Üniversite	**: T.C. İstanbul Kültür Üniversitesi**
Enstitü	**: Lisansüstü Eğitim Enstitüsü**
Anabilim Dalı	**: Bilgisayar Mühendisliği**
Program	**: Bilgisayar Mühendisliği Tezli YL**
Tez Danışmanı	**: Prof. Dr. Akhan Akbulut**
Tez Türü ve Tarihi	**: Yüksek Lisans – OCAK 2026**

ÖZET

BÜYÜK DİL MODELLERİ (LLM) VE ALMAYLA ARTTIRILMIŞ ÜRETİM (RAG) TABANLI YAKLAŞIM İLE İNSAN KAYNAKLARINDA BELGELERDEN BİLGİ ÇIKARIMI

Büyük Dil Modelleri (LLM) ve Almayla Arttırılmış Üretim (RAG) teknolojileri, İnsan Kaynakları departmanlarının belge yoğunluklu iş süreçlerini köklü bir biçimde dönüştürmektedir. Karmaşık formatlara sahip ve farklı dillerde yazılmış özgeçmişlerden (CV) saniyeler içinde anlamlı veri çıkarmayı mümkün kılan bu teknolojiler; beraberinde "halüsinasyon" (yanlış bilgi üretimi) ve veri mahremiyeti (KVKK) gibi kritik riskleri de getirmektedir. Bu tez çalışması, söz konusu riskleri minimize ederek karar destek süreçlerini daha güvenilir kılmayı ve literatürde genellikle ayrı ele alınan sıralama kalitesi (nDCG) ile operasyonel verimlilik (gecikme süresi) arasındaki dengeyi optimize etmeyi amaçlamaktadır. Çalışma kapsamında, 100.800 adet belge ve 154 doğal dil sorgusundan oluşan geniş ölçekli bir veri seti üzerinde; ColBERT, GraphRAG, Saf Vektör (Dense) ve BM25 mimarileri karşılaştırmalı olarak analiz edilmiştir. Deneysel bulgular, İK terminolojisinde vektör tabanlı modellerin "Semantik Kayma" (Semantic Drift) problemine yol açtığını; örneğin "Deneyimli" ve "Deneyimsiz" gibi zıt kavramlar arasında %86.4 oranında hatalı benzerlik kurduğunu ortaya koymuştur. Veri güvenliğini en üst düzeyde tutmak adına, bulut tabanlı API'ler yerine Docker üzerinde izole edilmiş yerel Ollama altyapısı ve Mistral 7B modeli tercih edilmiştir.

Semantik kayma problemini aşmak için geliştirilen Ağırlıklı RRF (Weighted Reciprocal Rank Fusion, α=0.8) stratejisi, 0.654 nDCG@5 skoru ile saf vektör yaklaşımını (0.365) ve modern ColBERT mimarisini (0.318) geride bırakarak en yüksek başarıyı elde etmiştir. Sistemin en dikkat çekici yönü ise, Redis tabanlı semantik önbellekleme mekanizması sayesinde yanıt sürelerini 1550ms'den 50ms'ye düşürerek 31 kat hızlanma sağlamasıdır. Eklenen "konuşma geçmişi" özelliği ile bağlam koruma başarısı %95 seviyesine ulaşmış; sistem, operasyonel hedefleri tutturarak kurumsal ölçekte yüksek doğruluklu ve düşük maliyetli bir mimari

sunmuştur. Gelecek çalışmalarda, özgeçmişlerdeki görsel içeriklerin de işlenebildiği çok modlu (multi-modal) bir yapıya geçilmesi hedeflenmektedir.

Anahtar Kelimeler: Büyük Dil Modelleri, RAG, İnsan Kaynakları, belge işleme, semantik arama, aday–rol eşleştirme, çok dillilik, KVKK, GDPR, hibrit arama, vektör veritabanı, semantik önbellekleme.

University	**: T.C. İstanbul Kültür University**
Institute	**: Institute Of Graduate Studies**
Department	**: Computer Engineering**
Program	**: Computer Engineering**
Thesis Advisor	**: Prof. Dr. Akhan Akbulut**
Degree Awarded And Date	**: MA – January 2026**

ABSTRACT

A Large Language Model (LLM) and Retrieval-Augmented Generation (RAG)-Based Approach for Information Extraction from Human Resources Documents

Large Language Models (LLM) and Retrieval-Augmented Generation (RAG) technologies are radically transforming document-intensive business processes within Human Resources departments. While extracting meaningful data from complex, multilingual, and unstructured resumes (CVs) in seconds is now possible, this capability introduces critical risks such as "hallucination" (generation of false information) and data privacy (KVKK/GDPR) concerns. This study aims to minimize these risks to ensure reliable decision support processes and fills a gap in the literature by optimizing the balance between ranking quality (nDCG) and operational efficiency (latency).

Within the scope of the study, a comparative analysis was conducted on a large-scale dataset comprising 100,800 documents and 154 natural language queries. The architectures evaluated included ColBERT, GraphRAG, Pure Vector (Dense), and BM25. To ensure maximum data security, an isolated local Ollama infrastructure running on Docker and the Mistral 7B model were preferred over cloud-based APIs. Experimental findings revealed that vector-based models in HR terminology lead to a "Semantic Drift" problem; for instance, opposing concepts such as "Experienced" and "Inexperienced" exhibited an erroneous similarity rate of 86.4%.

To overcome this challenge, a Weighted Reciprocal Rank Fusion (Weighted RRF, $\alpha=0.8$) strategy was developed. This approach achieved an nDCG@5 score of 0.654, significantly outperforming both the pure vector approach (0.365) and the modern ColBERT architecture (0.318). Furthermore, thanks to the Redis-based semantic caching mechanism, response latency was reduced from 1550ms to 50ms, achieving a 31x speedup. With the addition of a conversation history feature, context preservation success reached 95%. Consequently, the system met operational targets, offering a high-accuracy, cost-effective, and secure architecture at an enterprise scale.

Future steps aim to transition to a multi-modal structure capable of analyzing photos and graphics within CVs.

Keywords: Large Language Models, RAG, Human Resources, document processing, semantic search, candidate-role matching, multilingualism, KVKK, GDPR, hybrid search, vector database, semantic caching.