k+ token)
- LLM tabanlı kod üretiminde açık kaynak alternatifi



# LLM'lerin Yetenekleri ve Sınırlamaları

## Yapabilecekleri

- **Metin Oluşturma:** Kullanıcının verdiği bir konuda makale, hikaye ya da şiir yazabilir.
- **Soru Yanıtlama:** Verilen bir bilgi kümesine dayanarak sorulara yanıt verebilir.
- **Özetleme:** Uzun metinleri daha kısa ve öz bir şekilde özetleyebilir.
- **Çeviri:** Bir dilden diğerine metin çevirisi yapabilir.
- **Çok aşamalı muhakeme (Multi-hop reasoning):** Birden fazla bilgi kaynağını ilişkilendirerek mantıksal çıkarımlarda bulunabilirler.

## Yapamayacakları

- **Gerçek Zamanlı Bilgiye Erişim:** Eğitim verileri sabit olduğundan, güncel olaylara dair bilgi sunamaz.
- **Duygusal Anlayış ve Empati:** Metin üzerinden duyguları anlama kapasitesi sınırlıdır.
- **Kapsamlı Bilgi Analizi:** Derinlemesine analizler yapma yeteneği yoktur, çünkü bağlam ve ön bilgi eksikliği bulunabilir.
- **Model yanlılıkları (Bias):** Eğitimdikleri veri setlerinde mevcut olan yanlılıkları öğrenirler ve bu, çıktılarına yansıyabilir.



# LLM'lerin Uygulama Alanları

**Eğitim:** Kişiselleştirilmiş öğrenme deneyimi: Öğrencinin seviyesine göre eğitim içeriklerini ayarlayabilir.

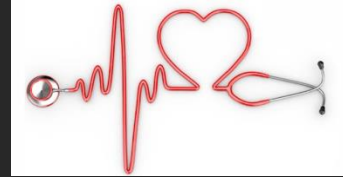
Otomatik test hazırlama ve değerlendirme: Sınavlar ve ödevler için içerik üretimi ve değerlendirme.



**Pazarlama:** İçerik oluşturma: Blog yazıları, sosyal medya paylaşımları, reklam metinleri gibi pazarlama içeriklerinin üretilmesi. Müşteri ilişkileri: Otomatik müşteri destek botları ve kişiselleştirilmiş müşteri iletişimleri.



**Sağlık:** Tıbbi danışmanlık: Hastaların sorularını yanıtlayabilir, genel sağlık önerileri verebilir. Tıbbi verilerin analizi: Tıbbi raporların özetlenmesi ve doktorlara karar destek sistemleri sağlanması.

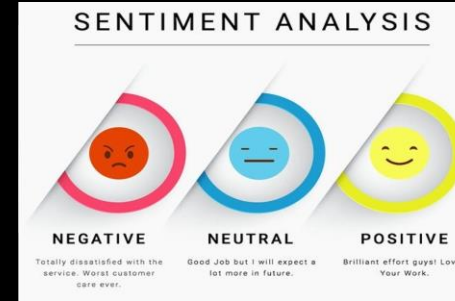


**Yaratıcılık:** Görsel ve müzik üretimi: Görsel ve ses verileri üzerinde çalışarak yeni sanat eserleri, müzikler veya tasarımlar oluşturabilir. Senaryo yazma: Sinema ve televizyon için yaratıcı hikaye ve senaryolar oluşturabilir.



# LLM'lerin Kullanım Alanları

Kullanım Alanı	Temel Model Ailesi	Anahtar Özellik
Metin Üretimi	GPT, Claude	Yaratıcı ve akıcı içerik
QA Sistemleri	BERT, T5, GPT + RAG	Doğruluğa dayalı yanıt üretimi
Kod Üretimi	Codex, StarCoder	Yazılım üretimi, hata düzeltme
Çeviri	T5, NLLB, GPT-4	Çok dilli çeviri
Duygu Analizi	BERT, RoBERTa	Sınıflandırma, polarite tespiti
Domain Adaptation	LLaMA + LoRA, BioBERT	Alan-özel öğrenme



# LLM'lerin Zorlukları ve Limitleri

## Ön Yargı (Bias)

Modellerin veri kaynaklarından gelen önyargıları içermesi ve modelin yanıtlarını olumsuz yönde etkilemesidir.



## Hafıza ve Hesaplama Gücü Gereksinimleri

Yüksek işlem gücü ve maliyet gerektirirler.



## Veri Gizliliği Sorunları

Hassas veri kullanımı durumunda gizlilik endişeleri oluşabilir. Kişisel verilerin korunması en öncelikli etik sorumluluklardan biridir.



## Genel yetenek Sınırlamaları

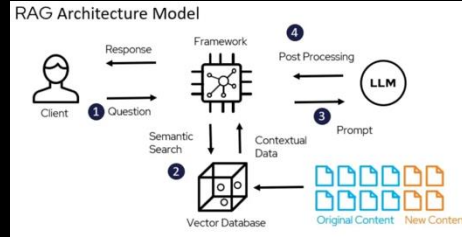
Özellikle karmaşık ya da uzun bağlamlarda anlama ve yanıtlar üretmede zorlukları yaşayabilirler.



# GELECEĞE YÖNELİK PERSPEKTİFLER

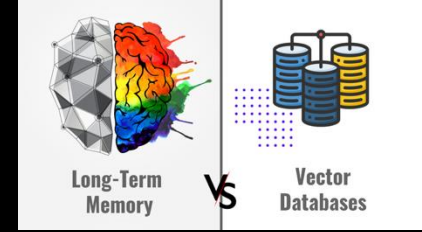
## 🧠 Retrieval-Augmented Generation (RAG)

- LLM, bilgiyi doğrudan parametrelerinden değil, dış bilgi kaynaklarından geri getirerek kullanır.
- Avantaj:** Hallucination azalır, gerçek bilgiyle yanıt üretilir.
- Kullanım:** Dokümana dayalı QA sistemleri, kurumsal arama



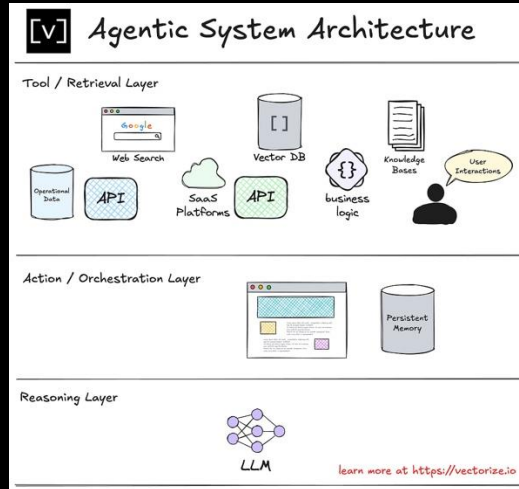
## 🧠 Uzun Süreli Bellek & Kişiselleştirme

- Kullanıcının geçmişyle bağlantılı hafıza sistemleri
- Vektör tabanlı kişisel veri hafızaları (e.g., memory embeddings)



## 🤖 LLM Tabanlı Ajanlar (Agentic AI)

- LLM'ler, planlama yapabilen, araç kullanabilen, döngüsel görevleri yürüten yapay ajanlara dönüşüyor.
- Örnek:** AutoGPT, LangGraph, OpenAI GPTs
- Yaklaşım:** Düşün → Planla → Eyleme geç → Geri bildir

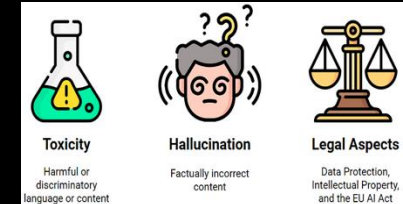


## 🔑 LLMOps ve Üretim Ortamları için Gelişmiş Altyapı

- Model yönetimi, versiyonlama, denetim sistemleri
- Değerlendirme metrikleri: Toxicity, factuality, latency, cost

## 🚀 Multimodal Modeller

- Tek Girdide Çoklu Modalite:** Metin + Görsel + Ses + Video + Yapısal veri
- Amaç:** İnsan algısına benzer çok yönlü anlama yeteneği





Q&A

???

# TEŞEKKÜRLER!



**Kubilay Tuna**

*Senior Data Scientist*

[kubilaytuna26@hotmail.com](mailto:kubilaytuna26@hotmail.com)