



3. Tez İzleme Komite Raporu (TİK-3)

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: Mayıs 2025

1. Bu Dönem Çalışmalarının Özeti

İkinci tez izleme raporunun ardından, doktora çalışmam kapsamında yürüttüğüm araştırmalar aşağıda özetlenmiştir:

- Türkçe doğal dil işleme (NLP)** alanında özgün bir tokenizasyon yaklaşımı geliştirildi.
- TR-MMLU** adını verdiğim Türkçe çoktan seçmeli sınav tabanlı benchmark oluşturularak 39 farklı büyük dil modeli (LLM) test edildi.
- Türkçe sağlık alanına özel **hasta-soru doktor-cevap** veri kümesi çıkarılarak medikal LLM eğitimi gerçekleştirildi.
- Yürütülen çalışmalardan elde edilen sonuçlar doğrultusunda **2 konferans bildirisi** SIU 2025'e kabul edildi.
- 4 bilimsel makale** farklı dergilere gönderildi.

2. Konferans Bildirileri

SIU 2025 – Sinyal İşleme ve İletişim Uygulamaları Kurultayı

Tarih: 25-28 Haziran 2025

Yer: Işık Üniversitesi, Şile Yerleşkesi

Bildiri 1: Tokenizasyon Standartları ve Ölçümü: Türkçe Üzerinden Büyük Dil Modellerinin Karşılaştırmalı Analizi

- Tokenizer benchmark'ı geliştirilmiş, %TR ve %Pure gibi yeni metriklerle değerlendirme yapılmıştır.
- 15'ten fazla tokenizer karşılaştırılmıştır.

Bildiri 2: AI Destekli Türkçe Çoktan Seçmeli Sınav Veri Kümesi: Model Değerlendirme ve Karşılaştırmalı Analiz

- TR-MMLU veri seti tanıtılmıştır: 6200 soru, 62 kategori, 39 model.
- Modeller Hugging Face üzerinden yayımlanan açık veri ve değerlendirme sistemleri ile analiz edilmiştir.
- GPT-4o, Claude 3.5 gibi ileri modellerin yüksek performansı, Türkçe özelinde modellerin alan bazlı başarısı tartışılmıştır.

3. Dergi Makaleleri

1. Tokenization Standards for Linguistic Integrity: Turkish as a Benchmark

- **Dergi:** ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)
- **Gönderim Tarihi:** 23 Nisan 2025
- **Durum:** İnceleme Aşamasında
- **İçerik:** Tokenizasyon kalitesinin Türkçedeki model doğruluğuna etkisi gösterilmiş, benchmark metrikleri ve değerlendirme sistemi sunulmuştur.

2. Healthcare-Focused Turkish Medical LLM: Training on Real Patient-Doctor Question-Answer Data for Enhanced Medical Insight

- **Dergi:** ACM TALLIP
- **Gönderim Tarihi:** 29 Kasım 2024
- **Durum:** İnceleme Aşamasında
- **İçerik:** Gerçek Türkçe hasta-doktor verisiyle finetune edilmiş olan medikal LLM'in yapısı, doğruluğu ve sınıflandırma başarısı anlatılmıştır.

3. Tokens with Meaning: A Hybrid Tokenization Approach for NLP

- **Dergi:** Language Resources and Evaluation (Springer Nature)
- **Gönderim Tarihi:** 23 Nisan 2025
- **Durum:** Editöre Atanma Aşamasında
- **İçerik:** Morfolojik yapıya duyarlı, hibrit bir tokenizer geliştirilmiş ve TR-MMLU üzerinde test edilmiştir.

4. Setting Standards in Turkish NLP: TR-MMLU for Large Language Model Evaluation

- **Dergi:** International Journal of Pattern Recognition and Artificial Intelligence
- **Gönderim Tarihi:** 23 Nisan 2025
- **Durum:** İnceleme Sürecinde
- **İçerik:** TR-MMLU veri seti, değerlendirme yöntemi, liderlik tablosu ve 39 modelin performansı sunulmuştur.

4. Teknik Katkılar

✓ Morfolojik Tokenizer

- Kök-ek ayrımı, ses olayı düzeltmeleri ve istatistiksel verilerle desteklenen tokenizer geliştirildi.
- Tokenizer, hem encoding hem decoding süreçlerinde **çok-anlamlılık** ve **morfolojik tutarlılık** sağlamaktadır.
- BPE ile entegre edilmiş, TR-MMLU benchmark üzerinde üstün başarı göstermiştir.

✓ TR-MMLU Benchmark Sistemi

- Türkiye'deki sınav sistemlerinden (AUZEF, ÖSYM, Açık Öğretim) toplanan 6200 soruyla oluşturulmuştur.
- 62 farklı kategori (hukuk, tıp, tarih, mantık, fen bilimleri, edebiyat, vs.).
- Değerlendirme sonuçları Hugging Face üzerinde açık kaynak olarak yayımlanmıştır.

- Tokenizer ve model karşılaştırmaları bilimsel makalelerde detaylandırılmıştır.

5. Devam Eden Çalışmalar

Model Eğitimi

- Geliştirilen tokenizer ve özel embedding yapıları ile küçük ölçekli bir LLM veya BERT modeli ön-eğitilecektir.
- Bu model, Türkçe sınıflandırma görevlerinde fine-tune edilecektir, özellikle medikal sınıflandırma görevleri için.

6. Planlanan Çalışmalar

Türkçe Sağlık Triage Uygulaması (Sonraki TİK'e Kadar)


- Geliştirilecek uygulama ile:
 - Kullanıcıdan alınan semptomlara göre doğru poliklinik yönlendirmesi yapılabilecek.
 - Ön tanı ve doktor randevusu öncesi bilgilendirme yapılabilecek.

7. Genel Değerlendirme

Bu dönemde yürütülen çalışmalar, doktora tezimin hem **teorik altyapısını güçlendirmiş** hem de **uygulamaya dönük çıktılar üretmiştir**. Türkçeye özgü NLP sistemleri, değerlendirme araçları ve modeller literatürde ilk kez bu düzeyde bütüncül şekilde ortaya konmuştur. Yapılan yayınlar ile ulusal ve uluslararası akademik katkı hedeflenmiştir. Bir sonraki tez izleme döneminde, bu çalışmaların tamamlayıcı uygulamaları sunularak tez tamamlanma aşamasına getirilecektir.

Hazırlayan:

 M. Ali Bayram – malibayram20@gmail.com

 Yıldız Teknik Üniversitesi, Bilgisayar Mühendisliği ABD

 GitHub: <https://github.com/malibayram>

 LinkedIn: <https://www.linkedin.com/in/mehmetalibayram/>

 Hugging Face: <https://huggingface.co/alibayram>

 Arxiv: Tokenization Standards for Linguistic Integrity: Turkish as a Benchmark

 Arxiv: Setting Standards in Turkish NLP: TR-MMLU for Large Language Model Evaluation