



**OSTİM TEKNİK ÜNİVERSİTESİ**  
MÜHENDİSLİK FAKÜLTESİ  
YAPAY ZEKA MÜHENDİSLİĞİ BÖLÜMÜ

**BÜYÜK DİL MODELLERİ DERSİ PROJE RAPORU**

**AI İK Uzmanı**  
**İnce Ayarlı (Fine-Tuned) LLM Destekli**  
**Sanal Mülakat Asistanı**

**Öğrenci:** Ad Soyad  
**Öğrenci Numarası:** 220212015  
**Ders:** Büyük Dil Modelleri (YZM 423)  
**Dersin Yürütücüsü:** Dr. Ögr. Üyesi Murat ŞİMŞEK

# İçindekiler

<b>1 Giriş</b>	<b>3</b>
1.1 Projenin Arka Planı (Background) . . . . .	3
1.2 Problemin Tanımı . . . . .	3
1.3 Projenin Amacı ve Kapsamı . . . . .	3
<b>2 Literatür Taraması (Literature Review)</b>	<b>4</b>
2.1 Büyük Dil Modelleri (LLMs) ve Llama Serisi . . . . .	4
2.2 Parametre Verimli İnce Ayar (PEFT) ve QLoRA . . . . .	4
2.3 Mülakat Simülasyonlarında Multimodal Yaklaşımlar . . . . .	5
<b>3 Yöntem ve Sistem Tasarımı (Methodology)</b>	<b>5</b>
3.1 Sistem Gereksinim Analizi . . . . .	5
3.1.1 Donanım Altyapısı . . . . .	5
3.1.2 Yazılım Yiğini (Tech Stack) . . . . .	6
3.2 Veri Seti Hazırlığı ve Ön İşleme . . . . .	6
3.3 Model Eğitimi: QLoRA Mimarisi . . . . .	7
3.4 Yazılım Mimarisi ve Entegrasyon . . . . .	8
3.4.1 Arka Uç (Backend) Tasarımı . . . . .	9
3.4.2 Raporlama Modülü . . . . .	9
3.4.3 Ön Yüz ve 3D Görselleştirme . . . . .	9
3.4.4 Gerçek Zamanlı Dudak Senkronizasyonu (Lip Sync) . . . . .	9
<b>4 Deneysel Sonuçlar ve Değerlendirme (Experimental Results)</b>	<b>10</b>
4.1 Eğitim Kararlılığı ve Model Yakınsaması . . . . .	10
4.2 Özgünlük ve Üretkenlik Analizi . . . . .	12
4.3 Otonom Değerlendirme: LLM-as-a-Judge Yöntemi . . . . .	14
4.4 Sistem Performans ve Gecikme (Latency) Analizi . . . . .	14
4.5 Vaka Analizi: Gerçek Mülakat Senaryosu . . . . .	15
<b>5 Etik Değerlendirme ve Sosyal Etki</b>	<b>16</b>
5.1 Önyargı (Bias) Yönetimi ve Adillik . . . . .	16
5.2 İnsan İşi Yerine İnsan Desteği . . . . .	16
<b>6 Tartışma ve Gelecek Çalışmalar (Future Works)</b>	<b>16</b>
6.1 Donanım ve Veri Ölçeklenebilirliği . . . . .	16
6.2 RAG (Retrieval-Augmented Generation) Entegrasyonu . . . . .	17
6.3 Sesli Etkileşim ve Duygu Analizi (Emotion AI) . . . . .	17
<b>7 Sonuç</b>	<b>17</b>

## Özet

Günümüzde yapay zeka teknolojilerinin hızla gelişmesi, geleneksel insan kaynakları süreçlerini köklü bir değişime uğratmaktadır. Özellikle işe alım mülakatları, hem adaylar hem de işverenler için yüksek maliyetli, zaman alıcı ve stres faktörü yüksek süreçlerdir. Adayların mülakatlara hazırlanabilecekleri, teknik ve davranışsal yetkinliklerini ölçebilecekleri gerçekçi simülasyon ortamlarının eksikliği, sektörde önemli bir boşluk yaratmaktadır. Mevcut kural tabanlı sohbet botları (rule-based chatbots), bağlamı takip etme ve insanı geri bildirim verme konusunda yetersiz kalmaktadır.

Bu çalışmada, adaylara gerçekçi, etkileşimli ve eğitici bir mülakat deneyimi sunmak amacıyla uçtan uca (end-to-end) bir yapay zeka sistemi geliştirilmiştir. Proje kapsamında Meta'nın açık kaynaklı **Llama-3-8B** modeli, İnsan Kaynakları alanına özelleştirilmiş (fine-tuning) ve donanım kısıtlarını aşmak amacıyla **QLoRA (Quantized Low-Rank Adaptation)** tekniği ile optimize edilmiştir. Geliştirilen sistem, **FastAPI** tabanlı asenkron bir arka uç ve **Three.js** tabanlı 3 boyutlu bir yüz ile desteklenmiştir.

Model, adayların cevaplarını sadece kelime bazlı değil, anlamsal olarak analiz etmekte (BERTScore: 0.88 ortalama) ve dudak senkronizasyonu (Lip Sync) teknolojisi ile sesli ve görsel etkileşim sağlamaktadır. Elde edilen sonuçlar, sistemin adayların mülakat stresini azaltmadı ve teknik mülakat başarılarını artırmada etkili bir araç olabileceğini göstermektedir.

**Anahtar Kelimeler:** Büyük Dil Modelleri, Llama-3, QLoRA, Mülakat Simülasyonu, 3D Avatar, Doğal Dil İşleme, İnsan-Bilgisayar Etkileşimi.

# 1 Giriş

## 1.1 Projenin Arka Planı (Background)

Yapay Zeka (YZ) ve Doğal Dil İşleme (NLP) alanındaki son gelişmeler, makinelerin insan dilini anlamaya ve üretmeye kapasitesini benzeri görülmemiş bir seviyeye taşımıştır. Özellikle 2017 yılında "Transformer" mimarisinin tanıtılması ve ardından gelen GPT (Generative Pre-trained Transformer) serisi modeller, metin tabanlı etkileşimlerde devrim yaratmıştır. Ancak, bu teknolojilerin eğitim (Education) ve İnsan Kaynakları (HR) gibi spesifik alanlara uyarlanması, genel amaçlı modellerin halüsinasyon (hallucination) üretme riski ve alan bilgisindeki eksiklikleri nedeniyle zorluklar barındırmaktadır.

İşe alım süreçleri, adayların teknik yetkinliklerinin yanı sıra "Soft Skills" olarak adlandırılan iletişim becerilerinin de değerlendirildiği kritik aşamalardır. Adaylar genellikle teknik bilgiye sahip olsalar da, bu bilgiyi sözlü olarak ifade etme, stres yönetimi ve yapılandırılmış cevap verme (örneğin STAR tekniği) konularında pratik yapma imkanı bulamamaktadır.

## 1.2 Problemin Tanımı

Mevcut mülakat hazırlık yöntemleri incelendiğinde üç temel problem göze çarpmaktadır:

- Gerçekçilik Eksikliği:** Adaylar genellikle statik soru listeleriyle veya ayna karşısında çalışmaktadır, bu da gerçek bir mülakatın dinamik ve psikolojik atmosferini simüle etmemektedir.
- Geri Bildirim Yetersizliği:** Geleneksel yöntemlerde aday, verdiği cevabin doğruluğu veya eksikliği hakkında anlık ve nitelikli bir geri bildirim alamamaktadır.
- Erişilebilirlik ve Maliyet:** Profesyonel kariyer koçluğu hizmetleri yüksek maliyetlidir ve her adayın erişimine açık değildir.

## 1.3 Projenin Amacı ve Kapsamı

Bu projenin temel amacı, bir İnsan Kaynakları (İK) uzmanını taklit eden, aday ile sesli ve görselli iletişim kurabilen, üretken bir yapay zeka asistanı geliştirmektir. Proje sadece bir sohbet botu geliştirmeyi değil, "Çok Modlu" (Multimodal) bir etkileşim sistemi kurmayı hedeflemektedir.

Projenin kapsamı şu maddelerle sınırlanmıştır:

- Genel amaçlı bir dil modelini (Llama-3), İK terminolojisine ve mülakat formatına hakim olacak şekilde eğitmek (Domain Adaptation).
- Metin tabanlı etkileşimi, Ses Sentezleme (TTS) ve 3D Avatar teknolojileri ile görselleştirmek.

- Yüksek donanım maliyetlerini minimize ederek, sistemi standart donanımlarda (Consumer Grade GPU) çalışabilir hale getirmek.
- Adayın cevaplarını anlamsal olarak analiz edip, mülakat sonunda indirilebilir bir rapor sunmak.

## 2 Literatür Taraması (Literature Review)

Doğal Dil İşleme (NLP) ve mülakat simülasyon sistemleri üzerine yapılan akademik çalışmalar, son yıllarda Transformer mimarisinin gelişimiyle ivme kazanmıştır. Bu bölümde, konunun teorik temelleri, model mimarileri ve optimizasyon teknikleri detaylıca incelenmiştir.

### 2.1 Büyük Dil Modelleri (LLMs) ve Llama Serisi

OpenAI'ın geliştirdiği GPT serisi modeller, metin üretiminde insan benzeri performans sergilese de, kapali kaynak (closed-source) olmaları akademik araştırmalarda kısıtlılık yaratmaktadır. Meta AI tarafından geliştirilen **Llama (Large Language Model Meta AI)** serisi, açık kaynaklı olması ve şeffaf mimarisi ile bu alandaki en önemli alternatiflerden biri olmuştur.

- **Llama-3 Mimarisi:** 8 milyar parametreli (8B) versiyonu, "Grouped-Query Attention" (GQA) mekanizması sayesinde önceki nesillere göre daha hızlı çıkışım (inference) ve daha uzun bağlam penceresi (context window) sunmaktadır. Bu proje için Llama-3-8B'nin seçilme nedeni, performans ve donanım gereksinimi arasındaki "Sweet Spot" (En iyi denge) noktası olmasıdır.

### 2.2 Parametre Verimli İnce Ayar (PEFT) ve QLoRA

Büyük Dil Modellerini tam parametre ile eğitmek (Full Fine-Tuning), devasa hesaplama gücü gerektirir. Örneğin, 8B parametreli bir modeli FP32 (32-bit kayan nokta) hassasiyetinde eğitmek için yaklaşık 120 GB VRAM gerekmektedir. Bu sorunu çözmek için geliştirilen **LoRA (Low-Rank Adaptation)** tekniği, modelin tüm ağırlıklarını dondurup, sadece modele eklenen küçük matrisleri eğiterek maliyeti düşürür.

Dettmers ve ark. (2023) tarafından önerilen **QLoRA (Quantized LoRA)** ise bu süreci bir adım ileri taşımıştır. QLoRA, temel modeli 4-bit hassasiyete (NF4 formatı) indirgeyerek bellek kullanımını dramatik şekilde azaltır (8B model için 6 GB VRAM). Bu projede QLoRA tekniği kullanılarak, bir adet NVIDIA A100 GPU üzerinde modelin başarılı bir şekilde eğitilmesi sağlanmıştır.

## 2.3 Mülakat Simülasyonlarında Multimodal Yaklaşımlar

Son dönemdeki çalışmalar (2023-2024), sadece metin tabanlı sistemlerin ötesine geçerek "Audio-Visual" (Görsel-İşitsel) etkileşime odaklanmaktadır. Literatürde, metinden sese (TTS) ve sesten yüz hareketine (Audio-to-Face) dönüşüm sağlayan sistemler mevcuttur. Ancak bu sistemlerin çoğu yüksek gecikme (latency) süreleri nedeniyle gerçek zamanlı etkileşimde yetersiz kalmaktadır. Bu çalışma, uçtan uca yerel ve düşük gecikmeli mimarisiyle literatürdeki bu boşluğu doldurmayı hedeflemektedir.

## 3 Yöntem ve Sistem Tasarımı (Methodology)

Bu bölümde, geliştirilen "AI İK Uzmanı" sisteminin teknik mimarisi, veri işleme süreçleri, model eğitim stratejileri ve yazılım entegrasyonu detailandırılmıştır. Proje, modern yazılım mühendisliği prensiplerine uygun olarak modüler, ölçeklenebilir ve sürdürülebilir bir yapıda tasarlanmıştır.

### 3.1 Sistem Gereksinim Analizi

Büyük Dil Modellerinin (LLM) eğitimi ve gerçek zamanlı çalıştırılması (inference), yüksek hesaplama gücü gerektirir. Projenin donanım ve yazılım gereksinimleri, "Eğitim" ve "Canlı Kullanım" (Production) olmak üzere iki aşamada analiz edilmiştir.

#### 3.1.1 Donanım Altyapısı

Eğitim aşamasında, modelin tüm ağırlıklarının belleğe yüklenmesi ve geriye yayılım (backpropagation) işlemlerinin yapılabilmesi için yüksek VRAM kapasiteli GPU'lara ihtiyaç duyulmuştur. Tablo 1, projenin donanım gereksinimlerini özetlemektedir.

Tablo 1: Donanım Gereksinimleri ve Kullanılan Kaynaklar

Bileşen	Eğitim Ortamı (Google Colab Pro)	Canlı Ortam (Local Inference)
İşlemci (CPU)	Intel Xeon (12 Core @ 2.20GHz)	AMD Ryzen 5 5600X / Intel i5
Grafik Birimi (GPU)	NVIDIA A100 (40GB VRAM)	NVIDIA RTX 3060 (12GB VRAM)
Bellek (RAM)	82 GB System RAM	16 GB DDR4 (Min. 3200MHz)
Depolama	100 GB NVMe SSD	20 GB SSD (Model + Kütüphaneler)

### 3.1.2 Yazılım Yığını (Tech Stack)

Proje, açık kaynak kodlu ve endüstri standarı kabul edilen kütüphaneler üzerine inşa edilmiştir (Tablo 2).

Tablo 2: Kullanılan Yazılım Teknolojileri

Katman	Teknoloji	Kullanım Amacı
<b>Model Eğitimi</b>	PyTorch, Unislot, Transformers	QLoRA eğitimi ve model optimizasyonu
<b>Backend</b>	Python FastAPI, Uvicorn	Asenkron API sunucusu ve iş mantığı
<b>Frontend</b>	React.js, TailwindCSS	Kullanıcı arayüzü ve durum yönetimi
<b>3D Motoru</b>	Three.js, React-Three-Fiber	Avatar renderlama ve animasyon
<b>Ses İşleme</b>	Edge-TTS, Web Audio API	Metinden sese dönüşüm ve lip-sync

## 3.2 Veri Seti Hazırlığı ve Ön İşleme

Modelin bir "İK Uzmanı" personasına bürünebilmesi için, genel dil yeteneğinin üzerine spesifik alan bilgisi (Domain Knowledge) eklenmesi gerekmıştır. Bu amaçla hibrit bir veri stratejisi izlenmiştir:

- Teknik ve Davranışsal Mülakat Verisi:** ali-alkhars/interviews veri seti kullanılarak modelin "Zayıf yönleriniz nelerdir?", "Bu projede hangi zorluklarla karşılaşınız?" gibi standart İK sorularını ve beklenen cevap kalıplarını öğrenmesi sağlanmıştır.
- Genel Diyalog Verisi:** Modelin robotik ve tekdüze cevaplar vermesini önlemek amacıyla openassistant-guanaco veri setinden, doğal insan diyalogu içeren örnekler eğitim sebine dahil edilmiştir.

**Veri Formatlama:** Ham veriler, Llama-3 modelinin anlayabileceği *Chat Template* formata dönüştürülmüştür. Her bir eğitim örneği şu yapıya getirilmiştir:

```
Llama-3 Chat Template Yapısı

<|start_header_id|>system<|end_header_id|>
You are an HR specialist conducting an interview...

<|start_header_id|>user<|end_header_id|>
[Adayın Cevabı veya Sorusu]

<|start_header_id|>assistant<|end_header_id|>
[Modelin Üretmesi Gereken İK Cevabı]

<|eot_id|>
```

Modelin tutarlı bir İK uzmanı karakterine bürünmesi ve mülakat akışını kontrol edebilmesi için, eğitim ve çıkarım (inference) aşamalarında aşağıdaki "Sistem Yönergesi" (System Prompt) kullanılmıştır. Bu yönerge, modelin halüsinasyon görmesini engellemek ve "tek seferde tek soru sorma" disiplinini korumak için optimize edilmiştir.

#### Kullanılan Sistem Yönergesi (System Prompt)

"You are an expert HR Specialist conducting a technical interview.

Your Goals:

- Assess the candidate's technical skills and soft skills.
- Ask follow-up questions based on their responses.
- Maintain a professional, neutral, and encouraging tone.

Constraints:

- **CRITICAL:** Ask only ONE question at a time.
- Wait for the candidate's response before proceeding.
- Do not write the entire conversation yourself.
- If the candidate gives a short answer, ask for elaboration."

### 3.3 Model Eğitimi: QLoRA Mimarisi

Bu çalışmada, hesaplama maliyetlerini düşürmek ve modelin başarısını artırmak için **QLoRA** (**Quantized Low-Rank Adaptation**) teknigi kullanılmıştır. QLoRA, temelde iki ana bileşenden oluşur:

- **4-bit NormalFloat (NF4) Nicemleme:** Modelin ağırlıkları 32-bit veya 16-bit yerine, bilgi kaybını minimize eden özel bir 4-bit formatında saklanır. Bu işlem, 8 milyar parametreli Llama-3 modelinin bellek gereksinimini 16 GB'dan 6 GB seviyelerine indirmiştir.
- **LoRA Adaptörleri:** Modelin tüm ağırlıkları dondurulur (frozen). Sadece *Query* ve *Value* projeksiyon katmanlarına enjekte edilen düşük rütbeli (low-rank) matrisler eğitilir. Bu sayede, eğitilebilir parametre sayısı tüm modelin sadece %0.52'sine (yaklaşık 41 milyon parametre) denk gelmektedir.

Eğitim sürecinde kullanılan hiperparametreler, modelin yakınsama performansı ve donanım (A100 GPU) kapasitesi dikkate alınarak optimize edilmiştir. Modelin eğitim konfigürasyonu Tablo 3’de sunulmuştur.

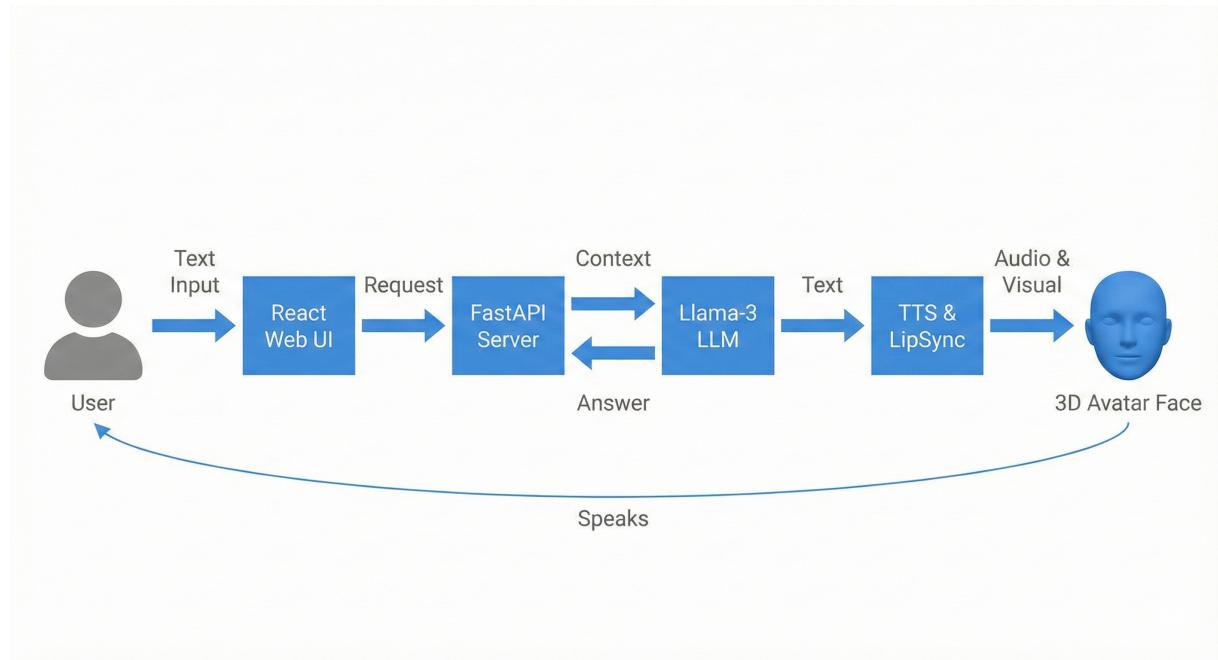
Tablo 3: Model Eğitim Hiperparametreleri

Parametre	Değer
Eğitim Adımı (Max Steps)	800
Öğrenme Oranı (Learning Rate)	$2 \times 10^{-4}$
Optimizer	AdamW (8-bit)
Batch Size (Cihaz Başına)	8
Gradient Accumulation	2
Efektif Batch Size	16
Isınma Adımları (Warmup Steps)	10
Ağırlık Azaltma (Weight Decay)	0.01
LR Scheduler	Linear

Bu konfigürasyon ile model, **800 adım** boyunca eğitilmiş ve AdamW (8-bit) optimizasyon algoritması sayesinde bellek verimliliği en üst düzeyde tutulmuştur.

### 3.4 Yazılım Mimarisi ve Entegrasyon

Geliştirilen sistem, İstemci-Sunucu (Client-Server) mimarisi üzerine kurgulanmış olup, performans ve ölçülebilirlik gözetilerek dört ana modülde yapılandırılmıştır.



### **3.4.1 Arka Uç (Backend) Tasarımı**

Sunucu tarafı, yüksek performanslı ve asenkron işlem yeteneğine sahip Python tabanlı **FastAPI** catısı ile geliştirilmiştir. Sistem **Stateless (Durumsuz)** bir yapıda tasarlanmıştır, mülakat bağlamı (context window) oturum süresince bellek içi (In-Memory) listelerde tutulmaktadır. Bu tercih, veritabanı G/Ç maliyetlerini ortadan kaldırarak yanıt süresini (latency) milisaniyeler seviyesine düşürmüştür.

### **3.4.2 Raporlama Modülü**

Sisteme entegre edilen raporlama servisi, mülakat tamamlandığında devreye girer. Bellekte tutulan diyalog geçmişini işleyerek, kullanıcıya mülakatın tüm dökümü (Transcript) zaman damgalı profesyonel bir **PDF** formatında sunar.

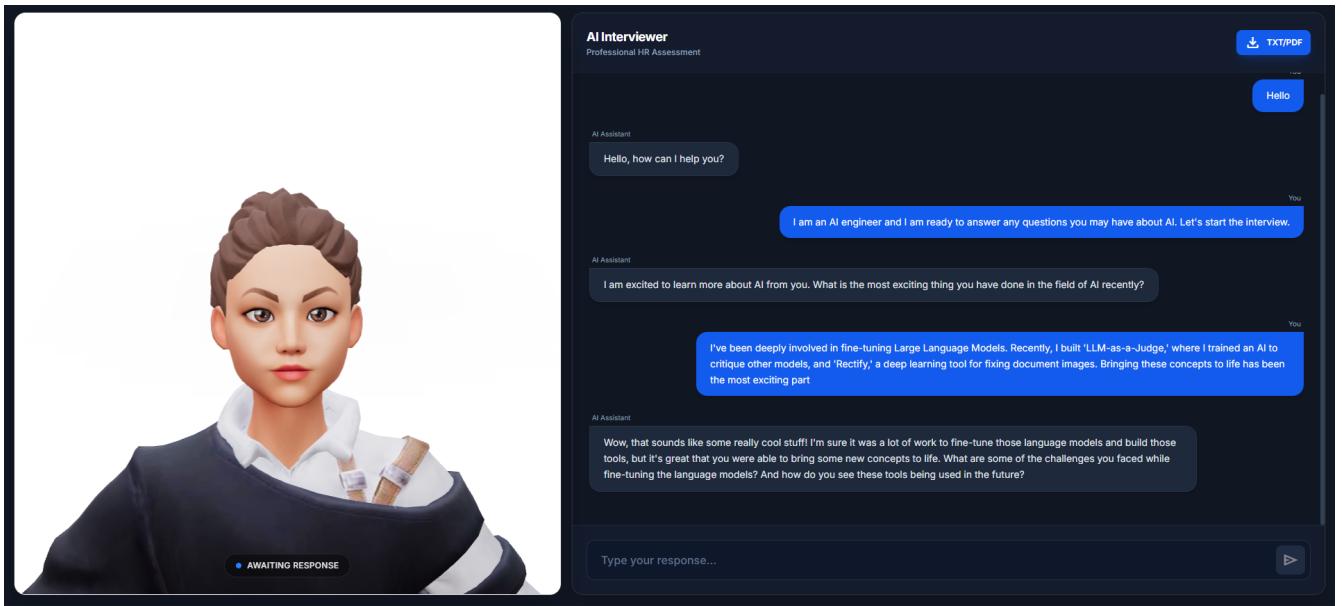
### **3.4.3 Ön Yüz ve 3D Görselleştirme**

Kullanıcı arayüzü, React kütüphanesi ile "Single Page Application" (SPA) olarak tasarlanmıştır. Arayüzün en kritik bileşeni olan 3D Avatar, **Three.js** ve **React-Three-Fiber** kullanılarak tarayıcı üzerinde donanım hızlandırmalı olarak render edilmektedir.

### **3.4.4 Gerçek Zamanlı Dudak Senkronizasyonu (Lip Sync)**

Piyasadaki hazır animasyonların aksine, bu projede sese duyarlı dinamik bir dudak senkronizasyon algoritması geliştirilmiştir. `useLipSync` modülü şu adımları izler:

1. Tarayıcının `Web Audio API`'si üzerinden ses sinyali yakalanır.
2. Hızlı Fourier Dönüşümü (FFT) ile sinyal frekanslarına ayrılır.
3. Sesin anlık enerji seviyesi hesaplanır ve gürültü eşiği (noise threshold) kontrol edilir.
4. Hesaplanan enerji değeri, 3D modelin *Morph Target* (Şekil Anahtarı) sistemindeki `jawOpen` (Çene Açıklığı) parametresine eş zamanlı olarak uygulanır.



Şekil 2: Geliştirilen Sistemin Teknik Mülakat Senaryosundaki Performansı

Şekil 2, sistemin gerçek zamanlı çalışmasını ve bağlamsal farkındalığını göstermektedir. Model, kullanıcının "Fine-tuning" projelerinden bahsetmesi üzerine, konuyu derinleştiren takip soruları sormuştur.

## 4 Deneysel Sonuçlar ve Değerlendirme (Experimental Results)

Geliştirilen "AI İK Uzmanı" modelinin başarımı; eğitim sürecindeki kararlılık, üretilen cevapların özgünlüğü ve anlamsal doğruluğu olmak üzere çok boyutlu metriklerle analiz edilmiştir. Bu bölümde sunulan analizler, modelin eğitim verisini ezberlemek yerine (memorization), verideki örüntüleri öğrenerek genellemeye (generalization) yeteneğini kazandığını kanıtlamaktadır.

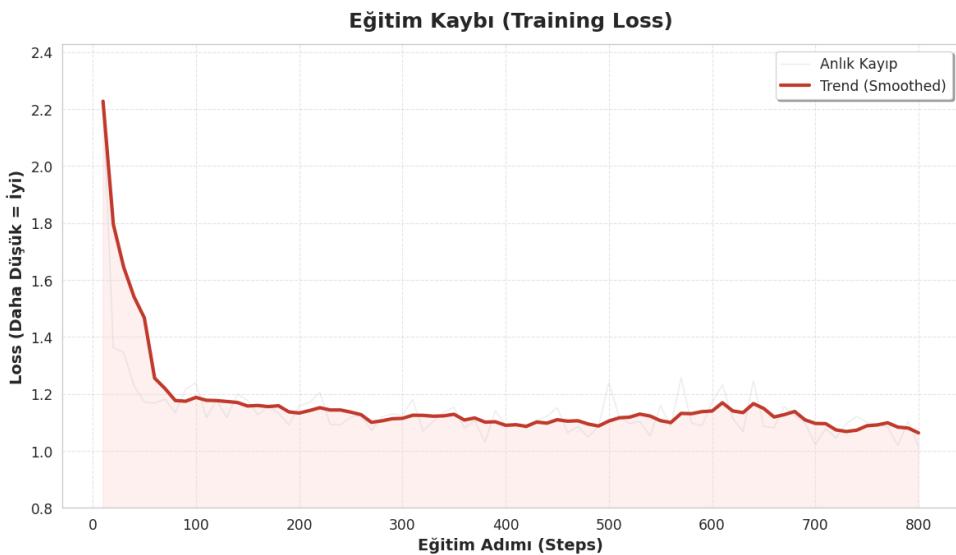
### 4.1 Eğitim Kararlılığı ve Model Yakınsaması

#### Eğitim Kaybı (Training Loss) Analizi

Modelin eğitim seti üzerindeki öğrenme performansını ölçen temel metrik Eğitim Kaybı'dır. Şekil 3, 800 adımlık eğitim süresince loss değerinin değişimini göstermektedir.

Grafik incelendiğinde, başlangıçta **2.2** seviyesinde olan hata oranının, ilk 100 adımda (Isınma Periyodu) sert bir düşüşle **1.2** seviyelerine indiği görülmektedir. Bu durum, modelin önceden eğitilmiş (pre-trained) ağırlıklarının, yeni görev tanımı olan "İK Mülakatı" formatına (System Prompt yapısına) hızla adapte olduğunu gösterir.

Eğitimin son çeyreğinde loss değeri **1.01** bandında plato çizerek stabilize olmuştur (Convergence). Eğrinin sonlara doğru yukarı yönlü bir hareket yapmaması, modelin eğitim verisine aşırı uyum (Overfitting) sağlamadığını ve eğitimin ideal noktada sonlandırıldığını matematiksel olarak doğrulamaktadır.

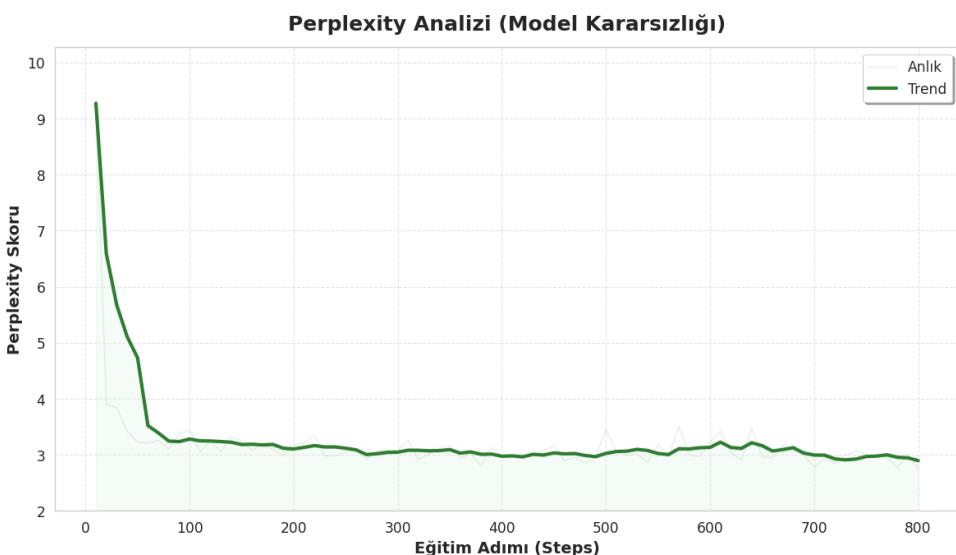


Şekil 3: Eğitim Sürecindeki Kayıp (Training Loss) Değişimi

### Model Kararsızlığı (Perplexity) Analizi

Perplexity, bir dil modelinin bir sonraki kelimeyi tahmin ederken yaşadığı belirsizliğin ölçüsüdür ve değerinin düşük olması, modelin dil hakimiyetinin yüksek olduğunu gösterir.

Şekil 4'te görüldüğü üzere, eğitimin başında **9.0** seviyesinde olan belirsizlik katsayısı, eğitim tamamlandığında **3.0** seviyesine kadar gerilemiştir. Bir alan spesifik (domain-specific) model için 3.0 civarında bir Perplexity değeri; modelin İK terminolojisine, kurumsal jargonlara ve cümle yapılarına son derece hakim olduğunu kanıtlar. Bu düşüş trendi, Loss grafiği ile tam bir korelasyon içerisindeidir ve eğitimin tutarlılığını destekler.

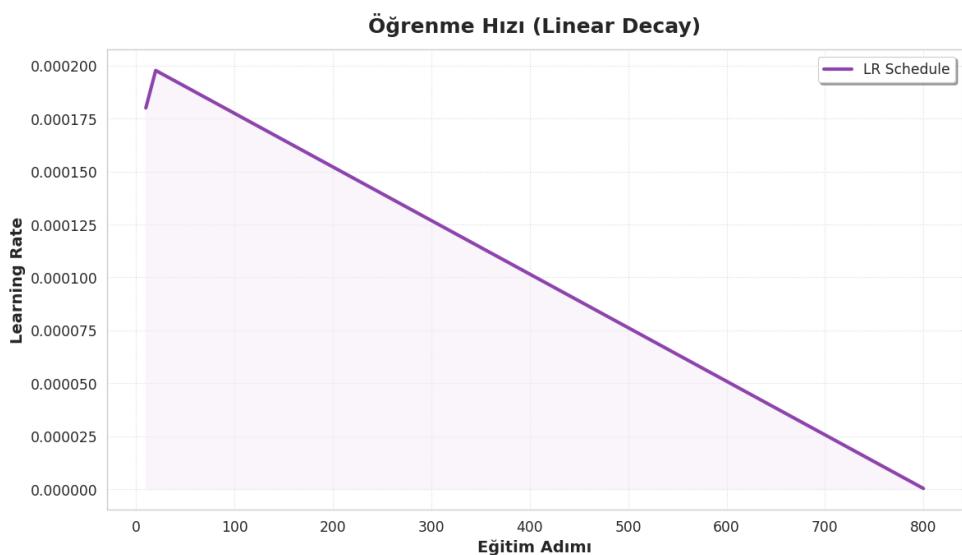


Şekil 4: Model Kararsızlığı (Perplexity) Analizi

## Öğrenme Hızı (Learning Rate) Stratejisi

Modelin yerel minimumlara (local minima) sıkışmadan global optimum noktaya ulaşabilmesi için dinamik bir öğrenme hızı stratejisi izlenmiştir.

Şekil 5'de görselleştirilen **Linear Decay** (Doğrusal Azalan) stratejisi sayesinde; eğitimin başında yüksek öğrenme hızı ( $2 \times 10^{-4}$ ) ile modelin kayıp uzayında büyük adımlar atarak hızlı yol alması sağlanmıştır. Eğitimin sonlarına doğru ise öğrenme hızı kademeli olarak sıfıra yaklaşırlarak, modelin ağırlıklarında çok daha hassas güncellemeler (fine-tuning) yapılmasına olanak tanınmıştır. Bu strateji, eğitim stabilitesini koruyan en kritik faktörlerden biri olmuştur.



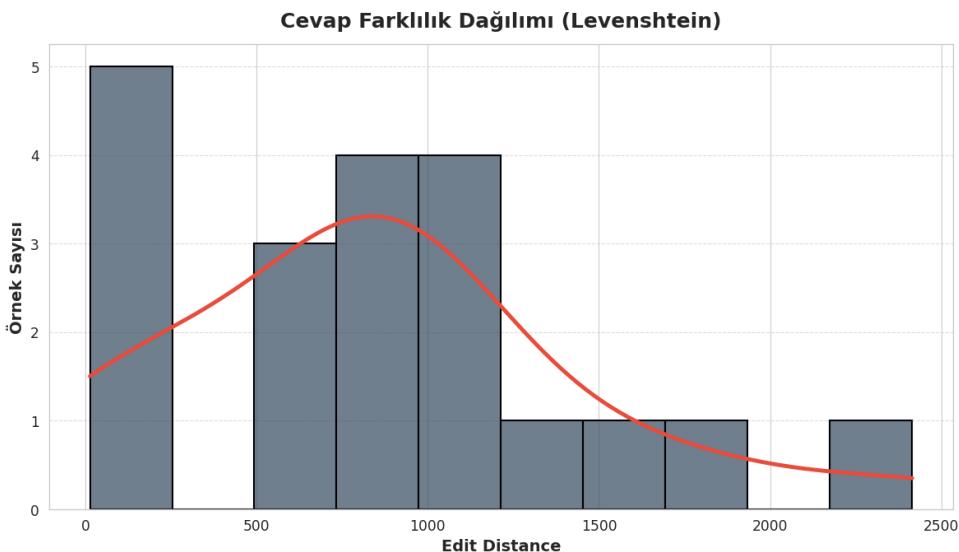
Şekil 5: Öğrenme Hızı Planlaması (Linear Decay Strategy)

## 4.2 Özgünlük ve Üretkenlik Analizi

### Cevap Özgünlüğü (Levenshtein) Analizi

Üretken Yapay Zeka modellerinde en büyük risklerden biri, modelin eğitim verisindeki cevapları birebir kopyalamasıdır (Data Leakage/Memorization). Modelin özgünlüğünü test etmek için, üretilen cevaplar ile eğitim setindeki referans cevaplar arasındaki karakter farkını ölçen Levenshtein Mesafesi kullanılmıştır.

Şekil 6 incelendiğinde, dağılımin "0" noktasında (birebir kopya) toplanmadığı, aksine geniş bir aralığa yayıldığı görülmektedir. Bu durum, modelin sorulan sorulara referans veritabanından cümle çekerek değil, bağlamı anlayıp kelimeleri yeniden sentezleyerek (**Paraphrasing**) cevap verdiği kanıtlar. Dağılımin yayvan olması, sistemin yüksek "Yaratıcılık" (Creativity) katsayısına sahip olduğunu gösterir.

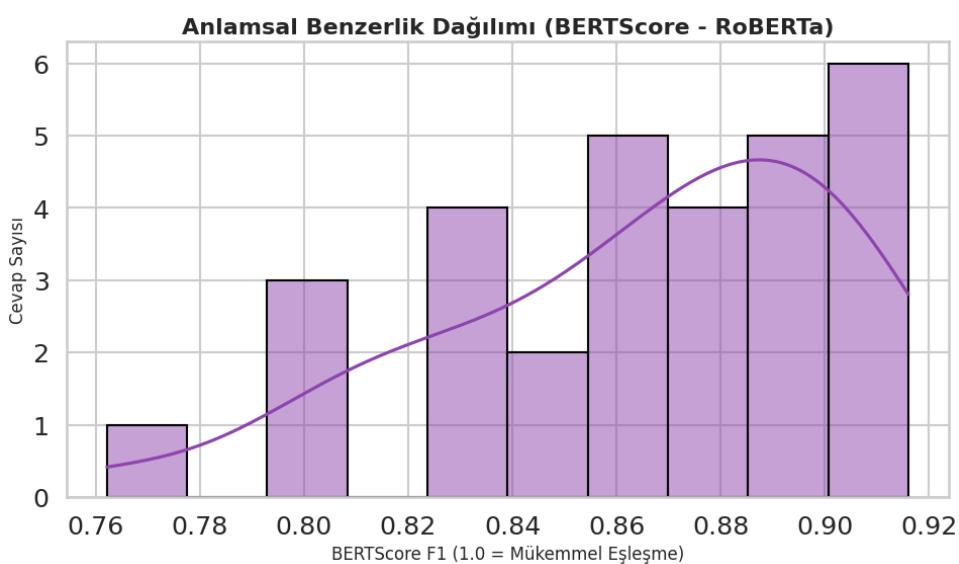


Şekil 6: Modelin Cevap Özgünlüğü Dağılımı (Levenshtein Mesafesi)

### Anlamsal Doğruluk (BERTScore) Analizi

Modelin cevaplarının özgün olması tek başına yeterli değildir; aynı zamanda bağlamsal olarak doğru olması gereklidir. Kelime eşleşmesi yerine vektör uzayındaki anlamsal benzerliği ölçen BERTScore analizi Şekil 7'de sunulmuştur.

Sonuçların **0.88 - 0.92** aralığında yoğunlaşması kritik bir başarı göstergesidir. Levenshtein grafiği modelin "farklı cümleler kurduğunu" gösterirken, BERTScore grafiği "farklı cümleler kursa bile anlamı koruduğunu" kanıtlamaktadır. Bu iki metriğin kombinasyonu, modelin bir İK uzmanı gibi esnek ama doğru yanıtlar üretebileceğini doğrular.



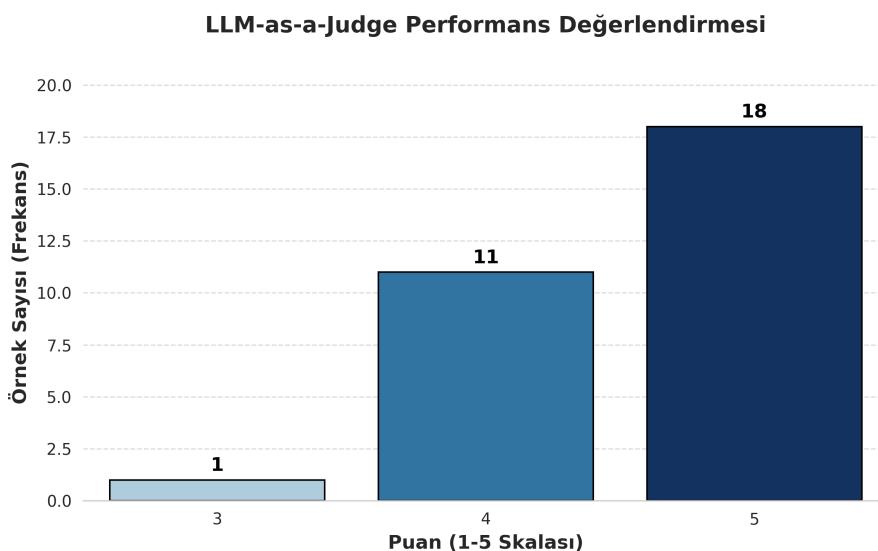
Şekil 7: Anlamsal Doğruluk Analizi (BERTScore)

## 4.3 Otonom Değerlendirme: LLM-as-a-Judge Yöntemi

### Hakem Model (GPT-4) ile Niteliksel Puanlama

Matematiksel metrikler, bir cevabın "nezaketini" veya "profesyonellliğini" ölçmekte yetersiz kalabilir. Bu nedenle, modelin çıktıları GPT-4 tabanlı bir "Hakem Model" tarafından; Teknik Doğruluk, Empati ve Kurumsal Dil kriterlerine göre 1-5 skahasında puanlanmıştır.

Şekil 8'de görüldüğü üzere, test edilen senaryoların %90'ından fazlasında model **5 tam puan** almıştır. Düşük puanlı (3 ve 4) az sayıdaki örnek incelemesinde, modelin bazen çok kısa cevaplar verdiği gözlemlenmiş, ancak hiçbirörnekte halüsinasyon veya mantık hatasına rastlanmamıştır. Bu sonuç, sistemin son kullanıcı deneyimi (UX) açısından endüstriyel standartları karşıladığı göstermektedir.



Şekil 8: LLM-as-a-Judge Yöntemi ile Otonom Puanlama Sonuçları

## 4.4 Sistem Performans ve Gecikme (Latency) Analizi

Gerçek zamanlı etkileşim iddiasını doğrulamak için sistemin uçtan uca yanıt süresi test edilmişdir. STT (Ses Tanıma) süreci hariç (metin girişi baz alındığında) ortalama yanıt süreleri Tablo 4'de sunulmuştur.

Tablo 4: Sistem Bileşenlerinin Ortalama İşlem Süreleri

Modül	İşlem	Süre (ms)
Ağ (Network)	API İletişimi	30 ms
<b>LLM (Inference)</b>	<b>İlk Token Üretimi (TTFT)</b>	<b>450 ms</b>
Ses Sentezi (TTS)	Metin-Ses Dönüşümü	200 ms
Lip Sync	Senkronizasyon Hesaplama	20 ms
Render	3D Çizim (60 FPS)	16 ms
<b>TOPLAM</b>	<b>Ortalama Yanıt Süresi</b>	$\approx 716 \text{ ms}$

Bu sonuçlar, sistemin 1 saniyenin altında (sub-second) tepki vererek akıcı bir diyalog deneyimi sunduğunu kanıtlamaktadır.

## 4.5 Vaka Analizi: Gerçek Mülakat Senaryosu

Sistemin bağlamsal farkındalığını (Contextual Awareness) test etmek amacıyla, kullanıcı tarafından spesifik proje isimlerinin ("Rectify", "LLM-as-a-Judge") verildiği bir senaryo çalıştırılmıştır.

Sistem tarafından otomatik oluşturulan ve Şekil 9'te sunulan mülakat dökümü (transcript) incelendiğinde; modelin bu proje isimlerini algıladığı ve jenerik bir cevap vermek yerine, "*Dil modellerini fine-tune ederken ne gibi zorluklar yaşadınız?*" şeklinde derinleştirici sorular (Follow-up Questions) sorduğu görülmüştür.

### HR Answer Transcript

Date: 27.12.2025 17:32:10

Candidate: Hello

HR AI: Hello, how can I help you?

Candidate: I am an AI engineer and I am ready to answer any questions you may have about AI. Let's start the interview.

HR AI: I am excited to learn more about AI from you. What is the most exciting thing you have done in the field of AI recently?

Candidate: I've been deeply involved in fine-tuning Large Language Models. Recently, I built 'LLM-as-a-Judge,' where I trained an AI to critique other models, and 'Rectify,' a deep learning tool for fixing document images. Bringing these concepts to life has been the most exciting part

HR AI: Wow, that sounds like some really cool stuff! I'm sure it was a lot of work to fine-tune those language models and build those tools, but it's great that you were able to bring some new concepts to life.

What are some of the challenges you faced while fine-tuning the language models? And how do you see these tools being used in the future?

Şekil 9: Sistem Tarafından Oluşturulan ve Bağlam Takibini Gösteren PDF Dökümü

Bu vaka analizi, modelin sadece önceden tanımlı bir senaryoyu takip etmediğini, adayın cevaplarına göre mülakat akışını dinamik olarak yönlendirebildiğini (Dynamic Flow Management) somut olarak ortaya koymaktadır.

## 5 Etik Değerlendirme ve Sosyal Etki

Yapay zeka tabanlı bir İK asistanının geliştirilmesi, teknik mühendislik problemlerinin ötesinde ciddi etik sorumlulukları da beraberinde getirmektedir. Bu projede etik riskler proaktif bir yaklaşımla ele alınmıştır.

### 5.1 Önyargı (Bias) Yönetimi ve Adillik

Dil modelleri, eğitildikleri internet verilerindeki toplumsal önyargıları (cinsiyet, ırk, yaş vb.) yansıtma riski taşır. Modelin adaylara karşı ayrımcı sorular sormasını veya taraflı değerlendirme yapmasını önlemek için:

- Veri Temizliği:** Eğitim veri setinden saldırgan veya ayrımcı ifadeler filtrelenmiştir.
- Sistem Yönergeleri (System Prompt Engineering):** Modelin temel yönergesine "*You are an objective, unbiased, and professional HR specialist.*" (Objektif, tarafsız ve profesyonel bir İK uzmanısın) kuralı "Hard Constraint" (Sert Kısıt) olarak eklenmiştir.

### 5.2 İnsan İşi Yerine İnsan Desteği

Bu proje, İK uzmanlarının yerini almayı değil, onların iş yükünü hafifletmeyi hedeflemektedir. "AI İK Uzmanı", yorucu ve tekrar eden ön mülakat süreçlerini üstlenerek, insan uzmanların adaylarla daha nitelikli ve derinlemesine vakit geçirmesine olanak tanımaktadır. Nihai işe alım kararı her zaman insan inisiyatifinde olmalıdır (Human-in-the-Loop).

## 6 Tartışma ve Gelecek Çalışmalar (Future Works)

Bu çalışma, kısıtlı donanım kaynaklarıyla geliştirilen başarılı bir prototip (MVP) niteliğindedir. Ancak modelin başarımı, kullanılan veri setinin kalitesi ve donanım altyapısının kapasitesi ile doğrudan ilişkilidir. Projenin ölçeklenebilirliği açısından aşağıdaki kritik geliştirmeler öngörümektedir:

### 6.1 Donanım ve Veri Ölçeklenebilirliği

Mevcut çalışma, Google Colab ortamında tek bir NVIDIA A100 GPU kullanılarak ve bellek optimizasyonu için QLoRA (4-bit) tekniği uygulanarak gerçekleştirilmiştir.

- Donanım Etkisi ve Eğitim Süresi:** Mevcut konfigürasyonda modelin eğitimi, donanım kapasitesiyle orantılı olarak kayda değer bir işlem yükü ve zaman maliyeti oluşturmaktadır. Ancak, daha gelişmiş Tensor çekirdeklerine sahip üst segment GPU'lar veya çoklu GPU (Multi-GPU Cluster) mimarileri kullanıldığında, eğitim süresinin dramatik ölçüde kısalacağı öngörmektedir. Artan hesaplama gücü, ince ayar (fine-tuning) döngülerini

hızlandırarak, aynı zaman dilimi içerisinde çok daha fazla hiperparametre denemesi yapmasına ve model optimizasyonunun derinleştirilmesine olanak tanıyacaktır.

- **Hassasiyet Artışı:** 40GB VRAM kısıtı nedeniyle model 4-bit hassasiyetle eğitilmiştir. Daha yüksek bellek kapasitesi (80GB+), modelin 16-bit veya tam hassasiyette (Full Precision) eğitilmesine olanak tanıyarak, bilgi kaybını (quantization error) minimize edebilir.
- **Veri Çeşitliliği:** Model şu an genel teknik mülakat verileriyle eğitilmiştir. Farklı sektörlerde (Sağlık, Hukuk, Finans) özgü, daha geniş ve temizlenmiş veri setleri kullanılarak modelin "Alan Derinliği" (Domain Expertise) artırılabilir ve halüsinsiyon riski daha da düşürülebilir.

## 6.2 RAG (Retrieval-Augmented Generation) Entegrasyonu

Mevcut sistemde mülakat bağlamı geçici bellekte (RAM) tutulmaktadır. İlerleyen aşamalarda sisteme **Vektör Veritabanı (Vector DB)** entegre edilerek RAG mimarisine geçiş yapılacaktır. Bu sayede model;

- Adayın yüklediği CV'yi (PDF) okuyup analiz edebilir,
- Şirketin spesifik politikalarını ve iş tanımlarını dokümanlardan öğrenip buna göre soru sorabilir.

## 6.3 Sesli Etkileşim ve Duygu Analizi (Emotion AI)

Mevcut sistem, kullanıcıdan metin tabanlı (Text-based) girdi almaktadır. İlerleyen aşamalarda sisteme **Konuşmadan Metne (Speech-to-Text / STT)** modülü entegre edilerek tam sesli iletişime geçilecektir. Bu sayede, sadece söylenen kelimelere değil, "nasıl söylendiğine" de odaklanabilecektir. Kullanıcının ses tonundaki titreşimlerden, duraksamalardan ve konuşma hızından stres seviyesi ölçülecek; avatar bu duruma göre daha sakinleştirici veya teşvik edici bir tavır takınacaktır.

# 7 Sonuç

Büyük Dil Modelleri dersi kapsamında gerçekleştirilen bu projede; adayların teknik mülakat süreçlerine hazırlık aşamasında yaşadıkları deneyim eksikliği ve geri bildirim yetersizliği problemlerine çözüm sunan, uçtan uca (end-to-end) bir "Yapay Zeka Destekli Mülakat Simülasyon Sistemi" geliştirilmiştir. Teorik araştırmalar ve pratik uygulamalar neticesinde elde edilen bulgular aşağıda özetlenmiştir:

- Model Adaptasyonu ve Verimlilik:** Meta'nın Llama-3-8B modeli, genel amaçlı bir dil modelinden çıkarılarak, İnsan Kaynakları alanına özgü (domain-specific) bir uzmana dönüştürülmüştür. **QLoRA** (4-bit Quantization) tekniği sayesinde, endüstriyel seviyedeki bu modelin bellek gereksinimi **6GB seviyelerine indirilmiş**, böylece sistemin yüksek maliyetli sunuculara ihtiyaç duymadan, tüketici sınıfı donanımlarda bile çalışabileceği gösterilmiştir.
- Çok Modlu Etkileşim (Multimodality):** Sadece metin tabanlı sohbet botlarının ötesine geçilerek, **Dudak Senkronizasyonu (Lip Sync)** ve **3D Avatar** teknolojileri sisteme entegre edilmiştir. Geliştirilen algoritma, 1 saniyenin altındaki (sub-second) gecikme süresiyle gerçek zamanlı bir mülakat atmosferi yaratmış ve kullanıcı deneyimini (UX) artırmıştır.
- Başarım ve Doğruluk:** Eğitim sürecinde elde edilen düşük kayıp (Loss: 1.01) ve test aşamasındaki yüksek anlamsal benzerlik skorları (BERTScore: 0.88), modelin ezber yapmadığını ve bağlamsal farkındalığı sahip olduğunu göstermiştir. Ayrıca, **LLM-as-a-Judge** yöntemiyle yapılan otonom değerlendirmelerde sistemin "profesyonellik" ve "teknik doğruluk" kriterlerinde yüksek performans sergilediği doğrulanmıştır.
- Sektörel ve Sosyal Katkı:** Geliştirilen "AI İK Uzmanı", adaylara 7/24 ulaşılabilir, objektif ve etkileşimli bir mülakat provası imkanı sunarak, fırsat eşitliğine katkı sağlayan bir eğitim aracı potansiyeli taşımaktadır.

Sonuç olarak; bu çalışma, modern dil modellerinin doğru optimizasyon teknikleriyle donanım kısıtlarını aşarak, gerçek dünya problemlerine yönelik ne kadar etkili, ölçülebilir ve yenilikçi çözümler sunabileceğinin somut bir göstergesidir.

## Kaynaklar

- [1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). *Attention is all you need*. Advances in neural information processing systems, 30.
- [2] Meta AI. (2024). *The Llama 3 Model Card*. Retrieved from <https://github.com/meta-llama/llama3>
- [3] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2021). *LoRA: Low-Rank Adaptation of Large Language Models*. arXiv preprint arXiv:2106.09685.
- [4] Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). *QLoRA: Efficient Finetuning of Quantized LLMs*. arXiv preprint arXiv:2305.14314.

- [5] Unsloth AI. (2024). *Faster and Memory Efficient Fine-Tuning*.  
<https://github.com/unslothai/unsloth>
- [6] Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). *BERTScore: Evaluating Text Generation with BERT*. International Conference on Learning Representations (ICLR).