



4. Tez İzleme Komite Raporu (TİK-4)

Hazırlayan: M. Ali Bayram

Danışman: Prof. Dr. Banu Diri

Üniversite: Yıldız Teknik Üniversitesi, Fen Bilimleri Enstitüsü

Program: Bilgisayar Mühendisliği Doktora Programı

Tez Başlığı: Düşük Kaynaklı Diller ve Ortamlara Büyük Dil Modellerinin Adapte Edilmesi

Rapor Tarihi: Ekim 2025

1. Bu Dönem Çalışmalarının Özeti

Önceki tez izleme komitesinden bu yana geçen sürede, doktora tez çalışmamın hem akademik yayın hedeflerini tamamlama hem de teknik altyapısını güçlendirme noktasında kritik adımlar atılmıştır. Bu dönemin en önemli başarısı, **doktora mezuniyeti için gerekli olan tüm resmi şartların tamamlanmış olmasıdır.**

Dönemin öne çıkan gelişmeleri aşağıda özetlenmiştir:

- ✓ Mezuniyet Şartları Tamamlandı:**
 - Dergi Yayını:** "Healthcare-Focused Turkish Medical LLM" başlıklı makalemiz, **Q2** düzeyindeki **ACM TALLIP** dergisinde kabul edilerek en önemli mezuniyet koşulu sağlanmıştır.
 - Konferans Bildirisi:** Daha önce kabul alan 4 konferans bildirisi sunulmuş ve bildiri kitapçıklarında yayımlanmıştır.
 - TİK Süreci:** Bu rapor ile birlikte dördüncü TİK toplantısı gerçekleştirilerek gerekli komite sayısı tamamlanmıştır. Tez yazımı ve savunma süreci dışında bir engel kalmamıştır.
- 🇹🇷 Yeni Veri Seti Katkıları:** Türkçe doğal dil işleme ekosistemini zenginleştirmek amacıyla, çalışma arkadaşlarım **Umut Ertuğrul Daşgın, Alp Dikmen, Muhammed Enes Çam ve Ceyhun Batman** ile birlikte farklı alanlarda (tıp, haber, teyit analizi, kültür) **6 adet yüksek kaliteli ve büyük ölçekli veri seti** oluşturulmuş ve açık kaynak olarak yayımlanmıştır.
- 🧠 Kapsamlı Deneysel Altyapı Kurulumu:** Reddedilen bir makalemizdeki eksiklikleri gidermek ve tokenizasyon yaklaşımlarının etkisini derinlemesine analiz etmek üzere, çalışma arkadaşlarım **Ali Arda Fincan** ve **Ahmet Semih Gümüş** ile birlikte **toplam 8 adet küçük ölçekli dil modelinin eğitimi** sürecine başlanmıştır.
- 🔧 Teknik Araç Geliştirme ve Yayınlama:** Geliştirdiğimiz morfolojik Türkçe tokenizer'ın **decoder modülü** tamamlanmış, **turkish-tokenizer** adıyla **PyPI paketi** olarak ve Hugging Face üzerinde interaktif bir **Space** uygulaması olarak herkesin kullanımına sunulmuştur.
- 📱 Tezin Somut Çıktısı: Alan Odaklı Chatbot Uygulaması:** Tezin ana çıktılarından biri olan ve toplanan **sağlık ve hukuk veri setlerini** temel alan **LLM tabanlı chatbot uygulamasının geliştirme sürecinde son aşamaya gelinmiştir.**
- 👥 Akademik Topluluk Oluşturma:** **magibu.web.app** platformu üzerinden, alanında uzman isimlerin (OpenAI, Huawei) ve lisansüstü öğrencilerin katıldığı aktif bir araştırma grubu kurularak

2. Yayın Durumu ve Akademik Katkıları

Bu dönemde yayın hedefleri başarıyla tamamlanmış ve devam eden çalışmaların altyapısı güçlendirilmiştir.

✅ Kabul Edilen Makale (Mezuniyet Şartı Sağlandı)

- **Başlık:** *Healthcare-Focused Turkish Medical LLM: Training on Real Patient-Doctor Question-Answer Data for Enhanced Medical Insight*
- **Dergi:** ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP) - Q2
- **Durum:** **Kabul Edildi** (Minor revision sonrası).
- **Önemi:** Bu makalenin kabulü ile doktora mezuniyeti için gerekli olan **Q2 dergi yayın şartı sağlanmıştır**.

🔄 İnceleme Sürecindeki Makaleler

1. **Başlık:** *Tokenization Standards for Linguistic Integrity: Turkish as a Benchmark* (ACM TALLIP)
2. **Başlık:** *Setting Standards in Turkish NLP: TR-MMLU for Large Language Model Evaluation* (International Journal of Pattern Recognition and Artificial Intelligence)

🔬 Revize Edilen ve Geliştirilen Makale

- **Başlık:** *Tokens with Meaning: A Hybrid Tokenization Approach for NLP*
- **Durum:** DeneySEL yetersizlikler nedeniyle **Reddedildi**.
- **Aksiyon:** Red kararına yol açan eksiklikleri gidermek amacıyla, aşağıda detaylandırılan kapsamlı ve kontrollü yeni bir deney tasarımı oluşturulmuş ve model eğitimlerine başlanmıştır.

3. Teknik Katkıları ve Altyapı Geliştirmeleri

3.1. Yeni ve Kapsamlı Tokenizasyon Deneyleri

İşbirliği: Ali Arda Fincan, Ahmet Semih Gümüş

Geliştirdiğimiz morfolojik tokenizer'ın etkinliğini kanıtlamak amacıyla, kontrollü bir deney ortamı tasarlanmıştır. Bu kapsamda, her biri ~50 milyon parametrelili olan toplam **8 farklı dil modeli** eğitilmektedir. Bu deneyler, tokenizer seçiminin model performansı üzerindeki etkisini nicel olarak ölçecektir.

3.2. Türkçe Morfolojik Tokenizer'ın Tamamlanması ve Yayınlanması

Encoder modülüne ek olarak **decoder modülü** de tamamlanan tokenizer, `pip install turkish-tokenizer` komutuyla kurulabilen bir **PyPI paketi** ve **Hugging Face Space** uygulaması olarak yayınlanmıştır.

3.3. Yeni Türkçe Veri Setleri

Türkçe NLP araştırmaları için yüksek kaliteli veri ihtiyacını karşılamak amacıyla 6 yeni veri seti oluşturulmuş ve açık lisanslarla yayımlanmıştır:

- **Tıp:** *Turkish Medical Articles from 14 Hospital Websites* ve *Doktorsitesi.com Turkish Medical Articles* (**Umut Ertuğrul Daşgın** ile)
- **Haber:** *Turkish_HQ_Articles* (**Alp Dikmen** ile)
- **Teyit Analizi:** *teyit-analizleri* (**Muhammed Enes Çam** ile)
- **Kültür/Edebiyat:** *turkce_masallar_extended* ve *tarihtebugun* (**Ceyhun Batman** ile)

3.4. Tezin Uygulama Çıktısı: Sağlık ve Hukuk Odaklı LLM Chatbot

Tez kapsamında geliştirilen teorik ve altyapısal çalışmaların somut bir çıktısı olarak, alan odaklı bir LLM chatbot uygulaması geliştirilmiştir. Geliştirme sürecinin son aşamasına geline bu uygulama, aşağıdaki temel bileşenlerden oluşmaktadır:

- **Modeller:** Toplanan sağlık ve hukuk verileriyle **fine-tune edilmiş** özel dil modelleri.
- **Bilgi Kaynağı:** Tez sürecinde derlenen yüksek kaliteli ve güvenilir veri setleri.
- **Teknoloji: Retrieval-Augmented Generation (RAG)** mimarisi kullanılarak, modelin halüsinasyon üretmesi engellenmekte ve verdiği cevapları doğrudan bilgi kaynağından referans göstererek sunması sağlanmaktadır.
- **Amaç:** Kullanıcılara özellikle sağlık ve hukuk gibi hassas konularda güvenilir, doğru ve kanıta dayalı bilgiler sunan bir danışma aracı geliştirmek.

4. Akademik Topluluk Oluşturma ve İşbirlikleri

Magibu AI Research Group (magibu.web.app) çatısı altında OpenAI'dan **Sercan Karakaş** ve Huawei'den **Çağrı Yeşil** gibi sektör liderlerinin de katılımıyla düzenli akademik tartışma oturumları gerçekleştirilmektedir.

5. Devam Eden ve Planlanan Çalışmalar

- **Deneyler ve Makale:** Tokenizasyon deneylerini tamamlayıp sonuçlarla "Tokens with Meaning" makalesini yeniden gönderime hazırlamak.
- **Uygulama:** LLM chatbot uygulamasının son testlerini tamamlamak ve bir demo sürümü yayınlamak.
- **Entegrasyon ve Yazım:** Tüm deneysel sonuçları, araçları ve uygulama çıktılarını tez metnine entegre ederek yazımı tamamlamak.
- **Savunma Hazırlığı:** Tezi savunmaya hazır hale getirmek ve sunumu hazırlamak.


6. Genel Değerlendirme

Bu tez izleme dönemi itibarıyla, doktora mezuniyeti için gerekli olan tüm akademik ve idari şartlar **başarıyla tamamlanmıştır**. Yayın hedeflerine ulaşılmış, tezin teorik ve pratik altyapısı yeni veri setleri, tamamlanmış bir tokenizer aracı ve kapsamlı bir deney tasarımı ile en üst seviyeye taşınmıştır.

Ayrıca, geliştirilen **LLM chatbot uygulaması**, yapılan tüm teorik ve altyapı çalışmalarının gerçek dünya problemlerine çözüm üreten **somut bir çıktıya dönüştüğünü göstermektedir**. Geline noktada, tüm odak tez metninin bütüncül bir şekilde kaleme alınması ve savunma sürecine hazırlanılmasıdır.

Hazırlayan:

 M. Ali Bayram – malibayram20@gmail.com

 Yıldız Teknik Üniversitesi, Bilgisayar Mühendisliği ABD

 GitHub: [@malibayram](#) | LinkedIn: [@mehmetalibayram](#) | Hugging Face: [@alibayram](#)

 Arxiv: [Tokenization Standards for Linguistic Integrity: Turkish as a Benchmark](#)

 Arxiv: [Setting Standards in Turkish NLP: TR-MMLU for Large Language Model Evaluation](#)