T.C.
YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

TOKENİZASYON, VERİ KALİTESİ VE ALAN UYARLAMASI EKSENİNDE TÜRKÇE DİL MODELLERİNİN GELİŞTİRİLMESİ VE DEĞERLENDİRİLMESİ

Mehmet Ali BAYRAM

DOKTORA TEZİ
Bilgisayar Mühendisliği Anabilim Dalı
Bilgisayar Mühendisliği Programı

Danışman
Prof. Dr. Banu DİRİ

İkinci Danışman
Doç. Dr. Savaş YILDIRIM

Nisan, 2026

T.C.
YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

TOKENİZASYON, VERİ KALİTESİ VE ALAN UYARLAMASI EKSENİNDE TÜRKÇE DİL MODELLERİNİN GELİŞTİRİLMESİ VE DEĞERLENDİRİLMESİ

Mehmet Ali BAYRAM tarafından hazırlanan tez çalışması 02.04.2026 tarihinde aşağıdaki jüri tarafından Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı Bilgisayar Mühendisliği Programı **DOKTORA TEZİ** olarak kabul edilmiştir.

Prof. Dr. Banu DİRİ
Yıldız Teknik Üniversitesi
Danışman

Doç. Dr. Savaş YILDIRIM
İstanbul Bilgi Üniversitesi
İkinci Danışman

Jüri Üyeleri

Prof. Dr. Banu DİRİ, Danışman
Yıldız Teknik Üniversitesi _____

Prof. Dr. Mehmet Fatih AMASYALI, Üye
Yıldız Teknik Üniversitesi _____

Doç. Dr. Ayşe Berna ALTINEL GİRGİN, Üye
Marmara Üniversitesi _____

Prof. Dr. Tunga GÜNGÖR, Üye
Boğaziçi Üniversitesi _____

Dr. Öğr. Üyesi Göksel BİRİCİK, Üye
Yıldız Teknik Üniversitesi _____

Danışmanım Prof. Dr. Banu DİRİ sorumluluğunda tarafımca hazırlanan "Tokenizasyon, Veri Kalitesi ve Alan Uyarlaması Ekseninde Türkçe Dil Modellerinin Geliştirilmesi ve Değerlendirilmesi" başlıklı çalışmada veri toplama ve veri kullanımında gerekli yasal izinleri aldığımı, diğer kaynaklardan aldığım bilgileri ana metin ve referanslarda eksiksiz gösterdiğimi, araştırma verilerine ve sonuçlarına ilişkin çarpıtma ve/veya sahtecilik yapmadığımı, çalışmam süresince bilimsel araştırma ve etik ilkelerine uygun davrandığımı beyan ederim.

Bu tez çalışması sürecinde Üretken Yapay Zekâ (ÜYZ) araçları kullanılmamıştır. Tez çalışmasında yer alan tüm bilgi, veri, analiz, iddia ve referansların doğruluğu, güncelliği ve bilimsel geçerliliğine ilişkin sorumluluk tamamen tarafıma aittir. Beyanımın aksinin tespiti hâlinde doğabilecek her türlü idari ve hukuki sorumluluğu kabul ederim.

Mehmet Ali BAYRAM

İmza

Aileme ve arkadaşlarıma adanmıştır

TEŞEKKÜR

Bu tez çalışmasının her aşamasında değerli katkılarıyla yanımda olan, neredeyse tüm çalışmalarda birlikte yol aldığımız Ali Arda FİNCAN ve Ahmet Semih GÜMÜŞ başta olmak üzere Magibu Araştırma Grubu'nun tüm üyelerine teşekkür ederim. Süreç boyunca daima yanımda olup, desteğini esirgemeyen Gizem EYİ'ye en içten teşekkürlerimi sunarım. Gülşah Gökhan GÖKÇEK başta olmak üzere Yeditepe Üniversitesi'ndeki çalışma arkadaşlarıma müteşekkirim. Özel bir teşekkürü de lisans ve yüksek lisansım boyunca yanımda olan ve desteklerini esirgemeyen Zeynep BAL'a borç bilirim.

Mehmet Ali BAYRAM

İÇİNDEKİLER

SİMGE LİSTESİ ix

KISALTMA LİSTESİ xi

ŞEKİL LİSTESİ xiii

TABLO LİSTESİ xiv

ÖZET xv

ABSTRACT xvii

1 GİRİŞ 1

1.1 Alanın Gelişimi ve Literatür Özeti 4
1.2 Araştırma Problemi ve Motivasyon 7
1.3 Tezin Amacı ve Hipotezleri 8
1.4 Tezin Özgün Katkıları 11
1.5 Tezin Organizasyonu 13
1.6 Bölüm Özeti 13

2 TEMEL KONSEPTLER VE TANIMLAR 14

2.1 NLP ve Transformer Temelleri 14
2.2 Tokenizasyon ve Morfolojik Zorluklar 15
2.3 Parametre Verimli Alan Uyarlaması 16
2.4 Değerlendirme Çerçevesi 17

3 İLGİLİ ÇALIŞMALAR 18

3.1 Türkçe ve Düşük Kaynaklı Doğal Dil İşleme 19
3.2 Tokenizasyon ve Biçimbilim 21
3.3 Veri Kümeleri, Korpuslar ve Karşılaştırıcılar 24
3.4 Değerlendirme, Metrikler ve Test Süreçleri 27
3.5 Temel Modeller ve Mimariler 30
3.6 Alana Özgü ve Uygulamalı Büyük Dil Modelleri 34
3.7 Dil Modeli Uyarlaması ve Aktarımı 36

3.8 Çok Modlu ve Yeni Eğilimler . . . 39
3.9 Son Gelişmeler ve Geleceğe Yönelik Eğilimler (2024–2025) . . . 42
3.10 Bölüm Özeti . . . 45

4 VERİ SETLERİ VE KAYNAKLAR 46
4.1 TR-MMLU Ekosistemi: Türkçe Çok Alanlı Değerlendirme . . . 46
4.2 Alan-Özel Veri Setleri: Sağlık . . . 47
4.2.1 Veri Şeması . . . 47
4.2.2 Veri Toplama ve Temizlik . . . 47
4.2.3 Unvan ve Uzmanlık Dağılımı . . . 48
4.2.4 Soru ve Yanıt Uzunluk Dağılımı . . . 48
4.2.5 Kullanım Amacı . . . 49
4.3 Kontrollü Tokenizasyon Deneyi İçin Birleşik Veri Seti . . . 49
4.4 Tamamlayıcı Veri Setleri: Duygu, Teyit ve Kültürel Metinler . . . 50
4.5 Embedding Model Eğitim Kaynakları . . . 50
4.6 Embedding Değerlendirme Veri Setleri: STSbTR, TR-MTEB ve TabiBench . . . 52
4.7 Veri Yönetişimi, Etik ve Lisanslama . . . 52
4.8 Bölüm Özeti . . . 53

5 YÖNTEM VE ÖNERİLEN YAKLAŞIMLAR 54
5.1 Hipotezlerle İlişki . . . 54
5.2 Tokenizasyon Standartları ve Değerlendirme Yaklaşımı . . . 56
5.3 MFT Tokenizasyonu . . . 58
5.3.1 Sözlük Bileşimi ve Kimlik Aralıkları . . . 58
5.3.2 Kodlama (Encode) Kararı . . . 60
5.3.3 Çözme (Decode) ve Fonolojik Kurallar . . . 61
5.3.4 Kodlama-Çözme Çevrimi Doğruluğu . . . 63
5.4 DQ-ALR: Kavramsal Çerçeve . . . 64
5.5 Alan-Özel Model Uyarlama: LoRA ve Slerp . . . 66
5.6 Uzun Bağlamlı Türkçe Sentence Embedding Modeli (embeddingmagibu-200m) . . . 68
5.7 Deneysel Altyapı ve Tekrarlanabilirlik . . . 71
5.8 Kontrollü Tokenizasyon Deneyi: Alt Görev Kurulumu . . . 72
5.8.1 Veri Hazırlama . . . 72
5.8.2 Modeller ve Karşılaştırıcılar . . . 73
5.9 Bölüm Özeti . . . 74

6 DENEYSEL ÇALIŞMALAR 75
6.1 Hipotezlerle İlişki . . . 75

6.2 Deneysel Kurulum ve Ortak Protokoller . . . 76
6.3 TR-MMLU Üzerinde Model Karşılaştırmaları . . . 76
6.4 Tokenizasyon Yaklaşımlarının Deneysel Değerlendirmesi . . . 78
6.4.1 Tokenizasyon Ölçüm Hattı . . . 78
6.4.2 Tokenizasyon Verimliliği (Süre ve Token Ayak İzi) . . . 81
6.5 Alt Görevlerde Tokenizasyon Etkisi (Kontrollü Deney) . . . 82
6.5.1 STSbTR . . . 82
6.5.2 MTEB-TR Özeti . . . 84
6.5.3 TurBLiMP: Minimal Çiftlerle Dilsel Duyarlılık . . . 85
6.6 DQ-ALR: Veri Kalitesi Perspektifi . . . 87
6.7 Türkçe Sentence Embedding Modeli Deneyleri (embeddingmagibu-200m) . . . 88
6.7.1 STSbTR (Resmi Train+Test Karşılaştırması) . . . 88
6.7.2 TR-MTEB (26 Model, 26 Görev) Özet Karşılaştırması . . . 88
6.7.3 TabiBench Notu (İkincil/Supplementary) . . . 88
6.8 Sağlık Alanı Deneyleri: Türkçe Tıbbi LLM . . . 93
6.9 Hipotezlerin Deneysel Durumu . . . 94
6.10 Bölüm Özeti . . . 95

7 GÖSTERİM VE DENEYİM ORTAMLARI 97
7.1 Etkileşimli Deneyim Ortamları . . . 97
7.1.1 Tokenizer Karşılaştırma Ortamı . . . 97
7.2 Erişilebilirlik ve Yeniden Üretilebilirlik . . . 99
7.3 Açık Kaynak Destekleri ve Topluluk Etkisi . . . 99
7.4 Bölüm Özeti . . . 99

8 SONUÇ 100
8.1 Araştırmanın Genel Değerlendirmesi . . . 100
8.2 Tezin Katkıları . . . 102
8.3 Hipotezlerin Nihai Değerlendirmesi . . . 103
8.4 Sonuçların Genel Yorumları . . . 103
8.5 Gelecekteki Çalışmalar . . . 104
8.6 Genel Sonuç . . . 104
8.7 Bölüm Özeti . . . 105

KAYNAKÇA 106

A EKLER 131
A.1 Tez Kapsamında Geliştirilen Veri Setleri ve Paylaşım Bilgileri . . . 131
A.2 Açık Bilim İlkeleri ve Tekrarlanabilirlik . . . 132

A.3 Ek Kod, Model ve Kaynak Erişim Bilgileri 133
A.3.1 Kurulum 133
A.3.2 Tokenizer Eğitimi (SentencePiece BPE) 133
A.3.3 Öğretmen Modelin Yeni Tokenizer ile Klonlanması (Embedding Eşleştirme) 134
A.3.4 Öğretmen Embedding'lerinin Önceden Hesaplanması (Çevrimdışı Veri Seti) 134
A.3.5 Damıtma Eğitimi (Kosinüs Kaybı) 134
A.3.6 Yayımlanan Ek Çıktılar 134
A.4 Embedding Sonuçları: Tam Döküm 134
A.5 TR-MTEB Tam Sıralama (26 model) 135
A.6 TR-MTEB Tam Görev Matrisi (26x26) 135
A.7 TabiBench Sonuçları (İkincil / Supplementary) 141
A.8 Bölüm Özeti 141

TEZDEN ÜRETİLMİŞ YAYINLAR **142**

SİMGE LİSTESİ

A_1, A_2	Ağırlık uzayındaki uç nokta vektörleri/tensörleri (modeller)
D_{test}	Ayrılmış test veri kümesi
N_{tokens}	Cümle başına toplam token sayısı (tokenizasyon sonrası)
D_{eval}	Değerlendirme/doğrulama veri kümesi
$\%TR$	Dile özgü token yüzdesi (dilsel kapsam metriği)
η_i	DQ-ALR'de örnek bazlı öğrenme oranı
ΔL	Eğitim adımları/dönemleri arasındaki kayıp değişimi
D_{train}	Eğitim veri kümesi
$\mathbf{x}$	Girdi metin dizisi (tokenize edilmiş)
L_{task}	Göreve özgü kayıp fonksiyonu (örn. çapraz entropi)
β	Güvenilirlik ağırlıklandırma faktörü (genel)
$\mathbf{y}$	Hedef etiketler/yanıtlar
R_i	i. örneğe atanan güvenilirlik skoru (veri kalitesi)
W	İnce ayar sonrası uyarlanmış ağırlık matrisi
r	LoRA rütbesi ($r \ll \min(d, k)$)
P_{valid}	Modelin bir yanıtın geçerli olma olasılığı tahmini (öz-değerlendirme)
W_0	Önceden eğitilmiş ağırlık matrisi
$\alpha_1, \alpha_2, \alpha_3$	R_i güvenilirlik bileşenleri (kaynak, tutarlılık, etiket) ağırlıkları
λ	R_i'nin η_i üzerindeki etkisini belirleyen ölçekleme katsayısı
α	Slerp enterpolasyon parametresi (veya genel katsayı)
θ	Slerp'te A_1 ile A_2 arasındaki açısal ayrım
V	Sözlük boyutu
η_0	Temel öğrenme oranı (küresel)

$\%Pure$	Token anlamsal saflık oranı (morfolojik bütünlük metriği)
E_{emb}	Token gömme vektör temsili
γ	Token saflığı normalizasyon katsayısı (metrik kalibrasyonu)
ρ	Tokenizer metrikleri ile görev doğruluğu arasındaki Pearson korelasyonu
L_{total}	Toplam amaç fonksiyonu (düzenlileştirme terimleri dahil)
d, k	$W_0 \in \mathbb{R}^{d \times k}$ matrisinin girdi/çıktı boyutları
A, B	$\Delta W = BA$ olacak şekilde düşük rütbeli LoRA matrisleri

KISALTMA LİSTESİ

AGI	Artificial General Intelligence
AI	Artificial Intelligence
AUZEF	İstanbul Üniversitesi Açık ve Uzaktan Eğitim Fakültesi
BART	Bidirectional and Auto-Regressive Transformers
BERT	Bidirectional Encoder Representations from Transformers
BLEU	Bilingual Evaluation Understudy (generation metric)
BPE	Byte Pair Encoding
CC BY-NC	Creative Commons Attribution–NonCommercial License
CIR	Composed Image Retrieval
DOI	Digital Object Identifier
DQ-ALR	Data Quality-Aware Learning Rate
F1	Harmonic mean of precision and recall
FAIR	Findable, Accessible, Interoperable, Reusable (data principles)
FST	Finite-State Transducer
GloVe	Global Vectors for Word Representation
GLUE	General Language Understanding Evaluation
GPQA	Graduate-level Google-proof Q&A
GPT	Generative Pre-trained Transformer
GPU	Graphics Processing Unit
ICL	In-Context Learning
LLM	Large Language Model
LM	Language Model
LoRA	Low-Rank Adaptation

LSTM	Long Short-Term Memory
MFT	Morfolojik Farkındalıklı Tokenizer
MLM	Masked Language Modeling
MMLU	Massive Multitask Language Understanding
MoE	Mixture of Experts
MTEB	Massive Text Embedding Evaluation
NER	Named Entity Recognition
NLI	Natural Language Inference
NLP	Natural Language Processing
NMT	Neural Machine Translation
NSP	Next Sentence Prediction
ÖSYM	Ölçme, Seçme ve Yerleştirme Merkezi (TR national exams)
PEFT	Parameter-Efficient Fine-Tuning
RLHF	Reinforcement Learning from Human Feedback
RNN	Recurrent Neural Network
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
Slerp	Spherical Linear Interpolation (weight-space model merging)
SLM	Small Language Models
SOP	Standard Operating Procedure
SQuAD	Stanford Question Answering Dataset
STS	Semantic Textual Similarity
STSbTR	Semantic Textual Similarity Benchmark Turkish
T5	Text-to-Text Transfer Transformer
TAT	Text-Anchored Tuning
TR-MMLU	Turkish Massive Multitask Language Understanding
TR-MTEB	Turkish Massive Text Embedding Evaluation
TRM	Tiny Recursive Model

ŞEKİL LİSTESİ

Şekil 4.1 Sağlık verisi: unvan dağılımı 48
Şekil 4.2 Sağlık verisi: uzmanlık dağılımı 49
Şekil 4.3 Sağlık verisi: uzunluk dağılımı 49
Şekil 5.1 Tezde önerilen yöntemsel çerçeve. 55
Şekil 5.2 MFT tokenizasyon sürecinin algoritmik akış diyagramı. 63
Şekil 5.3 Sağlık alanı: uyarlama akışı 68
Şekil 5.4 embeddingmagibu-200m: eğitim dinamikleri 71
Şekil 6.1 Korelasyon matrisi 80
Şekil 6.2 Çok boyutlu karşılaştırma 80
Şekil 6.3 STSbTR (kontrollü) grafik 83
Şekil 6.4 STSbTR Pearson revizyon 84
Şekil 6.5 MTEB-TR ortalamalar 85

TABLO LİSTESİ

Tablo 5.1 MFT sözlük bileşimi 58
Tablo 5.2 Kontrollü alt görev deney özeti 72
Tablo 6.1 TR-MMLU üzerinde seçili model doğrulukları (%). 77
Tablo 6.2 Tokenizer karşılaştırması 79
Tablo 6.3 Hibrit tokenizer metrikleri 81
Tablo 6.4 Tokenizasyon verimliliği 81
Tablo 6.5 STSbTR (kontrollü) sonuçlar 83
Tablo 6.6 MTEB-TR (kontrollü) özet 84
Tablo 6.7 MTEB-TR (kontrollü) görev bazlı 86
Tablo 6.8 TurBLiMP (kontrollü) sonuçlar 87
Tablo 6.9 STSbTR sonuçları 89
Tablo 6.10 TR-MTEB (26 görev) özet 91
Tablo 6.11 TR-MTEB 200m-152m özet 92
Tablo 6.12 Sağlık alanı deneyleri: özet 94
Tablo 6.13 Hipotezlerin deneysel durumu 95
Tablo A.1 Tez veri setleri: platform, DOI ve sürüm bilgileri. 132
Tablo A.2 TR-MTEB tam sıralama 135
Tablo A.3 MTEB sütun kısaltmaları 136
Tablo A.4 TR-MTEB tam görev matrisi - Bölüm I 137
Tablo A.5 TR-MTEB tam görev matrisi - Bölüm II 139
Tablo A.6 TabiBench karşılaştırması 141
Tablo A.7 TabiBench tür ortalamaları 141

ÖZET

Tokenizasyon, Veri Kalitesi ve Alan Uyarlaması Ekseninde Türkçe Dil Modellerinin Geliştirilmesi ve Değerlendirilmesi

Mehmet Ali BAYRAM

Bilgisayar Mühendisliği Anabilim Dalı
Doktora Tezi

Danışman: Prof. Dr. Banu DİRİ
İkinci Danışman: Doç. Dr. Savaş YILDIRIM

Büyük dil modellerinin İngilizce gibi yüksek kaynaklı dillerde elde ettiği başarılar, Türkçe gibi eklemeli ve biçimbilimsel olarak zengin dillere aynı ölçüde yansımamaktadır. Bu tez, bu açığı tokenizasyon, veri kalitesi ve alan uyarlaması eksenlerinde, değerlendirme standardizasyonu ile birlikte ele alan tekrar üretilebilir bir araştırma çerçevesi sunmaktadır. İlk olarak, Türkiye'nin ulusal sınav sisteminden derlenen yaklaşık 280.000 sorudan seçilen 6.200 çoktan seçmeli soru ile TR-MMLU karşılaştırıcısı oluşturulmuş ve açık liderlik tablosu 85 model değerlendirmesine ulaşmıştır. Tokenizasyon ekseninde, Türkçenin kök-ek yapısını, ses olaylarını ve biçimbirimsel bütünlüğünü koruyan Morfolojik Farkındalıklı Tokenizer geliştirilmiş; önerilen Türkçe Token Oranı ve Saf Token Oranı metrikleriyle TR-MMLU metni üzerinde sırasıyla %90,29 ve %85,80 değerleri elde edilmiştir. Aynı mimari, veri ve başlangıç ağırlıkları altında yürütülen kontrollü deneylerde MFT, STSbTR ve 26 görevli TR-MTEB değerlendirmelerinde diğer Türkçe BPE tabanlı tokenizer akışlarına göre en yüksek skorları vermiş; `decode(encode(w))` doğruluğu %99,48 ile tersinir bir Türkçe metin çözümleme hattı olarak konumlandırılmıştır. Veri kalitesi tarafında, örnek güvenilirliği sinyallerini öğrenme sürecine taşıyan DQ-ALR yaklaşımı kavramsal olarak tartışılmıştır. Alan uyarlaması tarafında ise 167.000'den fazla gerçek hasta-doktor etkileşiminden derlenen Türkçe tıbbi veri seti üzerinde düşük-rank uyarlama ve ağırlık uzayı birleştirmesiyle alan-özel

modeller geliştirilmiş; ince ayarlı model uzman yanıtlarına benzerlikte taban modele göre %11,33 artış sağlamıştır. Ayrıca **embeddingmagibu-200m** modeli, STSbTR ve TR-MTEB sonuçlarıyla düşük maliyetli Türkçe cümle gömme modeli olarak değerlendirilmiştir. Tüm veri setleri, modeller ve değerlendirme çıktıları açık bilim ilkeleriyle paylaşılmıştır.

Anahtar Kelimeler: Türkçe dil modelleri, tokenizasyon, veri kalitesi, alan uyarlaması, cümle gömme

ABSTRACT

Developing and Evaluating Turkish Language Models within a Tokenization, Data Quality, and Domain Adaptation Framework

Mehmet Ali BAYRAM

Department of Computer Engineering
Doctor of Philosophy Thesis

Supervisor: Prof. Dr. Banu DİRİ
Co-supervisor: Assoc. Prof. Dr. Savaş YILDIRIM

The strong performance of large language models in high-resource languages such as English does not transfer equally to agglutinative, morphologically rich languages such as Turkish. This thesis presents a reproducible research framework that addresses this gap through tokenization, data quality, domain adaptation, and evaluation standardization. First, TR-MMLU is introduced as a Turkish benchmark built from 6,200 multiple-choice questions selected from roughly 280,000 items in Turkey's national examination system; its public leaderboard has reached 85 model evaluations. On the tokenization axis, the Morphology-Aware Tokenizer is developed to preserve root-suffix structure, phonological processes, and morphemic integrity. Using the proposed Turkish Ratio and Purity Ratio metrics, MFT achieves 90.29% and 85.80%, respectively, on TR-MMLU text. In controlled experiments where architecture, data, and initial weights are fixed and only the tokenizer varies, MFT obtains the highest scores on STSbTR and across 26 TR-MTEB tasks compared with Turkish BPE-based tokenizer streams. It also functions as a reversible Turkish text processing pipeline with 99.48% `decode(encode(w))` accuracy. On the data quality axis, the Data Quality-Based Adaptive Learning Rate approach is discussed as a conceptual strategy for integrating sample reliability signals into learning. On the domain adaptation axis, domain-specific models are trained on a Turkish medical dataset of more than 167,000 real patient-doctor interactions using low-rank adaptation and weight-space merging; the fine-tuned

model improves similarity to expert responses by 11.33% over the base model. Finally, **embeddingmagibu-200m** is evaluated on STSbTR and TR-MTEB as a cost-efficient Turkish sentence embedding model. All datasets, models, and evaluation outputs are released under open-science principles.

Keywords: Turkish language models, tokenization, data quality, domain adaptation, sentence embeddings

1
GİRİŞ

Yapay zeka teknolojilerinin özellikle son on yılda kaydettiği gelişmeler, doğal dil işleme (Natural Language Processing, NLP) alanının hem kuramsal hem de uygulamalı boyutlarını köklü bir biçimde dönüştürmüştür. 2018 sonrasında Transformer mimarisinin ve GPT benzeri büyük dil modellerinin (Large Language Models, LLM) yaygınlaşması, dil modellemesinin ölçeğini, derinliğini ve etki alanını yeniden tanımlamıştır. Bu modeller, devasa veri kümeleri üzerinde eğitilerek bağlamı küresel düzeyde temsil edebilmekte, sıfır örnekli ve az örnekli öğrenme gibi yetenekler sayesinde insan düzeyine yaklaşan performanslar sergileyebilmektedir [1]. Ancak, bu etkileyici başarıya rağmen, mevcut literatür ve model tasarımları büyük ölçüde İngilizce gibi yüksek kaynaklı dillere odaklanmaktadır. Bu durum, düşük kaynaklı ve morfolojik olarak karmaşık dillerde performansın önemli ölçüde düşmesine neden olmakta ve eşitsiz bir araştırma alanı oluşturmaktadır.

Bu giriş bölümü, tez boyunca izlenen yöntemsel hattın motivasyonunu ve problem tanımlarını netleştirir. Türkçenin yapısal özelliklerinden kaynaklanan temsil sorunları, veri kalitesindeki farklılıklar ve alan-özel uyarlama gereksinimleri birlikte ele alınır; böylece çalışmanın kapsamı ile katkı hedefleri açık bir çerçeveye oturtulur. Bu yaklaşım, teorik arka plan ile pratik uygulama gereksinimlerini aynı bağlamda görünür kılar.

Türkçe, eklemeli yapısı, yüksek morfolojik çeşitliliği, yoğun ek dizilimleri ve bağlama duyarlı anlam kaymaları nedeniyle Büyük dil modelleri (Large Language Models, LLM) için kritik bir sınama alanı oluşturmaktadır. Bir Türkçe kelime kökünden yüzlerce farklı biçimbirimsel varyasyon türeyebilmekte ve bu çeşitlilik, tokenizasyon tabanlı temsil yöntemlerinin yapısal sınırlılıklarını belirgin hale getirmektedir. Özellikle Byte Pair Encoding (BPE) ve SentencePiece gibi istatistiksel tokenizasyon yöntemleri, yüzeysel parçalama stratejileri nedeniyle kelimenin kökünü ve anlam taşıyan alt birimlerini bozarak modelin semantik bütünlüğü öğrenmesini güçleştirmektedir. Ramshaw ve Marcus'un dönüşüm

tabanlı öğrenmeye dayanan erken dönem tokenizasyon çalışmaları [2], modern yöntemlere öncülük etse de, günümüz LLM mimarileri için Türkçe gibi dillerde yetersiz kalmaktadır. Morfolojik farkındalığın eksikliği, temsil uzayında anlamsal bozulmalara neden olmakta ve Türkçede gözlemlenen performans kayıplarının önemli bir bölümünü açıklamaktadır.

Bu nedenle tez, tokenizasyonu yalnızca "model öncesi bir adım" olarak değil, ölçülebilir bir tasarım değişkeni olarak ele alır. Tezde kullanılan tokenizer karşılaştırma yaklaşımı, farklı tokenizerların aynı metin normalizasyonu altında kıyaslanabilmesini ve sonuçların tekrar üretilebilir biçimde paylaşılmasını hedefler. Bu kapsamda %TR ve %Pure gibi dilbilimsel metrikler ile token ayak izi ölçümleri birlikte raporlanarak, tokenizer tercihinin hem temsil bütünlüğü hem de hesaplama maliyeti açısından ürettiği değişim görünür kılınır (bkz. Bölüm 5, Bölüm 6 ve Bölüm 7).

Bu temsil sorunlarının kökeni, Türkçenin yalnızca "çok ek alan" bir dil olmasından değil; eklerin ve köklerin yüzey biçimlerini sistematik biçimde değiştiren morfofonolojik süreçlerden de kaynaklanır. **Ünlü uyumu (alomorfi)** aynı dilbilgisel ekin bağlama göre farklı yüzey biçimleriyle gerçekleşmesi sonucunda ortaya çıkar (örn. çoğul *-lAr*; ayrılma hâli *-dAn*/*-tAn*); **ünsüz dönüşümleri** ünlüyle başlayan eklerden önce ünsüz yumuşaması gibi düzenli değişimlerle görünür hale gelir (örn. *kitap* → *kitabı*); **eklemeli yapı** ise tek bir kelimede çoklu ek dizilerinin birikmesiyle belirginleşir (örn. *kitaplarımızdan*: *kitap* + *-lar* + *-ımız* + *-dan*). Saf alt-birim tokenizasyon yaklaşımları bu tür yüzey varyantlarını çoğu zaman ilişkili birimler olarak değil, bağımsız "parça"lar olarak ele alma eğilimindedir; bu da sözlükte gereksiz tekrarlar ve temsillerde semantik parçalanma üretebilmektedir [3].

Dil modellerinin Türkçede karşılaştığı zorluklar yalnızca tokenizasyonla sınırlı değildir. Veri kalitesi de model başarısını belirleyen temel bileşenlerden biridir. Literatürde yer alan Türkçe veri setlerinin önemli bir kısmı web kazıması niteliğinde olup doğrulanabilirlik, tutarlılık ve uzmanlık gerektiren alan bilgisini yeterince yansıtmamaktadır. Özellikle sağlık gibi bilgi yoğun ve hata toleransı düşük alanlarda, düşük kaliteli verilerle eğitilmiş modellerin üreteceği çıktılar ciddi riskler barındırmaktadır. Bu nedenle veri toplama süreçlerinin yeniden ele alınması, kaynağı bilinen ve uzman doğrulamasına yakın yapılar kullanılması kritik bir gereklilik haline gelmiştir.

Bu tez, Türkçenin yalnızca üzerinde deney yapılan bir "örnek dil" değil, aynı zamanda düşük kaynaklı dillerin temsil sorunlarını görünür kılan metodolojik bir çerçeve olduğunu varsayan bir yaklaşım benimsemektedir. Çalışma, mevcut

literatürde ayrı ayrı ele alınan üç temel bileşeni bütüncül bir yapıda bir araya getirmektedir: dilbilimsel olarak yapılandırılmış tokenizasyon, veri kalitesine duyarlı öğrenme stratejileri ve alana özgü model adaptasyonu yöntemleri. Bu bütünleşik yaklaşım, hem Türkçe NLP alanındaki tarihsel eksiklikleri gidermeyi hem de düşük kaynaklı diller için genellenebilir bir standart oluşturmayı hedeflemektedir.

Bu bağlamda tez kapsamında kapsamlı bir ekosistem geliştirilmiştir. **Massive Multitask Language Understanding (MMLU)** ailesinin Türkçe uyarlaması olarak kurgulanan **Turkish Massive Multitask Language Understanding (TR-MMLU)** karşılaştırıcısı, Türkçe eğitim sisteminden elde edilen 6.200 çoktan seçmeli soruyla LLM'lerin değerlendirilmesi için standart bir çerçeve sunmaktadır. Bu karşılaştırıcı, daha önce literatürde sağlanmamış olan dilsel, kavramsal ve kültürel uyumluluğu garanti eden yüksek kaliteli bir değerlendirme ortamı sağlamaktadır. Çalışmanın çekirdek deney fazında 39 farklı LLM bu veri seti üzerinde sistemli şekilde test edilmiş ve sonuçlar açık kaynak olarak yayımlanmıştır. Açık liderlik tablosu ise 05 Şubat 2026 itibarıyla 85 model değerlendirmesine ulaşmış; bu kapsam, TR-MMLU'yu Türkiye'de kamuya açık Türkçe MMLU karşılaştırıcıları içinde en fazla model değerlendirmesi barındıran ölçüm zemini haline getirmiştir [4]. Aynı dönemde sağlık alanında, 167.000'den fazla gerçek hasta–doktor etkileşiminden oluşan büyük ölçekli bir veri seti hazırlanmış ve Türkçe tıbbi modellerin eğitimi için kullanılmıştır. Bu veri seti, hem genişliği hem de içerdiği uzmanlık seviyesi nedeniyle Türkçe sağlık alanındaki en kapsamlı kaynaklardan biri olmuştur. Ayrıca, Türkçe odaklı ve çok dilli tokenizer sözlüğüne sahip cümle gömme modeli **embeddingmagibu-200m** geliştirilmiş ve STSbTR ile TR-MTEB üzerinde sistematik olarak değerlendirilmiştir.

Tezde ayrıca, Türkçe'nin morfolojik yapısına duyarlı **Morfolojik Farkındalıklı Tokenizer (MFT)** geliştirilmiştir. Bu yöntem, kök-ek ayrıştırmasını önceliklendiren, ses olaylarını hesaba katan ve anlamsal bütünlüğü koruyan hibrit bir yapı sunmaktadır. MFT, yalnızca bir LLM tokenizersi olarak değil; aynı zamanda *tokenize-detokenize* (kodlama-çözme) çevrimiyle metni yüksek doğrulukla geri birleştirebilen üretim odaklı bir Türkçe **metin çözümleyici** olarak konumlanır. Cosmos-Corpus veri kümesinden [5] seçilen 66.547 kelimede kelime-bazlı `decode(encode(w))` tam eşleşme doğruluğu %99,48 olarak ölçülmüş; tüm metnin tek seferde encode-decode edilip kelime hizalamasıyla ölçülen tanılayıcı metrik ise %99,84 bulunmuştur [6]. Benzer biçimde, aynı çalışmada raporlanan verimlilik ölçümleri, morfem hizalı tokenizasyonun daha yüksek token yoğunluğu üretse bile pratik bir hız ayak izini koruyabildiğini göstermektedir [6]. Kamuya açık Türkçe araçlar arasında bu düzeyde tersinirlik

ile üretim-odaklı tokenizasyon hızını birlikte raporlayan bir yaklaşım sınırlıdır; bu yönüyle MFT, Türkçe NLP için özgün bir metin çözümleme hattı sunmaktadır. Geliştirilen **MFT** yaklaşımı, derinlemesine analiz edilmek üzere 15'ten fazla tokenizer ile karşılaştırılmış; ortaya konan %TR ve %Pure gibi metrikler literatüre yeni değerlendirme ölçütleri olarak kazandırılmıştır. Bu çalışmalar, tokenizasyon kalitesinin model başarımıyla doğrudan ilişkili olduğunu nicel olarak göstermiş ve alanda önemli bir boşluğu doldurmuştur.

Model adaptasyonu tarafında ise Low-Rank Adaptation (LoRA) ve Spherical Linear Interpolation (Slerp) tabanlı yöntemler kullanılarak tıp gibi alanlara özgü modeller geliştirilmiştir. Geliştirilen DoctorLlama ve DoctorGemma serileri, hem modüler yapıları hem de düşük maliyetli uyarlama stratejileriyle Türkçe sağlık modelleri için önemli bir örnek teşkil etmektedir. Tez kapsamında ayrıca **Data Quality-Based Adaptive Learning Rate (DQ-ALR)** yaklaşımı kavramsal olarak önerilmiş; veri setindeki hekim unvanı ve uzmanlık alanı gibi meta-bilgilerin eğitim sürecinde dikkate alınabileceği bir optimizasyon perspektifi tartışılmıştır.

Bu giriş bölümü, çalışmanın kavramsal temelini oluşturmakta ve Türkçe gibi morfolojik olarak karmaşık, fakat araştırma ekosistemi sınırlı dillerde büyük dil modellerinin geliştirilmesi ve değerlendirilmesine yönelik ihtiyaçları ortaya koymaktadır. Tez, yalnızca Türkçe NLP alanına katkı sunmakla kalmayıp düşük kaynaklı diller için uygun yöntemlerin nasıl tasarlanacağına, nasıl test edileceğine ve nasıl sürdürülebilir bir araştırma ekosistemi kurulacağına ilişkin bütüncül bir yaklaşım önermektedir. Böylece çalışma, hem kuramsal hem de pratik açıdan geniş bir etki potansiyeline sahip bir araştırma çerçevesi sunmaktadır.

1.1 Alanın Gelişimi ve Literatür Özeti

Doğal dil işlemenin gelişim tarihi, dilin matematiksel modellerle temsil edilmesi yönündeki çabaların sürekli olarak yeniden yapılandığı bir süreç olarak okunabilir. 1990'lı yıllarda istatistiksel yöntemlerin yükselişiyle birlikte kelime frekanslarına ve n-gram temelli modellere dayanan erken NLP yaklaşımları, dilin yüzeysel yapısını yakalamakta belirli ölçüde başarılı olsa da, sınırlı bağlam penceresi nedeniyle karmaşık sözdizimsel ilişkileri ve uzun menzilli bağımlılıkları öğrenmede yetersiz kalmıştır. Bu dönemde Ramshaw ve Marcus'un dönüşüm tabanlı öğrenme modeli [2], cümle segmentasyonu ve sözdizimsel birimlerin belirlenmesine yönelik ilk sistematik çabalardan biri olarak öne çıkmıştır. Bu çalışma, daha sonra yaygınlaşacak alt birim temelli tokenizasyon yöntemlerinin kavramsal temelini oluştursa da, modern büyük dil modellerinin ihtiyaç duyduğu morfolojik ayrıntıları

ve anlamsal bağlamı yakalamaktan uzaktır.

2000'li yılların ortalarına gelindiğinde Türkçe ve diğer eklemeli diller için açık kaynak araçların geliştirilmesi önem kazanmaya başlamıştır. Zemberek [7], Türkçe'nin biçimbilimsel çözümlemesi için geliştirilen ilk kapsamlı sistemlerden biri olarak literatürde önemli bir yer edinmiştir. Ancak, bu dönemin yöntemleri kural tabanlı yapılar üzerine inşa edildiğinden, günümüz LLM mimarilerinin gerektirdiği ölçeklenebilir temsil gücü ve istatistiksel genelleme kapasitesi açısından sınırlı kalmıştır. Bu araçlar, özellikle büyük veri kümeleri üzerinde eğitilen modellerle entegre kullanım için uygun biçimde tasarlanmadığı için, istatistiksel tokenizasyon yöntemleri pratikte baskın hale gelmiştir.

Alt birim temelli yöntemlerin yükselişi, tokenizasyonu tüm NLP modellerinin performansına doğrudan etki eden kritik bir bileşen haline getirmiştir. BPE [8], [9], WordPiece [10] ve SentencePiece [11] gibi yaklaşımlar, dilin kelime düzeyindeki çeşitliliğini daha yönetilebilir parçalara ayırarak veri seyrekliğini azaltmayı hedeflemiştir. Bu yöntemler yüksek kaynaklı dillerde etkili sonuçlar vermiş olsa da, eklemeli ve morfolojik olarak zengin dillerde ortaya çıkan anlamsal kopukluklar literatürde giderek daha fazla eleştirilmeye başlamıştır. Türkçe kelimelerin yüzlerce olası varyasyonla ortaya çıkabilmesi, bu yöntemlerin yüzeysel parçalama stratejileriyle birleştiğinde kök ve eklerin anlam bütünlüğünü bozmasına neden olmaktadır. Toraman ve ark. [12], biçimbilimsel tokenizasyon yöntemlerinin Türkçe modellerde geleneksel altbirim yöntemlerine göre daha yüksek performans sağlayabildiğini göstermiştir. Kayalı ve ark. [13], Türkçe özetleme görevlerinde hibrit tokenizasyon stratejilerinin etkisini incelemiştir. Erkaya [14], Türkçeye özgü altbirim stratejilerinin kelime haznesi ve performans ilişkisini analiz etmiş; Pan ve ark. [15] ise Uygurca ve Türkçe gibi eklemeli dillerde morfolojik segmentasyonun Neural Machine Translation (NMT) performansını iyileştirdiğini göstermiştir.

2018 sonrası dönem, LLM'lerin yükselişi ile NLP'nin ikinci büyük kırılma anı olarak değerlendirilmiştir. Transformer mimarisi ile GPT, **Bidirectional Encoder Representations from Transformers (BERT)** ve **Text-to-Text Transfer Transformer (T5)** gibi modellerin sunduğu bağlamsal temsiller farklı görevlerde insan seviyesine yaklaşan performanslar sağlamıştır. Ancak, bu başarıların büyük çoğunluğu İngilizce gibi yüksek kaynaklı dillerde elde edilmiştir. Morfolojik açıdan karmaşık dillerde kayda değer performans kayıpları yaşanmış; literatürdeki birçok çalışma bu düşüşün temel nedenleri arasında tokenizasyon hataları, yetersiz veya düşük kaliteli veri kaynakları ve dilsel yapıyı dikkate almayan mimari tasarımları işaret etmiştir. Bu bağlamda TR-MMLU [16] ve diğer son dönem Türkçe değerlendirme çalışmaları [17] çok dilli modellerin Türkçede belirgin başarım kaybı

yaşadığını göstermiştir.

Türkçe NLP literatürünün gelişimi de aynı dönemde hız kazanmış olsa da, veriye dayalı yaklaşımların kalıcılığı çoğunlukla veri setlerinin niteliği ve kapsayıcılığıyla sınırlı kalmıştır. TR-News [18] gibi büyük ölçekli veri kümeleri önemli adımlar sunmuş olsa da, mevcut Türkçe karşılaştırıcı çalışmalarının önemli kısmı İngilizce veri setlerinin çevirilerine dayanmıştır. Bu durum hem anlamsal kaymalara hem de kültürel bağlamın bozulmasına yol açmıştır. Bezir'in MMLU-pro-TR yayını [19] bu eksikliği gidermek üzere geliştirilmiş bir örnek olarak dikkat çekmiştir. MMLU, SuperGLUE veya Stanford Question Answering Dataset (SQuAD) gibi uluslararası standartların Türkçeye doğrudan aktarılması durumunda hem soru türlerinde hem de kavramsal derinlikte kayıplar yaşandığı gösterilmiştir.

Son yıllarda tıp gibi bilgi yoğun alanlarda LLM tabanlı çözümlere yönelik artan ilgi, alan-özel modeller ve veri setleri gereksinimini daha görünür kılmıştır. Tıbbi modellerde BioBERT, ClinicalBERT ve PubMedGPT gibi sistemler [20], [21], [22] önemli başarılar göstermiştir; ancak, Türkçe için karşılaştırılabilir ölçekte alan modelleri neredeyse hiç bulunmamıştır. Osmanlı tarihi için THQuAD [23], Türkçe haber özetleme için TR-News [18], duygu analizi için TRSAv1 [24] gibi çalışmalar sınırlı alanlarda veri sunmaktadır, fakat kapsamlı alan modellerinin eğitimi için yeterli değildir.

Literatürde göze çarpan bir başka eksiklik, eğitim verisinin nitelik farklılıklarını öğrenme sürecinde gözetmeye yönelik yaklaşımların sınırlılığıdır. Klasik eğitim stratejilerinde, farklı uzmanlık düzeylerindeki kaynaklardan gelen örnekler arasında bir ayrım yapılmamaktadır. Veri kalitesi odaklı öğrenme yaklaşımlarına dair erken örnekler [25] bulunsa da, Türkçe literatürde bu konuda henüz sistematik bir tartışma sunulmamıştır.

Tüm bu gelişmeler değerlendirildiğinde, Türkçe NLP alanının hem güçlü adımlar attığı hem de büyük ölçüde parçalı ilerlediği görülmüştür. Tokenizasyon, veri seti oluşturma, çok dilli aktarım ve alan-özel adaptasyon gibi çalışmaların önemli kısmı birbirinden bağımsız ilerlemiştir. Toraman ve ark. [26], çok çeşitli uyarlama stratejilerini karşılaştırarak Türkçe dil modellerinin parçalı gelişim yapısını ortaya koymuştur. Açık kaynaklı Turkish NLP platformları [27] önemli katkılar sunmuş olsa da, veri ekosistemi bütünlüklü bir araştırma çerçevesi sunmaktan uzak kalmıştır. Bu tez, hem kapsam hem metodoloji hem de ölçek açısından literatürdeki bu boşluğu doldurmayı; tokenizasyondan veri kalitesine, alan uyarlamasından değerlendirme metriklerine uzanan çok katmanlı bir yaklaşım ortaya koymayı amaçlar.

1.2 Araştırma Problemi ve Motivasyon

Büyük dil modellerinin son yıllarda elde ettiği başarılar, doğal dil işlemenin birçok alanında köklü bir dönüşüm yaratmış olsa da, bu başarıların dilsel çeşitliliğe eşit düzeyde yansımadığı giderek daha belirgin hale gelmektedir. İngilizce gibi yüksek kaynaklı dillerde insan seviyesine yaklaşan performanslar sergileyen modeller, Türkçe gibi eklemeli ve morfolojik olarak karmaşık dillerde benzer doğruluk seviyelerine ulaşmakta zorlanmaktadır. Bu performans farkı yalnızca veri miktarıyla açıklanamaz; asıl sorun, modelin karşılaştığı girdilerin yapısal özelliklerinin temsil sürecinde bozulması, veri kalitesindeki heterojenlik ve alana özgü bilginin öğrenilmesini zorlaştıran adaptasyon eksikliklerinden kaynaklanmaktadır. Dolayısıyla Türkçe gibi dillerde LLM başarımını sınırlayan temel mekanizmaların sistemli biçimde incelenmesi, hem yerel hem de küresel araştırma ekosistemi açısından kritik bir ihtiyaçtır.

Bu tezin temel problemi, Türkçe dilinin morfolojik yapısı ile modern LLM mimarilerinin varsayımları arasındaki uyumsuzluğun ortaya çıkardığı temsil hatalarıdır. Alt birim temelli tokenizasyon yöntemleri, yüzeysel frekans dağılımlarına dayandığından kelimenin anlam taşıyan çekirdek birimlerini tutarlı biçimde koruyamamaktadır. Bir Türkçe kelime, kök ve ek ilişkisiyle karmaşık bir anlam yapısı taşıdığı için, yanlış parçalanmış bir token dizisi modelin kavramsal haritasını sistematik olarak çarpıtmakta ve genel performansta belirgin düşüşlere yol açmaktadır. Tokenizasyon aşamasında ortaya çıkan bu yapısal bozulma, daha ileri seviyelerde bilgi uydurmaya, yanlış eşleştirmelere ve özellikle uzmanlık gerektiren alanlarda ciddi güvenilirlik sorunlarına zemin hazırlamaktadır.

Model başarımını sınırlayan ikinci temel problem, Türkçedeki veri ekosisteminin heterojenliği ve niteliğe ilişkin belirsizliklerdir. Mevcut veri setlerinin önemli bir bölümü çevrimiçi platformlardan kazınmış, doğrulanabilirlik düzeyi düşük, kaynak çeşitliliği sınırlı ve çoğu zaman bağlamsal tutarsızlık barındıran içeriklerden oluşmaktadır. Sağlık gibi alanlarda kullanılacak modellerde bu tür verilerin yol açacağı hatalar yalnızca akademik bir problem değildir; bu hatalar doğrudan insan sağlığı ve hasta güvenliği üzerinde etkiler yaratabilecek niteliktedir. Veri kalitesinin modelin öğrenme dinamiklerini nasıl etkilediği literatürde yeterince incelenmemiştir ve mevcut çalışmalar verinin niteliğini model eğitim sürecine yansıtan bütüncül bir yaklaşım sunmamaktadır. Bu durum, Türkçe için güvenilir modeller geliştirilmesini daha da güçleştirmektedir.

Türkçeye özgü üçüncü büyük sorun, alana özgü bilgi temelli uyarlama yöntemlerinin sınırlı olmasıdır. Özellikle tıp gibi uzmanlık gerektiren alanlarda dil modellerinin başarısı yalnızca genel dil bilgisinden değil, uzmanlık alanının

kavramsal yapısını doğru biçimde temsil edebilmesinden geçmektedir. İngilizce literatürde BioBERT ve ClinicalBERT gibi modellerle bu yönde önemli adımlar atılmış olsa da, Türkçe için benzer ölçekte yüksek kaliteli uzmanlık modellerinin eksikliği ciddi bir boşluk yaratmaktadır. Bu eksiklik yalnızca veri yetersizliğiyle değil, aynı zamanda model adaptasyonu için kullanılan yöntemlerin Türkçe bağlamında sistemli biçimde değerlendirilmemiş olmasıyla da ilişkilidir. Alan bilgisinin modele etkin biçimde aktarılabileceği yöntemlerin belirlenmesi ve bu yöntemlerin Türkçenin yapısal özellikleriyle uyumlu biçimde yapılandırılması akademik açıdan önemli bir gerekliliktir.

Bu üç temel sorunun ortak noktası, modern LLM mimarilerinin Türkçe üzerinde çalışırken karşılaştığı çok katmanlı zorlukların tekil çözümlerle aşılmasının mümkün olmamasıdır. Tokenizasyon, veri kalitesi ve model adaptasyonu birbirinden bağımsız bileşenler değildir; aksine birinde yapılacak bir iyileştirme diğerleriyle birlikte ele alınmadığında etkisini büyük ölçüde yitirmektedir. Literatürde bu üç ekseni aynı çatı altında inceleyen bütüncül bir metodolojik yaklaşıma henüz rastlanmamış olması, bu tezin ortaya çıkış motivasyonunun temelini oluşturmaktadır.

Bu çalışma, Türkçe doğal dil işlemede gözlemlenen bu yapısal sınırlılıkları gidermeyi ve dil modellerinin Türkçede daha güvenilir, daha doğru ve daha ölçeklenebilir biçimde çalışmasını mümkün kılacak bir çerçeve sunmayı amaçlamaktadır. Tokenizasyona dilbilimsel temelli bir yaklaşım getirerek temsil uzayını iyileştirmek, veri kalitesini eğitim sürecine doğrudan entegre eden bir öğrenme stratejisi oluşturmak ve alan-özel model adaptasyon tekniklerini sistemli biçimde karşılaştırmak, bu motivasyonun temel bileşenleri arasındadır. Tezin nihai amacı, yalnızca Türkçe için daha iyi dil modelleri geliştirmek değil; düşük kaynaklı diller için genellenebilir, sürdürülebilir ve açık bilim yaklaşımına dayanan bir araştırma ekosistemi oluşturmaktır.

1.3 Tezin Amacı ve Hipotezleri

Bu tez, büyük dil modellerinin Türkçe gibi morfolojik açıdan zengin ve düşük kaynaklı dillerde karşılaştığı temel zorlukları bütüncül bir şekilde ele alarak, hem temsil düzeyinde hem de model uyarlama sürecinde daha güvenilir, tutarlı ve ölçeklenebilir çözümler sunmayı amaçlamaktadır. Giriş bölümünde tartışıldığı üzere Türkçe, eklemeli yapısı ve yüksek biçimbirimsel çeşitliliği nedeniyle modern LLM mimarilerinin varsayımlarını zorlayan bir dil yapısı sunmaktadır. Bu nedenle tezin temel amacı, Türkçede daha doğru ve semantik olarak tutarlı temsil

sağlayacak yöntemlerin geliştirilmesi ve bu yöntemlerin model başarısına katkısının sistematik olarak değerlendirilmesidir. Çalışmanın kapsamı yalnızca yeni bir yöntem geliştirmek değil, aynı zamanda bu yöntemi destekleyecek ölçüm araçları, veri kaynakları ve değerlendirme protokollerinin tamamını içeren bütüncül bir araştırma çerçevesi oluşturmaktır.

Tezin ilk amacı, Türkçe için dilbilimsel olarak yapılandırılmış yeni bir tokenizasyon yaklaşımı geliştirmektir. Mevcut alt birim temelli yöntemlerin kök ve ek ayrımını tutarlı biçimde yansıtamaması, hem temsil uzayında hem de modelin kavramsal genelleme kapasitesinde bozulmalara yol açmaktadır. Geliştirilen morfolojik farkındalıklı tokenizasyon yöntemi, ses olayları, ek dizilimleri ve anlam taşıyıcı birimlerin bütünlüğünü dikkate alarak Türkçe metinlerin daha doğru şekilde parçalanmasını hedeflemektedir. Bu yöntem, yalnızca teorik bir öneri olarak değil, aynı zamanda TR-MMLU karşılaştırıcısı üzerinde yapılan deneyler aracılığıyla pratik bir değerlendirme zeminiyle desteklenmektedir.

Tezin ikinci amacı, Türkçe dilinde model başarısını etkileyen temel unsurlardan biri olan veri kalitesinin model eğitim sürecine doğrudan entegre edilmesini sağlamaktır. Veri kaynağı, uzmanlık düzeyi veya içerik güvenilirliği gibi niteliksel farklılıkların öğrenme oranına yansıtılmadığı klasik eğitim stratejileri, özellikle sağlık gibi alanlarda hatalı sonuçların ortaya çıkmasına neden olmaktadır. Bu bağlamda tezde kavramsal olarak önerilen veri kalitesi odaklı adaptif öğrenme oranı yaklaşımı, veri kaynağındaki uzmanlık düzeyi gibi meta-bilgilerin eğitim sürecinde dikkate alınabileceği bir perspektif sunmaktadır. Bu yaklaşım, özellikle sağlık gibi hata toleransı düşük alanlarda veri heterojenliğinin model davranışı üzerindeki olası etkilerini tartışmayı amaçlamaktadır.

Tezin üçüncü amacı, Türkçede tıp gibi bilgi yoğun alanlar için büyük dil modellerinin etkili biçimde uyarlanabileceği yöntemleri değerlendirmektir. LoRA gibi parametre verimli ince ayar yöntemleri (Parameter-Efficient Fine-Tuning, PEFT), büyük ölçekli modellerin alan-özel ince ayarını düşük maliyetle mümkün kılarken; Slerp gibi ağırlık uzayı birleştirme stratejileri, uzmanlık kazanımı ile genel yetenek arasındaki dengeyi yeniden ayarlamak için tamamlayıcı bir araç sunmaktadır. Ancak, bu stratejilerin Türkçe bağlamındaki etkileri sistematik olarak karşılaştırılmamış ve alan-özel görevlerdeki rolleri yeterince incelenmemiştir. Bu çalışma, uzmanlık gerektiren veri kümeleri üzerinde yapılan kapsamlı deneylerle bu yöntemlerin Türkçe model geliştirme süreçlerindeki etkinliğini ortaya koymayı hedeflemektedir. Böylece Türkçe için yüksek doğruluk sağlayan, modüler ve yeniden kullanılabilir alan modelleri geliştirmek mümkün hale gelmektedir.

Tezin dördüncü amacı, Türkçe'ye özgü değerlendirme standartlarının geliştirilmesi ve bu standartların Türkçe modeller arasındaki farkları ne ölçüde görünür kıldığının değerlendirilmesidir. TR-MMLU ve TR-MTEB gibi yerel kaynaklar, yalnızca model karşılaştırması için değil, aynı zamanda kültürel ve kavramsal uyumun ölçülebileceği güvenilir bir değerlendirme zemini sunmayı hedefler. Böylece Türkçe'deki performans farklılıklarının daha görünür ve ölçülebilir hale gelmesi amaçlanmaktadır.

Bu amaçlar doğrultusunda tez, literatürde karşılaşılan boşluklara dayanarak birkaç temel hipotez ortaya koymaktadır. **H1** morfolojik farkındalıklı bir tokenizasyon işleminin, Türkçe gibi morfolojik açıdan zengin dillerde daha tutarlı ve semantik olarak daha güçlü temsiller sağlayacağını öne sürer. **H2** tokenizasyon kalitesinin yalnızca model başarımını artırmakla kalmayıp, %TR ve %Pure gibi metriklerle ölçülebilen öngörücü bir gösterge olduğunu savunur. **H3** verinin kalitesinin öğrenme sürecine doğrudan entegre edilmesinin, özellikle uzmanlık gerektiren alanlarda aynı koşullardaki taban modele göre daha hızlı yakınsama ve karşılaştırmalı performans farkı üretebileceğini belirtir. **H4** PEFT temelli ince ayar ve ağırlık uzayı birleştirme gibi uyarlama stratejilerinin, doğru temsil ve veri altyapısıyla birlikte kullanıldığında uzmanlık kazanımı sağlarken genel yetenek kaybını sınırlayabileceğini varsayar. **H5** ise yerel ve kültürel olarak özgün karşılaştırıcıların, Türkçe modeller arasındaki performans farklarını daha görünür ve yorumlanabilir kılacağını ileri sürer.[1]

Bu hipotezler, tez boyunca "tanım $\rightarrow$ yöntem $\rightarrow$ deney $\rightarrow$ nihai değerlendirme" zinciriyle işlenmiştir. Başka bir ifadeyle, her hipotez yalnızca kavramsal bir önerme olarak bırakılmamış; ilgili yöntem bölümleriyle operasyonelleştirilmiş, deneysel protokollerle sınanmış ve sonuç bölümünde "desteklendi/kısmen desteklendi" düzeyinde açık bir karar cümlesiyle raporlanmıştır. Bu yapı, tezin genel mantığını şeffaflaştırır: hangi iddianın hangi veriyle, hangi metrikle ve hangi sınırlılıklar altında değerlendirildiği takip edilebilir hale gelir.

Hipotezlerin deneysel değerlendirmesi Bölüm 6'te (Tablo 6.13), nihai yorumu ise Bölüm 8.3'de sunulmuştur. Genel sonuç olarak H1, H2 ve H4 deneysel olarak güçlü biçimde desteklenmiş; H3 ve H5 ise kısmen desteklenmiştir.

[1] Hipotezlerin deneysel sonuçları Bölüm 6 içinde raporlanmıştır: H1-H2 için Bölüm 6.4 ve Bölüm 6.5, H3 için Bölüm 6.6, H4 için Bölüm 6.8, H5 için Bölüm 6.3. Hipotez bazlı toplu durum özeti için Bölüm 6.9, nihai yorum için Bölüm 8.3.

1.4 Tezin Özgün Katkıları

Bu tez, büyük dil modellerinin Türkçe gibi morfolojik açıdan zengin ve düşük kaynaklı dillerde karşılaştığı yapısal sınırlılıkları bütüncül bir yaklaşımla ele alması bakımından literatürde özgün bir konuma sahiptir. Dil modelleme, veri kalitesi ve alan uyarlaması gibi birbirinden bağımsız araştırma eksenlerini tek bir metodolojik çerçevede birleştiren bu çalışma, Türkçe doğal dil işlemenin hem teorik hem de uygulamalı yönlerine kapsamlı bir katkı sunmaktadır. Tezin özgün değeri, yalnızca yeni yöntemler önermesinden değil, aynı zamanda Türkçeye özgü kaynakların oluşturulması, değerlendirme protokollerinin geliştirilmesi ve tüm bulguların açık bilim anlayışıyla erişime sunulmasından kaynaklanmaktadır. Aşağıdaki paragraflarda bu katkıların bütüncül resmi detaylandırılmaktadır.

Literatürdeki çoğu çalışmanın aksine, tokenizasyon etkisi önceden eğitilmiş modellerin veri ve optimizasyon geçmişinden bağımsız olarak, bu tez için özel olarak kurgulanan kontrollü bir laboratuvar ortamında test edilmiştir. Sabit mimari, sabit veri ve aynı başlangıç ağırlıkları altında yalnızca tokenizer değişkeninin ayrıştırıldığı bu deney hattında, MFT yalnızca tek bir altbirim tabanlı tokenizera karşı değil; üç farklı Türkçe BPE tabanlı tokenizera (Tabi, Cosmos ve Mursit) karşı da karşılaştırılmıştır. Böylece gözlenen farkların belirli bir tokenizer seçimine özgü olup olmadığı doğrudan sınanmış ve sonuçların daha genellenebilir biçimde yorumlanması hedeflenmiştir.

Tezin en önemli katkılarından biri, Türkçe için dilbilimsel açıdan yapılandırılmış yeni bir tokenizasyon yaklaşımı geliştirilmiş olmasıdır. Mevcut alt birim temelli yöntemlerin aksine, önerilen morfolojik farkındalıklı tokenizasyon yaklaşımı Türkçe kelimelerdeki kök–ek ayrımını, ses olaylarını ve semantik ilişkileri dikkate alarak daha bütünlüklü bir temsil uzayı oluşturmaktadır. Bu çalışma, tokenizasyon kalitesini ölçmek amacıyla geliştirilen %TR ve %Pure gibi metriklerle literatüre yeni değerlendirme araçları kazandırmış; ayrıca, geliştirilen yöntemin yalnızca teorik düzeyde değil, Türkçe MMLU üzerinden yapılan deneylerle pratik olarak da üstünlük gösterdiğini ortaya koymuştur. Bu yönüyle tez, tokenizasyonun dil modellerinin genel performansı üzerindeki belirleyici rolünü nicel olarak ortaya koyarak alana önemli bir katkı sunmaktadır.

Tezin bir diğer özgün katkısı, Türkçe doğal dil işlemede uzun süredir eksikliği hissedilen büyük ölçekli ve yüksek kaliteli değerlendirme veri setlerinin geliştirilmiş olmasıdır. Türkiye'nin ulusal sınav sistemlerinden derlenen yaklaşık 280.000 sorudan seçilen 62 farklı kategoriye ait 6.200 çoktan seçmeli sorudan oluşan TR-MMLU karşılaştırıcısı, Türkçe için ilk kez bu kapsamda oluşturulmuş özgün bir değerlendirme çerçevesidir. TR-MMLU, dilsel çeşitlilik, kavramsal

derinlik ve kültürel bağlam açısından tamamen yerli kaynaklara dayanmakta ve Türkçenin özgün anlam dünyasını yansıtan tutarlı bir ölçme zemini sunmaktadır. Ayrıca, tüm modellerin değerlendirmeleri, ayrıntılı sonuçları ve liderlik tabloları açık kaynak biçimde yayımlanarak araştırmacıların Türkçe LLM performanslarını şeffaf bir biçimde karşılaştırabilmesi sağlanmıştır. Bu çalışma, Türkçe NLP araştırmalarında eksik olan metodolojik standartlaşmanın önünü açan önemli bir adımdır.

Tez aynı zamanda, Türkçe için daha önce bu ölçekte sunulmamış alan-özel veri kaynaklarının oluşturulması sayesinde, tıp gibi uzmanlık gerektiren alanlarda büyük dil modellerinin güvenilir biçimde eğitilebilmesine katkı sağlamaktadır. Gerçek hasta–doktor etkileşimlerinden derlenen 167.000'in üzerindeki tıbbi soru–cevap veri seti, alan modellerinin eğitimi için kritik öneme sahiptir. Bu veri seti, yalnızca büyüklüğü açısından değil, aynı zamanda kaynağı bilinir ve bağlamı korunur olması nedeniyle literatürde benzeri az bulunan bir yapı sunmaktadır. Bu sayede Türkçe için güvenilir uzmanlık modelleri geliştirmek mümkün hale gelmiş ve özellikle sağlık alanında DoctorLlama ve DoctorGemma gibi alan-özel modeller başarıyla eğitilmiştir.

Tezde kavramsal olarak önerilen veri kalitesi odaklı adaptif öğrenme oranı yaklaşımı, eğitim verisindeki kaynak güvenilirliği, uzmanlık düzeyi ve metin tutarlılığı gibi sinyallerin öğrenme sürecinde dikkate alınabileceğini tartışan bir çerçeve sunmaktadır. Bu yaklaşım, özellikle sağlık gibi hata toleransı düşük alanlarda veri heterojenliğinin farkındalığını artırmayı hedefler; ancak henüz kontrollü deneysel doğrulamaya tabi tutulmamış olup, gelecek çalışmalarda ampirik olarak sınanması gerekmektedir.

Tezin bir diğer özgün katkısı, Türkçe sağlık alanında alan-özel uyarlama stratejilerinin sistemli biçimde karşılaştırılmasıdır. LoRA gibi PEFT temelli ince ayar yöntemleri ile Slerp gibi ağırlık uzayı birleştirme stratejileri literatürde yaygın olarak kullanılmakla birlikte, Türkçe alan modelleri üzerinde kapsayıcı bir analize konu edilmemiştir. Bu tez, farklı uyarlama stratejilerinin Türkçe uzmanlık görevlerindeki etkisini deneysel olarak ortaya koymakta; alan-özel modellerin daha düşük maliyetle ve daha yüksek doğrulukla nasıl uyarlanabileceğine ilişkin metodolojik bir yol haritası sunmaktadır. Bu yönüyle çalışma hem uygulamaya dönük hem de kuramsal düzeyde değerlidir.

Son olarak, bu tezin açık bilim ilkelerine dayanan yaklaşımı da önemli bir özgün katkı olarak değerlendirilmelidir. Tüm veri setlerinin, modellerin, değerlendirme betiklerinin ve sonuçların açık olarak yayımlanması, Türkçe NLP

alanında uzun zamandır eksikliği hissedilen şeffaflık ve tekrarlanabilirlik kültürünü güçlendirmektedir. Böylece hem akademik hem de endüstriyel araştırmacıların bu çalışmaları temel alarak yeni modeller geliştirmesi ve alanı daha ileri taşıması mümkün hale gelmektedir.

Genel olarak değerlendirildiğinde, bu tez Türkçe NLP'nin temel bileşenleri olan tokenizasyon, veri kalitesi, değerlendirme çerçeveleri ve model adaptasyonu alanlarında birbirini tamamlayan katkılar sunarak literatürde ilk kez bu ölçek ve bütünsellikte bir çerçeve ortaya koymaktadır. Bu yönüyle çalışma, yalnızca Türkçe için değil, benzer dilsel özellikler taşıyan diğer düşük kaynaklı diller için de genellenebilir nitelikte bir araştırma yaklaşımı sağlamaktadır.

1.5 Tezin Organizasyonu

Tez, kuramsal arka plandan başlayıp veri kaynakları, yöntemler, deneyler ve gösterim/deneyim ortamlarına uzanan bütüncül bir akış izleyecek şekilde kurgulanmıştır. Bölüm 2 temel kavramları (Transformer, tokenizasyon, PEFT ve değerlendirme metrikleri) özetler. Bölüm 3 ilgili çalışmaları, Türkçe ve düşük kaynaklı diller bağlamında tokenizasyon, veri kalitesi, karşılaştırıcılar ve model uyarlaması eksenlerinde konumlandırır. Bölüm 4 tez kapsamında geliştirilen/verilen veri setlerini (TR-MMLU, sağlık ve yardımcı kültürel/duygusal kaynaklar) ayrıntılandırır. Bölüm 5 MFT, DQ-ALR ve LoRA/Slerp tabanlı uyarlama dahil olmak üzere yöntemsel çerçeveyi sunar. Bölüm 6 TR-MMLU, tokenizasyon, veri kalitesi stratejileri ve embedding değerlendirmeleri üzerinden deneysel bulguları raporlar. Bölüm 7 geliştirilen gösterim/deneyim ortamlarını ve açık kaynak paylaşımını özetler. Son olarak Bölüm 8 bulguları özetler ve gelecekteki araştırma yönlerini tartışır.

1.6 Bölüm Özeti

Bu bölüm, tezin motivasyonunu, araştırma problemini, amaç ve hipotezlerini ortaya koymuş; Türkçe için tokenizasyon, veri kalitesi ve alan uyarlamasını bütüncül bir çerçevede konumlandırmıştır. Ayrıca özgün katkılar ve bölüm akışı netleştirilerek, izleyen bölümlerin hangi ihtiyaçlara yanıt verdiği açıklanmıştır.

Ek olarak bölüm, Türkçenin morfolojik yapısının temsil öğrenmeye etkisini, veri güvenilirliğinin model eğitiminde neden kritik olduğunu ve alan-özel uzmanlaşma ihtiyacının hangi pratik sorunlardan doğduğunu tartışmıştır. Böylece tezin problem alanı, ölçülebilir hedefler ve doğrulanabilir hipotezler etrafında somutlaştırılmıştır.

2
TEMEL KONSEPTLER VE TANIMLAR

Bu bölümde, tezin temelini oluşturan NLP teknolojilerinin tarihsel gelişimi, modern dil modellerinin mimari yapıları ve tez kapsamında kullanılan temel yöntemler ele alınmaktadır. Özellikle LLM'ler, tokenizasyon stratejileri, parametre verimli ince ayar yöntemleri ve değerlendirme metrikleri, metodolojinin dayandığı kuramsal çerçeveyi oluşturmaktadır. Amaç, sonraki bölümlerde sunulan yöntemlerin dayandığı kavramları ortak bir dilde tanımlamak ve terimler arası ilişkileri netleştirmektir.

Bu bağlamda bölüm, model mimarilerinin hangi problemleri çözdüğünü, tokenizasyonun verimlilik ve temsil kalitesi üzerindeki etkisini ve parametre-verimli uyarlamanın neden kritik olduğunu sistematik bir bütünlük içinde açıklar. Böylece yöntemsel kararların yalnızca mühendislik tercihi değil, kuramsal gerekçelere dayanan tasarımlar olduğu gösterilir.

2.1 NLP ve Transformer Temelleri

NLP, bilgisayarların insan dilini anlama, yorumlama ve üretme yeteneği kazanmasını amaçlayan ve gelişimi süresince farklı paradigmalardan geçen bir alandır. Alanın ilk dönemlerinde kurallara dayalı sistemler hakimken, dilin karmaşıklığı karşısında bu yaklaşımlar yetersiz kalmıştır. İlerleyen yıllarda büyük veri kümelerinden öğrenen istatistiksel modeller geliştirilmiş olsa da, bağlamsal ilişkileri yakalamakta sınırlı kalmışlardır. Derin öğrenmenin yükselişiyle birlikte, kelimelerin vektör uzayında temsil edilmesi (word embeddings) alanda bir devrim yaratmıştır. Word2Vec ve Global Vectors for Word Representation (GloVe) gibi yöntemler, kelimelerin anlamsal ilişkilerini yoğun vektörlerle ifade etmeyi başarmış ve kelimeler arası anlamsal benzerliklerin matematiksel işlemlerle modellenmesine olanak tanımıştır.

Vektör temsillerinin başarısının ardından, dilin sıralı yapısını modellemek amacıyla Tekrarlayan Sinir Ağları (Recurrent Neural Network, RNN) ve Uzun Kısa Süreli

Bellek Ağları (Long Short-Term Memory, LSTM) kullanılmıştır. Ancak, bu yapılar metni soldan sağa veya sağdan sola sıralı işlediği için uzun mesafeli bağımlılıkları öğrenmede ve paralel hesaplamada sınırlılıklara sahiptir. Bu kısıtlamalar, 2017 yılında Vaswani ve arkadaşları tarafından önerilen Transformer mimarisi ve onun çekirdeğini oluşturan "Dikkat" (Attention) mekanizması ile aşılmıştır. Öz-dikkat (Self-Attention), bir cümledeki her kelimenin diğer tüm kelimelerle olan ilişkisini aynı anda hesaplayarak bağlamsal bilginin çok daha etkili bir şekilde öğrenilmesini sağlar. Matematiksel olarak, bir girdi dizisi Sorgu (Q), Anahtar (K) ve Değer (V) matrislerine dönüştürülür ve dikkat skorları hesaplanır:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.1}$$

Burada d_k anahtar vektörlerinin boyutudur. Formül 2.1, dizideki her tokenin diğer tüm tokenlerle ne kadar ilişkili olduğunu belirleyen ağırlıklı bir toplam üretir. Çok Başlı Dikkat (Multi-Head Attention) yapısı ise bu işlemi paralel olarak farklı alt uzaylarda gerçekleştirerek modelin farklı dilsel özellikleri (sözdizimi, anlamsal ilişkiler vb.) aynı anda öğrenmesine olanak tanır. Transformer mimarisi, farklı görevler için özelleşmiş alt yapılara ayrılarak gelişmiştir. BERT [28] gibi sadece kodlayıcı tabanlı modeller, metni her iki yönden okuyarak derinlemesine bağlam bilgisi çıkarırken, sınıflandırma ve varlık isimlendirme gibi görevlerde üstün başarı göstermektedir. Buna karşın, GPT ailesi [29] ve LLaMA gibi sadece kodçözücü modeller, bir sonraki kelimeyi tahmin etme prensibiyle çalışmakta olup metin üretimi ve genelleme yetenekleriyle öne çıkmaktadır. Tezde kullanılan DoctorLlama ve DoctorGemma modelleri de bu mimariye dayanmaktadır. Ayrıca, T5 ve Bidirectional and Auto-Regressive Transformers (BART) gibi kodlayıcı-kodçözücü modeller, metinden metne dönüşüm gerektiren çeviri ve özetleme görevleri için tasarlanmıştır.

2.2 Tokenizasyon ve Morfolojik Zorluklar

Büyük dil modellerinin başarımını, işlem maliyetini ve çok dilli yeteneklerini doğrudan etkileyen en kritik bileşenlerden biri tokenizasyondur. Ham metnin modelin işleyebileceği en küçük birimlere bölünmesi işlemi olan tokenizasyon, modern modellerde genellikle kelime dağarcığı boyutu (V) ile dizi uzunluğu (N) arasındaki dengeyi (trade-off) yönetmek amacıyla alt kelime tabanlı algoritmalarla gerçekleştirilir. Eğer V çok küçük seçilirse (örneğin karakter düzeyinde), metin çok uzun dizilere (N) bölünür ve dikkat mekanizmasının $O(N^2)$ olan hesaplama maliyeti artar. Tersi durumda, V çok büyük seçilirse, nadir kelimeler için embedding parametre sayısı ($V \times d_{model}$) aşırı büyür ve seyrek veri problemi

ortaya çıkar. Modern LLM'lerde yaygın olarak kullanılan BPE ve SentencePiece gibi algoritmalar bu dengeyi istatistiksel sıklığa göre kurmaya çalışır. Ancak, Türkçe gibi sondan eklemeli dillerde tek bir kelime köküne çok sayıda ek getirilerek karmaşık anlamlar türetilebilmesi, bu algoritmalar için önemli bir sorun teşkil etmektedir. Standart BPE algoritmaları, morfolojik yapıdan bağımsız olarak işlem yaptığı için kelimeleri "ev-ler-i-miz" gibi anlamlı parçalar yerine "evl-erim-iz" gibi anlamsız parçalara bölebilmektedir. Bu durum, modelin bağlamı anlamasını zorlaştırmakta ve "token verimsizliğine" yol açmaktadır. Tez kapsamında geliştirilen MFT, bu sorunu çözmek amacıyla morfolojik analizi istatistiksel yöntemlerle birleştirerek daha tutarlı ve verimli bir tokenizasyon yapısı sunmaktadır. Ek olarak MFT, çözücü (decoder) bileşeniyle tersinir tokenize-detokenize çevrimi tanımladığı için yalnızca bir tokenizer değil, pratikte bir Türkçe metin çözümleyici olarak da değerlendirilebilir.

2.3 Parametre Verimli Alan Uyarlaması

Milyarlarca parametreye sahip LLM'lerin belirli bir alana (alan uyarlaması) uyarlanması, tam ince ayar yönteminin yüksek hesaplama ve bellek maliyeti nedeniyle zorludur. Bu sorunu aşmak için PEFT yöntemleri geliştirilmiştir. Bu yöntemler, önceden eğitilmiş modelin tüm ağırlıklarını dondurup, sadece küçük bir ek parametre kümesini güncelleyerek benzer performansa ulaşmayı hedefler. Tezde etkin bir şekilde kullanılan **LoRA** [30], modelin ağırlık matrislerine (W) düşük rütbeli güncelleme matrisleri (B ve A) ekleyerek çalışır:

$$W_{\text{new}} = W_0 + \Delta W = W_0 + BA \tag{2.2}$$

Burada $W_0 \in \mathbb{R}^{d\times d}$ dondurulmuş ağırlık matrisi, $B \in \mathbb{R}^{d\times r}$ ve $A \in \mathbb{R}^{r\times d}$ ise eğitilebilir düşük rütbeli matrislerdir ($r \ll d$). Formül 2.2 ile verilen bu yaklaşım, eğitilen parametre sayısını binlerce kat azaltarak, büyük modellerin tek bir grafik işlem birimi (Graphics Processing Unit, GPU) üzerinde bile tıp gibi alanlara özelleştirilmesini mümkün kılar. Ayrıca, farklı yeteneklere sahip modellerin ağırlıklarının birleştirilmesi (model merging) yeniden eğitim gerektirmeden performans artışı sağlayabilmektedir. Geleneksel ağırlık ortalaması yerine kullanılan **Slerp** tekniği, yüksek boyutlu vektör uzayında modellerin geometrik yapısını koruyarak daha yumuşak ve tutarlı bir geçiş sağlar. İki ağırlık vektörü v_1 ve v_2 arasındaki interpolasyon şu şekilde ifade edilir:

$$\text{Slerp}(v_1, v_2; t) = \frac{\sin((1-t)\Omega)}{\sin\Omega} v_1 + \frac{\sin(t\Omega)}{\sin\Omega} v_2 \tag{2.3}$$

Burada Ω vektörler arasındaki açıyı, t ise interpolasyon katsayısını temsil eder. Formül 2.3 ile tanımlanan bu yöntem, DoctorLlama modellerinin oluşturulmasında kritik bir rol oynamıştır.

2.4 Değerlendirme Çerçevesi

Modellerin başarısını ölçmek için çeşitli metrikler kullanılmaktadır. Makine çevirisi için geliştirilen Bilingual Evaluation Understudy (BLEU) [31], üretilen metin ile referans metin arasındaki n-gram hassasiyetini (precision) ölçerken, özetleme görevlerinde yaygın olan Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [32], referans metindeki bilginin ne kadarının kapsandığını (recall) değerlendirir. Ancak, bu yüzeysel metrikler anlamsal doğruluğu ölçmekte yetersiz kalabildiğinden, günümüzde çoktan seçmeli sorulardan oluşan genel yetenek testleri standart hale gelmiştir. İngilizce için MMLU karşılaştırıcısı yaygın olarak kullanılırken, bu tez çalışmasında Türkçe için kültürel ve dilsel olarak özelleştirilmiş TR-MMLU karşılaştırıcısı geliştirilmiş ve modellerin yetenekleri bu set üzerinde sınanmıştır. Transformer mimarisi, tokenizasyon dengesi, LoRA gibi verimli adaptasyon teknikleri ve doğru değerlendirme yöntemleri, sonraki bölümlerde detaylandırılacak olan metodolojinin temel taşlarını oluşturmaktadır.

3
İLGİLİ ÇALIŞMALAR

Yapay zeka ve doğal dil işleme alanında son on yıl, ölçeklenme, çok dillilik ve biçimbilimsel farkındalık gibi temel kavramların yeniden tanımlandığı bir dönüşüm sürecine sahne olmuştur. Erken dönem istatistiksel yöntemlerden derin öğrenme tabanlı modellere, oradan da LLM'lere uzanan bu süreç, yalnızca teknik ilerlemeler değil, aynı zamanda veri kalitesi, etik sorumluluk ve dilsel çeşitlilik gibi boyutlarda da köklü değişiklikler getirmiştir. Türkçe gibi biçimbilimsel olarak zengin ve düşük kaynaklı diller, bu dönüşümde özel bir öneme sahiptir; çünkü bu diller, dil modellerinin yapısal genelleme kapasitesini, tokenizasyon stratejilerinin sınırlarını ve çok dilli modellerin adaletini test eden benzersiz birer örnek sunmaktadır.

Bu bölüm, Türkçe ve düşük kaynaklı diller için doğal dil işleme araştırmalarının gelişimini bütüncül bir bakış açısıyla ele aldı; biçimbilim, tokenizasyon, veri kümeleri, karşılaştırıcılar, değerlendirme metrikleri, temel model mimarileri, alan uyarlamaları, çok modlu sistemler ve güncel eğilimler ekseninde mevcut literatürü inceledi. Her alt bölüm, yalnızca geçmiş çalışmaları özetlemekle kalmayıp, aynı zamanda alanın mevcut yönelimlerini, metodolojik boşluklarını ve gelecekteki araştırma fırsatlarını da tartıştı.

Bu bölümün amacı, Türkçe dil modellerine ilişkin literatürü izole bir alan olarak değil, küresel LLM araştırmalarının parçası olarak değerlendirmek; tokenizasyon, veri kalitesi, etik yönetişim ve alan uyarlaması gibi kavramların birbirini tamamlayan yönlerini ortaya koymaktı. Böylece hem Türkçe NLP'nin güncel konumunu hem de biçimbilimsel olarak zengin dillerin büyük dil modelleri çağındaki rolünü anlamaya yönelik kuramsal ve uygulamalı bir zemin sağlandı.

Bu çerçeve, tezde önerilen yöntemlerin hangi boşluklara yanıt verdiğini ve hangi literatür tartışmalarıyla doğrudan ilişkilendiğini açık biçimde göstermeyi hedefledi. Böylece takip eden bölümlerde sunulan katkılar, yalnızca teknik bir geliştirme değil, alanın evrimine dair tutarlı bir uzantı olarak konumlandırıldı.

3.1 Türkçe ve Düşük Kaynaklı Doğal Dil İşleme

Türkçe, biçimbilimsel olarak zengin ve eklemeli (agglutinatif) yapısı nedeniyle NLP sistemleri açısından hem dilbilimsel hem de hesaplamalı zorluklar içeren bir dildir. Bu durum, LLM'lerin metni işleme, temsil etme ve anlamlandırma biçimlerini doğrudan etkilemektedir. Son yıllarda yapılan çalışmalar, İngilizce merkezli modellerin ötesine geçerek Türkçe gibi düşük kaynaklı ve morfolojik olarak karmaşık dillerde de yüksek başarı elde etmeyi hedeflemiştir. Modern tokenizasyonun temelleri Ramshaw ve Marcus [2] tarafından dönüşüm temelli öğrenme yaklaşımıyla atılmış, bu çalışma metni sözdizimsel birimlere ayırarak günümüzdeki veri güdümlü yöntemlerin öncüsü olmuştur. Türkçenin eklemeli yapısı nedeniyle tokenizasyon konusu özellikle önemlidir. Toraman ve ark. [12], biçimbilimsel tokenizasyon yöntemlerinin Türkçe modellerde geleneksel altbirim (altbirim) yöntemlerine göre daha yüksek performans gösterebildiğini göstermiştir. Kayalı ve ark. [13], Türkçe özetleme görevleri için BERTurk, mT5 ve mBART gibi modeller üzerinde BPE, Boşluk, Unigram ve WordPiece yöntemlerini harmanlayan hibrit bir strateji önermiştir. Erkaya [14], Türkçeye özel altbirim tokenizasyon yaklaşımlarını karşılaştırarak kelime haznesi boyutunun ve korpus büyüklüğünün model performansı üzerindeki etkisini incelemiştir. Pan ve ark. [15] ise Türkçe ve Uygurca gibi eklemeli dillerde sinirsel makine çevirisinde morfolojik kelime segmentasyonunun hem dilbilgisel hem de anlamsal bilgi kaybını azalttığını göstermiştir.

Türkçe büyük dil modellerinin değerlendirilmesi için kapsamlı çerçeveler son yıllarda hızla gelişti. Yüksel ve ark. [16] ise 9 farklı lise dersinden 10.000 sorudan oluşan TurkishMMLU veri setini tanıtmış, özellikle Türk Dili ve Edebiyatı ile Türkiye Cumhuriyeti Tarihi gibi kültürel olarak temsil gücü yüksek konularda çok dilli modellerin zayıf performans gösterdiğini ortaya koymuştur. Doğan ve ark. [17], yedi farklı Türkçe model üzerinde gerçekleştirdikleri deneylerde, doğrudan ince ayar yerine sürekli ön-eğitimin daha etkili olduğunu ve bağlam içi öğrenme başarımının soru-cevap performansıyla doğrudan örtüşmediğini göstermiştir.

Türkçe NLP araştırmalarının ilerlemesi için yüksek kaliteli veri kümeleri de kritik önem taşıdı. Baykara ve ark. [18] tarafından yayımlanan TR-News veri kümesi, Türkçe özetleme görevleri için hazırlanmış büyük ölçekli bir kaynaktır. Ayrıca, topluluk tarafından paylaşılan *Turkish Corpus Small* [33] ve *Combined Turkish Datasets v4* [34] gibi korpus koleksiyonları da araştırmacılara genel amaçlı kaynaklar sundu.

Türkçe NLP'nin alanlara özgü uygulamaları da giderek çeşitlendi. Soygazi ve ark. [23] tarafından hazırlanan THQuAD veri kümesi, Osmanlı tarihi ve Türk-İslam

bilim tarihi temalı okuduğunu anlama veri setlerini içeriyordu. Safaya ve ark. [35] "Mukayese" adını verdikleri çalışmada Türkçe NLP görevleri için çok boyutlu bir değerlendirme çerçevesi tanımlamıştır. Eryiğit ve ark. [27] ITU Turkish NLP Web Service platformunu tanıtarak biçimbilim, sözdizimi ve Named Entity Recognition (NER) gibi katmanlarda web tabanlı bir hizmet sundu. Akin [7] tarafından geliştirilen Zemberek ise Türkçe'nin morfolojik karmaşıklığını çözümleyebilen en yaygın açık kaynak kütüphanelerden biri haline gelmiştir.

Bu tür dil araçları, biçimbilimsel çözümleme, normalizasyon ve yazım varyantlarını ele alma açısından değerli altyapılar sunsa da, Transformer tabanlı modellerde ihtiyaç duyulan **tokenizasyon** problemi farklı bir hedefle tanımlanır: sabit bir sözlük altında metni *morfem sınırlarına mümkün olduğunca hizalı* biçimde tokenlere ayırmak ve gerektiğinde bu token dizisini yüksek doğrulukla tekrar yüzey metne döndürebilmek. Bu tezde önerilen hibrit yaklaşım, MFT ile dilbilimsel çözümlemeyi üretim odaklı bir tokenize-detokenize hattına bağlayarak, yalnızca "çözümleme" değil aynı zamanda **tersinir metin çözümleme** (yakın kayıpsız yeniden kurma + pratik hız) hedefini merkeze alır [6]. Bu ayrım, MFT'yi klasik morfolojik çözümleyicilerden ziyade, model ön-işlemede doğrudan kullanılabilir bir tokenizasyon bileşeni olarak konumlandırır.

Düşük kaynaklı dillere yönelik adaptasyon stratejileri kapsamında Toraman ve ark. [26], sürekli eğitim, yönerge temelli ince ayar ve kelime haznesi genişletme gibi yöntemleri kıyaslamış; sürekli eğitimin dil anlama performansını artırdığını, ancak kelime haznesi genişletmenin sınırlı katkı sağladığını göstermiştir. Köksal ve ark. [36], insan etiketleyicilere gerek duymadan yüksek kaliteli yönerge verisi üreten MURI yöntemini tanıtmış; 200 dilde 2 milyon yönerge–çıktı çiftiyle çok dilli performansı artırmıştır. Kesgin ve ark. [37], yalnızca Türkçe korpuslarla eğitilen cosmosGPT model ailesini tanıtmış, Cengiz ve ark. [38] ise 17 Türkçe karşılaştırıcı veri kümesinin %70'inin kalite standartlarını karşılamadığını belirlemiştir.

Çapraz dilli aktarım (cross-lingual transfer) araştırmaları da Türkçe için önemli katkılar sunmuştur. Zhao ve ark. [39], LLaMA modelinin ön-eğitim verisinin yalnızca yüzde biriyle bile mevcut en iyi transfer modellerine yakın performans sergilediğini göstermiştir. İsmayılzada ve ark. [40], Türkçe ve Fince gibi dillerde biçimbilimsel bileşimsel genelleme (morphological compositional generalization) yeteneklerini incelemiş ve karmaşık morfolojik yapılarda doğruluğun belirgin şekilde azaldığını gözlemlemiştir. Açıkgöz ve ark. [41], İngilizce'den Türkçeye bilgi aktarımı ve yalnızca Türkçe veriyle sıfırdan model eğitimi yaklaşımlarını karşılaştırarak, Türkçe yönerge tabanlı ince ayar (instruction-tuning) veri kümeleriyle yeni bir liderlik tablosu oluşturmuştur.

Değerlendirme ve kalite analizi alanında, Çetin ve ark. [42] 29 açık kaynaklı büyük dil modelini etik, güvenilirlik ve adalet (fairness) kriterleriyle incelemiş ve etik değerlendirmenin dil bağımsız yürütülebileceğini göstermiştir. Yüce ve ark. [43], KAFA-Merge adını verdikleri katman bazlı bir model birleştirme (model merging) yöntemiyle Türkçe versiyonlarda ek ince ayara gerek kalmadan yüksek doğruluk sağlamıştır. Ayrıca, *Turkish MMLU Leaderboard* [4] ve *OpenLLM Turkish Leaderboard v0.2* [44] gibi platformlar, Türkçe modellerin performansını şeffaf biçimde karşılaştırmalı olarak sundu ve araştırma topluluğu için ortak bir değerlendirme zemini oluşturdu.

3.2 Tokenizasyon ve Biçimbilim

Tokenizasyon, doğal dil işlemenin temel adımlarından biri olup ham metin ile dil modelleri arasında bir köprü kurar. Seçilen tokenizasyon stratejisi, modelin performansını, eğitim verimliliğini ve çok dilli genelleme yeteneğini doğrudan etkiler. Tokenizasyonun evrimi, başlangıçta veri sıkıştırma tekniklerinden doğmuş ve daha sonra sinirsel dil modelleme süreçlerinin vazgeçilmez bileşenlerinden biri haline gelmiştir. Gage [8], **BPE** yöntemini veri sıkıştırma amacıyla tanıtmış; bu yöntem, en sık görülen byte veya karakter çiftlerini yinelemeli biçimde birleştirerek açık kelime haznesi (açık sözcük dağarcığı) sorununa zarif bir çözüm sunmuştur. Sennrich ve ark. [9], BPE'yi sinirsel makine çevirisine uyarlamış ve nadir kelimeleri altbirim dizileri olarak temsil etmenin çeviri doğruluğunu artırdığını göstermiştir. WMT15 İngilizce–Almanca ve İngilizce–Rusça çeviri görevlerinde yapılan deneylerde, sözlük tabanlı yaklaşımlara kıyasla sırasıyla 1,1 ve 1,3 BLEU puanı iyileşme elde edilmiştir. Kudo [11], dil bağımsız bir altbirim çerçevesi olan SentencePiece'i tanıtmış; bu sistem, önceden kelimeye bölünmüş veri gerektirmeksizin doğrudan ham cümlelerden öğrenme yeteneğiyle farklılaşmıştır. Kudo ve ark. [45] bu yaklaşımı geliştirerek olasılıksal altbirim düzenlileştirme (*subword regularization*) yöntemi önermiş, farklı segmentasyon örneklemeleriyle model kararlılığını artırmış ve özellikle düşük kaynaklı senaryolarda önemli iyileşmeler sağlamıştır.

Boström ve ark. [46], BPE ile Unigram Dil Modeli (Language Model, LM) tokenizasyonunu sistematik biçimde karşılaştırmış ve Unigram LM'nin biçimbilimsel olarak daha tutarlı altbirimler ürettiğini göstermiştir. Provilkov ve ark. [47], BPE'ye rastgelelik katarak aynı kelimenin birden fazla segmentasyon biçimini üreten BPE-dropout yöntemini tanıtmış; bu teknik 2,3 BLEU'ya kadar performans artışı sağlamıştır. Schmidt ve ark. [48], tokenizasyonun yalnızca bir sıkıştırma problemi olarak görülmesini eleştirerek PathPiece adlı yeni bir

algoritma önermiştir. Çalışma, daha az tokenin her zaman daha yüksek doğruluk getirmediğini ve ön-tokenizasyon (pre-tokenization) sürecinin kritik rolünü vurgulamıştır. Bu yönelim, tokenizasyonun artık salt veri bölütlemesi değil, öğrenme sürecinin bir parçası olarak değerlendirilmesi gerektiğini göstermektedir.

Biçimbilimsel olarak zengin dillerde tokenizasyon, dilin yapısal özelliklerinin dikkatle ele alınmasını gerektirir. Park ve ark. [49], Korece için yaptıkları çalışmada biçimbilimsel segmentasyonla BPE'nin birleşimini önererek hem çeviri hem doğal dil anlama görevlerinde en yüksek başarıyı elde etmiştir. Schuster ve Nakajima [10], Japonca ve Korece gibi kelime sınırlarının belirgin olmadığı dillerde sesli arama uygulamaları için özel çözümler geliştirmiştir. Ahmadi [50], Kürtçenin Sorani ve Kurmançi lehçeleri için biçimbilimsel çözümleyiciye dayalı bir tokenizasyon yaklaşımı sunmuş ve denetimsiz yöntemlere göre çok daha yüksek doğruluk sağlamıştır. Jabbar ve ark. [51], morfolojik segmentasyona dayalı MorphPiece yöntemini tanıtmış; bu yöntem, GPT-tarzı modellerde daha yüksek doğruluk sağlamış ve biçimbilimsel karmaşıklığı yüksek dillerde GPT-2'yi aşan performans elde etmiştir. Hofmann ve ark. [52], türetimsel biçimbilimin (derivational morphology) dil modellerinin anlamsal çözümleme kapasitesini güçlendirdiğini; Diewald ve ark. [53] ise tokenizasyonun hız–doğruluk dengesinin hiçbir araçta tam anlamıyla optimize edilemediğini göstermiştir.

Son dönemde, geleneksel tokenizasyonun ötesine geçen yöntemler geliştirildi. Xue ve ark. [54], tamamen tokenizasyonsuz ByT5 modelini tanıtarak karakter veya byte düzeyinde eğitilen modellerin birçok görevde token tabanlı modellere eşit performans gösterebildiğini kanıtlamıştır. Tay ve ark. [55], Charformer modelinde karakter düzeyinde otomatik altbirim öğrenimi yapmış ve GBST (Gradient-Based Subword Tokenization) yöntemiyle geleneksel yaklaşımlara kıyasla %28–100 oranında hız kazancı sağlamıştır. Videau ve ark. [56], eğitim sırasında kendi tokenlarını öğrenebilen U-Net tabanlı bir yaklaşım önermiştir. Bu çalışmalar, tokenizasyonun artık statik bir ön işlem olmaktan çıkıp dinamik ve öğrenilebilir bir bileşen haline geldiğini göstermiştir.

Kelime haznesi oluşturma ve optimizasyon konusundaki araştırmalar, çok dilli modellerde tokenizasyonun adalet ve verimlilik üzerindeki etkisini ortaya koymuştur. Rust ve ark. [57], çok dilli dil modellerinde kelime haznesinin diller arası performans farkına neden olduğunu ve tek dilli tokenizer kullanımının birçok görevde üstünlük sağladığını göstermiştir. Xu ve ark. [58], optimal taşım kuramını kullanarak VOLT (Vocabulary Learning via Optimal Transport) yöntemini geliştirmiş; böylece BLEU skorunu artırırken hesaplama maliyetini 384 GPU saatinden 30 GPU saatine düşürmüştür. Ali ve ark. [59], 24 model

üzerinde yaptıkları deneylerde tokenizer seçiminin hem eğitim maliyetini hem de performansı belirgin biçimde etkilediğini; İngilizce merkezli tokenizasyonun bazı dillerde %68'e kadar maliyet artışına yol açtığını bildirmiştir.

Çok dilli ve çapraz dilli yaklaşımlarda Petrov ve ark. [60], farklı dillerde aynı metnin 15 kata kadar farklı sayıda tokenla temsil edilebildiğini göstererek tokenizasyon kaynaklı dilsel adaletsizliğe dikkat çekmiştir. Thakur ve ark. [61], çok dilli veri bileşim algoritmalarının token–kelime oranını azaltarak performansı artırabileceğini; Rashad ve ark. [62], Arapça için bir değerlendirme platformu geliştirmiştir. Rosa ve ark. [63], İskandinav dilleri için tokenizer karşılaştırıcısı yayımlamış; Koubaa ve ark. [64], Arapça dil modellemesi için Aranizer adlı yeni bir SentencePiece türevini tanıtmıştır. Držík ve ark. [65], Slovakça morfolojisini BPE ile birleştirerek SKMT tokenizersini geliştirmiş ve SlovakBERT'e göre %3,5'lik precision ve recall'ın harmonik ortalaması (Harmonic Mean of Precision and Recall, F1) artışı sağlamıştır. Vasiu ve ark. [66], Romence için İngilizce ve Latin kökenli tokenizatörlerin uyarlanmasının %94 üzeri doğruluk sağladığını, Zhang ve ark. [67] ise kullanıcı üretimli içeriklerde dinamik bağlama uyumlu tokenizasyon stratejilerinin yeni kelimeleri daha iyi yakaladığını göstermiştir.

Düşük kaynaklı dillerde tokenizasyonun özel önemi büyüktür. Dossou ve ark. [68], Fon dili için katılımcı tabanlı WEB yöntemini tanıtmış; bu sistem ton, diakritik ve gramer özelliklerini insan katkısıyla doğru biçimde yakalamıştır. Li ve ark. [69], Tibetçe–Çince çeviri için iki düzeyli altbirim segmentasyonu önererek, karakter tabanlı yaklaşımların düşük kaynaklı ortamlarda daha başarılı olduğunu göstermiştir. Ding ve ark. [70], Birmanca ve Khmerce için NOVA sistemini geliştirerek yalnızca dört etiketle (n, v, a, o) etkili bir tokenizasyon–etiketleme çerçevesi sunmuştur. Bu tür dillerde tokenizasyonun basit ama uyarlanabilir biçimbilimsel temellere dayandırılması, model başarısını anlamlı biçimde artırmaktadır.

Tokenizasyonun değerlendirilmesi ve biçimsel temellendirilmesi üzerine de önemli çalışmalar yapılmıştır. Berglund ve ark. [71], BPE'nin biçimsel analizini yaparak tokenizasyonun sonlu durumlu dönüştürücüler (FST) ile ifade edilebileceğini göstermiştir. Neubeck ve ark. [72], büyük ölçekli iş akışlarında hız ve esneklik kazandıran optimize edilmiş bir byte-pair tokenizer geliştirmiştir. Biçimbilimsel ayrıştırma yönünde Hakkani-Tür ve ark. [73], Türkçede biçimbilimsel belirsizliği giderme üzerine kapsamlı bir inceleme sunmuş; Kulick ve ark. [74] Arapça için kapalı sınıf–açık sınıf ayrımına dayalı bir model önermiştir.

Ayrıca, alan ve görev odaklı yaklaşımlar da gelişti. Rashad ve ark. [75],

Arapça kitap sayfalarını OCR yoluyla Markdown'a dönüştüren Arabic-Nougat sistemini tanıtmış; bu sistem Arapça yazı biçimlerine uygun özel bir tokenizasyon stratejisi geliştirmiştir. Nayak ve ark. [76], BERT'in tokenizasyon stratejisinin alan-özgü terimlerde yetersiz kaldığını göstermiştir. Hiraoka ve ark. [77], tokenizasyonun model parametreleriyle birlikte öğrenildiği "ortak optimizasyon" fikrini öne sürmüş; Zhang ve ark. [78], karakter, altbirim ve kelime düzeyini birleştiren çok ölçekli AMBERT modelini tanıtmıştır. Kaing ve ark. [79] ve Ding ve ark. [80], Khmer ve Birmanca dillerinde morfem düzeyinde tokenizasyonun yeterli olduğunu ancak öbek sınırlarının karmaşıklaştığını vurgulamıştır.

Son olarak, Lin ve ark. [81], tokenizasyonun yalnızca bir sıkıştırma veya ön işleme problemi değil, bilgi kuramı ve öğrenme dinamikleriyle bağlantılı bir temsil sorunu olduğunu belirtmiştir. Bu yaklaşım, gelecekte token sayısı veya kelime haznesi boyutuna değil, dilsel anlamlılık ve bilişsel uygunluk ölçütlerine dayalı yeni metriklerin geliştirilmesini önermiştir. Genel olarak bakıldığında, tokenizasyon alanı hızla evrildi ve birkaç umut verici yön ortaya çıktı: model parametreleriyle birlikte öğrenilen "öğrenilmiş tokenizasyon" yöntemleri, bağlama ve alana göre uyarlanabilir stratejiler, çok ölçekli temsiller ve dilbilimsel uygunluğu önceleyen yeni ölçütler. Tüm bu gelişmeler, tokenizasyonun halen çözülmemiş bir araştırma alanı olduğunu ve dilbilimsel farkındalığın büyük dil modellerinin başarısında belirleyici rol oynadığını açıkça göstermiştir.

3.3 Veri Kümeleri, Korpuslar ve Karşılaştırıcılar

Eğitim verisinin kalitesi ve çeşitliliği, **LLM**'lerin yeteneklerini, sınırlarını ve genelleme gücünü temelden belirler. Doğal dil işlemenin erken dönemlerinde görev odaklı küçük derlemeler kullanılırken, zamanla kapsamlı çok dilli ve çok görevli karşılaştırıcılar geliştirilmiş, böylece modellerin bilgi, muhakeme ve dilsel yetenekleri sistematik biçimde ölçülmeye başlanmıştır. Wang ve ark. [82], General Language Understanding Evaluation (GLUE) karşılaştırıcısını tanıtarak dil modellerinin duygu analizi ve doğal dil çıkarımı gibi görevlerde anlamlı karşılaştırmasını mümkün kılmıştır. Ancak, kısa sürede insan seviyesine ulaşan bu değerlendirme zemini Wang ve ark. [83] tarafından daha zorlu SuperGLUE sürümüyle genişletilmiştir. Rajpurkar ve ark. [84], Wikipedia temelli 100.000'den fazla sorudan oluşan SQuAD veri setini geliştirerek okuduğunu anlama (reading comprehension) alanında bir dönüm noktası oluşturmuştur.

Embedding tabanlı bilgi erişim (getirim) ve semantik arama çalışmalarının yaygınlaşmasıyla birlikte, değerlendirme odağı yalnızca üretken modellerle sınırlı

kalmamış; **cümle gömme** modelleri için de çok görevli değerlendirme çerçeveleri ortaya çıkmıştır. Massive Text Embedding Evaluation (MTEB) [85], farklı görev tiplerini (Semantic Textual Similarity (STS), sınıflandırma, kümeleme, getirim vb.) tek bir çatı altında toplayarak embedding modellerinin karşılaştırılmasına olanak sağlamış; çok dilli senaryolarda ise MMTEB [86] gibi genişletilmiş çerçeveler önerilmiştir. Benzer biçimde, yerel dil-özgü değerlendirme ihtiyacını karşılamak üzere MTEB-French [87], PL-MTEB [88] ve ruMTEB [89] gibi dil odaklı uzantılar geliştirilmiş; Türkçe için de TR-MTEB [90], [91] ile hem değerlendirme altyapısı hem de Türkçe embedding modeli ailesi sistematik olarak raporlanmıştır. Bu tür çerçeveler, embedding modellerinin görevler arası genellenebilirliğini ölçmek açısından kritik bir rol oynamaktadır.

İngilizce merkezli bu değerlendirmelerden sonra, çok dilli ve çapraz dilli değerlendirme setleri hızla önem kazanmıştır. Conneau ve ark. [92] tarafından tanıtılan XNLI, doğal dil çıkarımı görevini 15 dile genişleterek çok dilli aktarımın (transfer) ilk sistematik değerlendirmesini sunmuştur. Ponti ve ark. [93] tarafından geliştirilen XCOPA, nedensel akıl yürütmeyi 11 dilde ölçerken, Liang ve ark. [94] çok görevli ve çok dilli XGLUE değerlendirme setini önermiştir. Longpre ve ark. [95] ise 26 dilde hizalanmış MKQA veri setini geliştirerek bilgi temelli soru-cevap değerlendirmelerinde dil bağımsız bir çerçeve oluşturmuştur. Bu veri kümeleri, çok dilli modellerin özellikle düşük kaynaklı dillerde ciddi performans kayıpları yaşadığını göstermiştir.

Genel dil anlama görevlerinden sonra, daha karmaşık akıl yürütme karşılaştırıcıları geliştirilmeye başlanmıştır. Hendrycks ve ark. [96], 57 disiplini kapsayan MMLU karşılaştırıcısı ile geniş kapsamlı bir değerlendirme standardı oluşturmuş, daha sonra Wang ve ark. [97] bu yapıyı MMLU-Pro sürümüyle geliştirerek soru zorluğunu artırmıştır.

Akıl yürütme ve mantıksal problem çözme alanında, Cobbe ve ark. [98] GSM8K veri setini; Hendrycks ve ark. [99] MATH veri setini; Mirzadeh ve ark. [100] ise sembolik varyantı GSM-Symbolic'i geliştirmiştir. Bu veri kümeleri, dil modellerinin gerçek akıl yürütme yeteneği yerine ezberlenmiş kalıplara dayandığını ortaya koymuştur. Clark ve ark. [101], ARC veri setiyle ilkokul düzeyi bilimsel çıkarım yeteneğini; Zellers ve ark. [102], HellaSwag ile sağduyu temelli muhakemeyi; Sakaguchi ve ark. [103], WinoGrande ile ortak mantık (coreference) çözümünü; Suzgun ve ark. [104] ise BIG-Bench Hard ile zincirleme akıl yürütmeyi ölçmüştür.

Gerçeklik ve güvenilirlik açısından Lin ve ark. [105], modellerin doğru yanıt

verme becerilerini ölçen TruthfulQA değerlendirme setini geliştirmiş, GPT-3'ün %58 doğrulukta kalması modellerin yanlış inançları da öğrenebildiğini göstermiştir. Ethayarajh ve ark. [106] ise yalnızca doğruluk odaklı liderlik tablolarının eksik olduğunu, enerji verimliliği ve yanıt gecikmesi gibi faktörlerin de değerlendirilmesi gerektiğini vurgulamıştır. Alan-özgü veri setleri de çeşitlenmiş; Sun ve ark. [107], ilaç güvenliği olaylarını içeren PHEE'yi; Fan ve ark. [108] ise doktor–hasta etkileşimini simüle eden MVME değerlendirme setini yayımlamıştır.

Yönerge takibi ve diyalog kalitesine yönelik değerlendirmelerde Zhou ve ark. [109] tarafından IFEval, Chiang ve ark. [110] tarafından insan tercihlerine dayalı Chatbot Arena tanıtılmıştır. Kod ve programlama dillerinde Xu ve ark. [111], farklı modellerin dil-özgü stratejilerle önemli performans farkları gösterdiğini bulmuştur. Çok modlu değerlendirmelerde Bandarkar ve ark. [112], 122 dil varyantını kapsayan Belebele veri setini, Yue ve ark. [113] ise 39 dilde çok modlu PANGEA veri kümesini sunmuştur.

Düşük kaynaklı diller için çok dilli korpuslar büyük ilerleme sağlamıştır. Nguyen ve ark. [114], 167 dil ve 6,3 trilyon token içeren CulturaX veri kümesini yayımlamış; Ortiz-Suarez ve ark. [115] büyük ölçekli Common Crawl verisini düşük kaynaklı altyapılarda verimli biçimde işleyebilecek boru hattı geliştirmiştir. Bu devasa veri kaynakları, çok dilli model ön-eğitimi için kritik önem taşırken aynı zamanda veri temizliği ve etik sorunları da gündeme getirmiştir.

Karşılaştırıcı değerlendirmelerine yönelik meta-analizlerde Tedeschi ve ark. [116], modellerin "süperinsan" performans iddialarını eleştirmiş; Verga ve ark. [117], birden fazla küçük modelin oluşturduğu Panel of LLMs yaklaşımının daha tarafsız sonuç verdiğini göstermiştir. Bavaresco ve ark. [118], 20 görev üzerinde 11 modelin insan yargılarıyla tutarlılığının göreve göre değiştiğini bulmuştur. Srivastava ve ark. [119], 204 görevi içeren BIG-Bench veri setini tanıtmış; karşılaştırıcıların zamanla doygunluğa ulaştığını ve sürekli yenilenmeleri gerektiğini göstermiştir.

Mitchell ve ark. [120], model kartlarını önererek veri ve model dokümantasyonunun şeffaflık için gerekliliğini ortaya koymuştur. Bender ve Gebru [121], rastgele web verilerinin etik risklerini vurgulamış ve veri kaynaklarının açık biçimde belgelenmesini savunmuştur. Bu anlayış, veri setlerinin yalnızca teknik değil aynı zamanda etik değerlere dayalı biçimde oluşturulması gerektiğini ortaya koymuştur.

Türkçe NLP bağlamında da çeşitli özgün veri kaynakları geliştirilmiştir. Çöltekin [122], 36.232 tweet içeren Türkçe saldırgan dil korpusunu; Aydoğan ve ark. [24], 150.000 etiketli kullanıcı yorumu içeren TRSAv1 duygu analizi veri setini;

Beken Fikri ve ark. [123] ise İngilizce STSb veri setini Türkçeye çevirerek ilk Türkçe anlamsal benzerlik veri kümesini oluşturmuştur. Dilsel yetenekleri daha kontrollü biçimde sınamak için, Başar ve ark. [124] TurBLiMP karşılaştırıcısını önererek Türkçe için 16 farklı olguyu kapsayan minimal cümle çiftleri üzerinden biçimbilimsel süreçler ve sözdizimsel esneklik gibi olguları ölçmeyi hedeflemiştir. TR-MMLU karşılaştırıcısına ek olarak Bezir [19], MMLU-Pro'nun Türkçe uyarlaması olan MMLU-pro-TR'yi yayımlamış; aynı zamanda Aguilar [125] tarafından geliştirilen DoctorGPT projesi, tıbbi diyalog tabanlı sistemler için açık kaynaklı bir örnek sunmuştur.

Daha geniş bağlamda, çapraz dilli değerlendirme çerçevelerinde Cer ve ark. [126], SemEval-2017 STS karşılaştırıcısını geliştirerek çok dilli anlamsal benzerlik ölçümlerine öncülük etmiş; Hasan ve ark. [127] 44 dili kapsayan XL-Sum veri setini tanıtmıştır. Sprague ve ark. [128], doğal dil anlatılarında çok adımlı yumuşak akıl yürütmeyi değerlendiren MuSR veri setini geliştirirken; Rein ve ark. [129], biyoloji, fizik ve kimya alanlarında 448 "Google-proof" sorudan oluşan Graduate-level Google-proof Q&A (GPQA) veri setini sunmuştur. GPT-4'ün yalnızca %39 doğruluk elde ettiği GPQA, hem insanlar hem modeller için ileri düzey bilgi sınaması sağlamıştır.

Genel eğilimler değerlendirildiğinde, veri kümeleri ve karşılaştırıcılar alanı dört temel gerilim hattı etrafında şekillenmektedir: (i) veri miktarı ve kalite arasındaki ödünleşim, (ii) çok dillilik ve kapsayıcılık eksikliği, (iii) karşılaştırıcı doygunluğu ve (iv) etik sorumluluk. Büyük ölçekli web verileri güçlü modeller üretse de gürültü ve önyargı sorunlarını beraberinde getirirken, CulturaX gibi dikkatle temizlenmiş veri kaynakları yüksek kalite sunmaktadır. Düşük kaynaklı dillerin yeterince temsil edilmemesi, transfer öğrenme ve sentetik veri üretimi gibi alanların önemini artırmıştır. GLUE $\rightarrow$ SuperGLUE, MMLU $\rightarrow$ MMLU-Pro gibi ardışık sürümler, karşılaştırıcıların evrimsel doğasını gösterirken, test setlerine aşırı uyum (overfitting) riski de doğurmuştur. Son olarak, veri belgeleri [120], [121], etik kaynak kullanımı, kullanıcı rızası ve çevresel maliyet gibi konular artık yalnızca akademik değil, aynı zamanda toplumsal sorumluluk bağlamında da merkezi öneme sahiptir.

3.4 Değerlendirme, Metrikler ve Test Süreçleri

Büyük dil modellerinin değerlendirilmesi, geleneksel doğal dil işleme metriklerinin ötesine geçen çok boyutlu bir süreci kapsamaktadır. Bu modellerin yetenekleri yalnızca doğrulukla değil, aynı zamanda güvenilirlik, güvenlik, tarafsızlık ve

insan tercihlerine uyum gibi faktörlerle ölçülmelidir. Bu çerçevede değerlendirme yöntemleri, klasik otomatik metriklerden insan yargısına dayalı değerlendirmelere ve son dönemde ortaya çıkan LLM-tabanlı yargılama (LLM-as-a-Judge) yaklaşımlarına kadar evrilmiştir. Ancak, üretken modellerin doğası, bu metriklerin çoğunun geçerliliğini zorlamakta ve yeni paradigmalara ihtiyaç doğurmaktadır.

Otomatik metrikler uzun süre NLP değerlendirmesinin belkemiğini oluşturmuştur. BLEU [31] ve ROUGE [32] metrikleri, sırasıyla çeviri ve özetleme görevlerinde n-gram örtüşmesi üzerinden başarıyı ölçerek erken dönem ilerlemeleri mümkün kılmıştır. Ancak, bu metrikler, üretken modellerin anlamsal çeşitliliğini yakalayamamakta ve yüzeysel benzerliklere sıkışmaktadır. Beken Fikri ve ark. [123], Türkçe özetleme çalışmalarında ROUGE'un biçimbilimsel olarak zengin diller için yetersiz kaldığını ve anlamsal benzerlik temelli ölçütlerin insan değerlendirmeleriyle çok daha yüksek korelasyon gösterdiğini ortaya koymuştur. Zhang ve ark. [130] tarafından geliştirilen BERTScore, bağlamsal gömmeler üzerinden anlamsal eşdeğerliği ölçerek, 363 sistemin değerlendirmesinde insan yargılarına en yüksek uyumu sağlamıştır. Bu yaklaşım, anlam temsiline dayalı yeni bir değerlendirme paradigmasını başlatmıştır.

Otomatik ölçütlerin ölçeklenebilir olmasına rağmen, insan tabanlı değerlendirmeler halen "altın standart" olmaya devam etmektedir. Ancak, maliyet, süre ve öznel farklılıklar gibi zorluklar nedeniyle alternatif yollar araştırılmıştır. Snow ve ark. [131], crowdsourcing yöntemleriyle uygun yönergeler sağlandığında uzman dışı katılımcıların yüksek doğrulukta etiketleme yapabileceğini göstermiştir. Tedeschi ve ark. [116], insan ve model karşılaştırmalarında kullanılan protokollerin tutarsızlığını vurgulamış ve modellerin "insanüstü" performans iddialarının sıklıkla veri kalıplarını sömürmekten kaynaklandığını göstermiştir. Chiang ve ark. [110] tarafından geliştirilen Chatbot Arena platformu, 240.000 kullanıcı oyu üzerinden kitlesel değerlendirmeyi mümkün kılmış; kullanıcı tercihleri ile uzman değerlendirmeleri arasında yüksek korelasyon saptanmıştır. Bu bulgular, iyi tasarlanmış kalabalık kaynaklı değerlendirmelerin güvenilir sinyaller sağlayabileceğini göstermektedir.

LLM'lerin kendilerinin değerlendirici olarak kullanılması ise yeni bir dönemi başlatmıştır. Bavaresco ve ark. [118], 11 büyük dil modelini 20 görev üzerinde insan yargılarıyla karşılaştırmış ve bazı görevlerde modellerin insanlarla tutarlı, bazılarında ise belirgin sapmalar gösterdiğini bulmuştur. Verga ve ark. [117], tek bir model yerine birden fazla küçük modelden oluşan LLM Paneli (PoLL) yaklaşımını önererek önyargıyı azaltmış ve maliyeti yedi kat düşürmüştür. Aynetdinov ve ark. [132], LLM çıktılarının yönergeye duyarlı

biçimde değerlendirilebilmesi için semantik benzerlik temelli SemScore metriğini geliştirmiş ve insan yargılarıyla en yüksek korelasyonu elde etmiştir. Aldeen ve ark. [133], ChatGPT'nin insan etiketleyicilerle karşılaştırmalı performansını analiz ederek ortalama %78,2 doğruluk elde etmiş; ancak görev türüne göre ciddi farklılıklar gözlemlenmiştir. Bu sonuçlar, LLM'lerin etiketleme ve değerlendirme süreçlerinde yardımcı olabileceğini ancak insan doğrulamasının yerini tam olarak alamayacağını göstermektedir.

Alana özgü değerlendirme çalışmalarında, tıp gibi yüksek riskli alanlarda yalnızca doğruluk değil, güvenlik, klinik geçerlilik ve açıklanabilirlik de ön plana çıkmıştır. Nazi ve ark. [134], mevcut tıbbi karşılaştırıcıların klinik geçerlilik açısından sınırlı olduğunu, Mumtaz ve ark. [135] ise farklı branşların farklı değerlendirme ölçütleri gerektirdiğini vurgulamıştır. Bedi ve ark. [136], 519 tıbbi LLM çalışmasını analiz ederek yalnızca %5'inin gerçek hasta verisiyle test edildiğini göstermiştir. Bu örnekler, sağlık uygulamaları için alan-özgü güvenilirlik standartlarının geliştirilmesinin zorunlu olduğunu ortaya koymaktadır.

Çok dilli değerlendirme de ayrı bir zorluk alanıdır. Zhang ve ark. [137], modellerin diller arası bilgi aktarımında çeviri görevlerinde tutarlı, dil içi özgün görevlerde ise başarısız olduğunu göstermiştir. Wang ve ark. [138], XSafety karşılaştırıcısı ile çok dilli güvenlik değerlendirmesi yapmış ve İngilizce dışındaki dillerde %19,1 oranında daha fazla güvensiz yanıt üretildiğini belirlemiştir. Çetin ve ark. [42], Türkçe ve İngilizce modeller üzerinde etik değerlendirme gerçekleştirerek parametre sayısı arttıkça güvenlik ve adalet metriklerinde iyileşme olduğunu bulmuştur. Mirzadeh ve ark. [100], GSM-Symbolic veri setiyle yapılan deneylerde küçük yapısal değişimlerin bile akıl yürütme doğruluğunu ciddi biçimde düşürdüğünü göstermiştir. Huang ve ark. [139], sekiz boyutta güvenilirlik ölçen TrustLLM çerçevesini geliştirerek mülkiyet ve açık kaynak modellerin güvenilirlik farklarını ortaya koymuştur.

Belirli yeteneklerin değerlendirilmesinde, Zhou ve ark. [109] tarafından geliştirilen IFEval değerlendirme seti, modelin açık talimatları yerine getirme kapasitesini ölçmek için doğrudan doğrulanabilir yönergeler sunmuştur. Linzen ve ark. [140], mevcut değerlendirme paradigmasının modelleri istatistiksel örüntülere yönlendirdiğini, insan benzeri genelleme kabiliyetinin test edilmediğini belirtmiştir. Değerlendirme geçerliliği konusundaki eleştirilerde Ethayarajh ve ark. [106], yalnızca doğruluk ölçütlerine dayalı sıralamaların enerji tüketimi ve gecikme gibi pratik metrikleri göz ardı ettiğini vurgulamıştır. Gerçek uygulama koşullarında, Henderson ve ark. [141], enerji tüketimi ve karbon ayak izini raporlayan çevresel etki çerçevesi önermiştir. Bedi ve ark. [142], tıbbi değerlendirmelerde toksisite ve

tarafsızlık boyutlarının göz ardı edildiğini, daha çok boyutlu çerçevelerin gerekli olduğunu ortaya koymuştur.

Yeni değerlendirme paradigmaları, etik ve insani faktörleri giderek daha fazla merkeze almaktadır. Baeza-Yates [143], etik değerlendirmelerde ayrımcılık, anlamsal eksiklik, bilişsel önyargılar ve kültürel farklılıklar olmak üzere sekiz temel zorluk tanımlamış ve teknik ilerlemenin etik uyumla birlikte yürütülmesi gerektiğini belirtmiştir. Rein ve ark. [129], insan denetimi altında "ölçeklenebilir gözetim" yaklaşımını GPQA veri setiyle deneysel olarak göstermiştir. Bu tür çalışmalar, insan kontrolüyle model denetiminin birlikte kullanılabileceği yeni değerlendirme çerçevelerine işaret etmektedir.

3.5 Temel Modeller ve Mimariler

Büyük dil modellerinin mimarisi, eğitim yöntemleri ve optimizasyon stratejileri; modellerin yeteneklerini, ölçeklenebilirliğini ve verimliliğini belirleyen en temel faktörlerdir. Bu evrim, 2017 yılında Vaswani ve ark. [144] tarafından tanıtılan Transformer mimarisiyle başlamıştır. Transformer, yalnızca dikkat (attention) mekanizmalarına dayanarak tekrarlamalı (recurrent) veya evrişimsel (convolutional) bileşenleri ortadan kaldırmış ve dizilerdeki uzun menzilli bağımlılıkları paralel biçimde modellemeyi mümkün kılmıştır. Çok başlı öz-dikkat (multi-head self-attention) yapısı, aynı girdi dizisinin farklı temsil altuzaylarını eşzamanlı değerlendirebilmiş, bu da hem hesaplama verimliliğini hem de bağlam farkındalığını artırmıştır. Mimarinin başarısı, günümüzdeki neredeyse tüm büyük dil modellerinin temeli haline gelmiştir. Khatri ve ark. [145], dikkat mekanizmalarını kategorik (category-theoretic) bir biçimsel çerçevede analiz ederek farklı dikkat türlerinin yapısal rollerini açıklamıştır. Dikkat mekanizmasının karesel karmaşıklığı, uzun dizilerde hesaplama maliyetini artırdığı için, alternatif dikkat biçimleri geliştirilmiştir. Locatello ve ark. [146] tarafından önerilen Slot Attention, görev bağımlı soyut temsil birimleri ("slot"lar) öğrenerek uzun bağlamların nesne temelli çözümlemesine imkan tanımış ve dil modelleri için ilham kaynağı olmuştur.

Transformerlardan türeyen ilk büyük ilerleme, yalnızca kodlayıcı (encoder-only) modellerde gözlemlenmiştir. Devlin ve ark. [28], **BERT** modelini tanıtmış ve iki yönlü bağlamdan öğrenmeyi mümkün kılan maskelenmiş dil modelleme (MLM) tekniğini geliştirmiştir. BERT, sol ve sağ bağlamı aynı anda koşullandırarak çok daha zengin temsiller üretmiş, GLUE, MultiNLI ve SQuAD 2.0 gibi görevlerde dönemin en yüksek başarısını elde etmiştir. Liu ve ark. [147], BERT'in eksik

eğitildiğini göstererek, RoBERTa modelini geliştirmiş ve cümle tahmini (NSP) görevini kaldırarak eğitim stratejisinin performans üzerindeki belirleyici etkisini vurgulamıştır. BERT tabanlı modeller alanlara göre özelleşerek farklı türevlerin doğmasına yol açmıştır; örneğin Huang ve ark. [20] klinik verilerle eğitilen ClinicalBERT'i tanıtmıştır.

Bu dönemin paralelinde yalnızca çözücü (decoder-only) mimariler de gelişmiştir. Radford ve ark. [29], büyük metin koleksiyonlarında denetimsiz biçimde eğitilen GPT-2 modelini tanıtarak, sıfır-atış (zero-shot) görevlerde dahi yüksek performansın mümkün olduğunu göstermiştir. Bu çalışma, ölçeklenmenin (scaling) model başarısındaki merkezi rolünü açık biçimde ortaya koymuştur. OpenAI [148] tarafından tanıtılan GPT-4, metin ve görsel girdileri aynı anda işleyebilen çoklu modalite yapısıyla, profesyonel sınavlarda insan düzeyinde başarı göstermiştir. Grattafiori ve ark. [149] tarafından geliştirilen açık kaynaklı LLaMA 3 ailesi ise 405 milyar parametreye kadar ölçeklenerek çok dillilik, akıl yürütme ve araç kullanımı gibi görevlerde GPT-4 düzeyine yaklaşmıştır. LLaMA 3, açık erişim politikasıyla araştırma topluluğuna ileri seviye modelleri erişilebilir kılmıştır.

Encoder–decoder mimarileri de Transformer ekosisteminin önemli bir kolunu temsil etmektedir. Lewis ve ark. [150], BERT ve GPT'nin güçlü yönlerini birleştirerek BART modelini geliştirmiştir. BART, rastgele bozulmuş (noised) metinleri yeniden yapılandırmayı öğrenen bir yapı sunarak özetleme, diyalog ve soru yanıtlama görevlerinde yüksek başarı elde etmiştir. Uludoğan ve ark. [151], düşük kaynaklı Türkçe için UL2 temelli TURNA modelini geliştirerek, hem üretim hem anlama görevlerinde çok dilli modellere kıyasla üstün performans göstermiştir. Bu örnek, uygun ön-eğitim verisiyle düşük kaynaklı dillerde encoder–decoder yapıların etkinliğini kanıtlamıştır.

Model ölçeği büyüdükçe, hesaplama maliyetini azaltmak için uzman karışımı (Mixture-of-Experts, MoE) mimarileri ortaya çıkmıştır. Jiang ve ark. [152], Mixtral modelini tanıtarak katman başına 8 uzman (expert) içeren seyrek MoE yapısı önermiştir. Her adımda yalnızca iki uzman aktifleştiği için model, 47B parametreye sahip olmasına karşın yalnızca 13B parametreyle çıkarım yapabilmektedir. Zhong ve ark. [153], yumuşak ağırlıklı (differentiable) uzman seçimiyle Lory modelini sunmuş ve segment tabanlı yönlendirme (causal segment routing) tekniğiyle yüksek verim elde etmiştir. MoE sistemleri, ölçek ile verimlilik arasında yeni bir denge sağlamıştır.

Hibrit modeller, farklı mimari paradigmaları birleştirerek hem uzun bağlam işleme hem de düşük bellek maliyeti hedeflemiştir. Lieber ve ark. [154], Transformer

ve durum uzayı modellerini (state-space models) harmanlayan Jamba modelini tanıtmıştır. Mamba tabanlı katmanlar, dikkat mekanizmasına kıyasla çok daha düşük hesaplama maliyetiyle 256K token uzunluğundaki bağlamları işleyebilmiştir. Glorioso ve ark. [155], Mamba omurgasıyla tekil dikkat modülünü birleştiren Zamba modelini geliştirerek parametre maliyetini azaltmıştır. Park ve ark. [156], MambaFormer yapısıyla hibrit SSM–attention bileşiminin bağlam içi öğrenmede (In-Context Learning, ICL) daha yüksek başarı sağladığını göstermiştir.

Çok dilli modellerin yükselişi, BLOOM [157] gibi açık kaynaklı projelerle hız kazanmıştır. 176 milyar parametreli BLOOM modeli, 46 doğal dil ve 13 programlama dilini içeren ROOTS korpusu üzerinde eğitilmiş ve "Sorumlu Yapay Zeka Lisansı" ile yayımlanmıştır. Shliazhko ve ark. [158], 60 dilde eğitilmiş mGPT modelini tanıtmış; Lin ve ark. [159], 534 dili kapsayan MaLA-500 modelini geliştirerek düşük kaynaklı dillerde önemli performans artışları sağlamıştır. Bu modeller, dil temsillerinde adalet ve kapsayıcılığın önemini yeniden gündeme taşımıştır.

Ölçek artarken küçük ve verimli modeller de paralel bir araştırma yönü haline gelmiştir. Abdin ve ark. [160], yüksek kaliteli veriyle eğitilen 3,8B parametreli phi-3-mini modelinin 69% MMLU başarısı elde ederek büyük modellerle rekabet edebildiğini göstermiştir. Aynı ekibin phi-4 modeli [161], 14B parametreli olup GPT-4'ü STEM görevlerinde geride bırakmıştır. Hillier ve ark. [162], yalnızca 10–100 milyon parametreye sahip "süper küçük" dil modellerini tanıtarak veri verimliliği odaklı yeni bir paradigma önermiştir.

Uzmanlaşmış model aileleri de bu dönemde çeşitlenmiştir. Gemma ekibi [163], 2B–27B aralığında Gemma 2 ailesini tanıtmış; bilgi damıtımı (knowledge distillation) ve yerel–küresel dikkat yapılarıyla verimliliği artırmıştır. Jiang ve ark. [164], 7B parametreli Mistral modelini sunmuş ve LLaMA 2 13B'yi tüm görevlerde aşmıştır. Roziere ve ark. [165], kodlama görevleri için Code Llama ailesini geliştirerek Python sürümünde LLaMA 2 70B'yi dahi geride bırakmıştır.

Eğitim ve optimizasyon teknikleri, büyük modellerin erişilebilirliğini doğrudan etkilemiştir. Zhao ve ark. [166], düşük bellek kullanımıyla tam parametreli eğitim sağlayan GaLore yöntemini; Yang ve ark. [167], istemleri optimize ederek modelleri doğal dil tabanlı "öğrenen optimizatörler" haline getiren OPRO yaklaşımını önermiştir. Saxena ve ark. [25], veri kalitesiyle orantılı öğrenme oranı stratejileriyle gürültülü verilerde daha kararlı öğrenme sağlamıştır. PyTorch için k-bit kuantizasyon sunan bitsandbytes kütüphanesi [168], büyük modellerin donanım gereksinimlerini dramatik biçimde azaltmıştır. Yenduri ve ark. [169],

GPT mimarisine dair bütüncül bir inceleme sunarak bu modellerin başarısı kadar sınırlılıklarını anlamanın da etik geliştirme için gerekli olduğunu vurgulamıştır. Aynı şekilde, açık kaynaklı Free Dolly [170] ilk yönerge-takipli LLM olarak tarihe geçmiştir. Loodos'un Türkçe transformer modelleri [171] ve Tay dili için WangchanBERTa [172], yerel NLP ekosistemleri için önemli katkılar sunmuştur.

PEFT yöntemleri, büyük modellerin düşük kaynak ortamlarında uyarlanmasını kolaylaştırmıştır. Hu ve ark. [30] tarafından geliştirilen **LoRA** yöntemi, parametre sayısını 10.000 kat azaltırken performanstan ödün vermemiştir. Dettmers ve ark. [173], 65B modellerin 48GB GPU'da incelenmesini sağlayan QLoRA'yı tanıtmış; Jiang ve ark. [174] yüksek-rank güncellemelerle MoRA'yı önermiştir. Wu ve ark. [175], gizli temsiller üzerinde çalışan ReFT yöntemini tanıtmış ve LoReFT varyantıyla 50 kata kadar verim artışı elde etmiştir. Chen ve ark. [176], adapter tabanlı UniPELT mimarisine PromptTuning katmanı ekleyerek tam model ince ayarına yakın başarıyı çok daha az parametreyle sağlamıştır.

Sürekli öğrenme ve bilgi unutma konularında Gupta ve ark. [177], yeniden ısınma (re-warming) stratejilerinin uzun vadeli uyarlamalarda daha iyi sonuç verdiğini; Ke ve ark. [178] yumuşak maskeleme (soft-masking) mekanizmasıyla bilgi unutulmasını azalttığını göstermiştir. Luo ve ark. [179] ve Li ve ark. [180], sürekli ince ayar sırasında bilgi kaybının model boyutuyla arttığını vurgulamıştır. Bu konunun kökeni, French [181] tarafından tanımlanan "felaket unutma" kavramına dayanmaktadır.

Verimlilik ve çıkarım optimizasyonu yönünde, Kong ve ark. [182] LoRA-Switch yöntemiyle gecikmeyi 2,4× azaltmış, Addair ve ark. [183] ise çok sayıda modelin tek GPU'da verimli şekilde çalışmasını sağlayan LoRAX altyapısını geliştirmiştir. Yorumlanabilirlik çalışmaları arasında Tufanov ve ark. [184], Transformer katmanlarını görselleştiren LM Transparency Tool'u geliştirmiş; Jha ve ark. [185] ise gömme uzayları arasında veri olmadan dönüşüm sağlayan ilk yöntemi önermiştir. Bu araştırmalar, açıklanabilir yapay zeka yönelimlerinin dil modellerine taşındığını göstermektedir.

Genel olarak, mimari araştırmalar hızla hibrit, adaptif ve veriye duyarlı bir yöne evrilmektedir. Dikkat ve durum uzayı modellerinin birleştiği yapılar, uzun bağlamları verimli işleyebilen yeni dikkat biçimleri, öğrenilmiş tokenizasyon sistemleri, alan odaklı uzmanlaşmalar ve enerji açısından sürdürülebilir küçük modeller geleceğin temel yönelimlerini oluşturmaktadır. LoRA, QLoRA ve ReFT gibi yöntemler, devasa modellerin düşük maliyetle uyarlanmasını mümkün kılsa da, ölçeklenebilirlik ve enerji verimliliği arasındaki dengenin korunması halen

temel bir araştırma problemidir. Küçük modellerin yüksek kaliteli veriyle büyük modellerle yarışabilmesi [160], [161], "veri merkezli yapay zeka" anlayışını güçlendirmiştir. Son olarak, açık kaynaklı LLaMA, BLOOM, Gemma ve benzeri modellerin yaygınlaşması, büyük dil modeli araştırmalarını demokratikleştirmiştir. Ancak, açık erişimin beraberinde getirdiği güvenlik ve etik sorumluluklar da dikkate alınmalıdır. Geleceğin araştırmaları, açık bilimin faydalarıyla güvenliğin gerekleri arasında denge kurarak, temel modelleri toplumun tamamına hizmet edecek biçimde geliştirmeyi hedeflemelidir.

3.6 Alana Özgü ve Uygulamalı Büyük Dil Modelleri

Genel amaçlı büyük dil modelleri birçok farklı görevde etkileyici sonuçlar üretse de, yüksek riskli uygulamalarda alan bilgisi, terminoloji ve akıl yürütme kalıpları belirleyici hale gelmektedir. Bu nedenle son yıllarda özellikle sağlık alanına yönelik özelleştirilmiş büyük dil modelleri geliştirilmektedir.

Sağlık alanı, teknik doğruluk kadar güvenilirlik ve gizlilik gereksinimleri nedeniyle en hassas alanlardan biridir. Nazi ve ark. [134], sağlık alanında LLM'lerin gelişimini geleneksel dil modellerinden alana özgü sistemlere geçiş perspektifiyle incelemiş ve adlandırılmış varlık tanıma, ilişki çıkarımı, belge sınıflandırma, doğal dil çıkarımı ve çok modlu uygulamalar gibi görevlerde klinik doğrulama ve güvenli entegrasyonun halen temel zorluklar olduğunu göstermiştir. Mumtaz ve ark. [135], onkoloji, dermatoloji, nörodejeneratif hastalıklar, diş hekimliği ve ruh sağlığı gibi farklı branşlarda LLM kullanımını değerlendirmiş; her alanın kendine özgü gereksinimleri bulunduğunu vurgulamıştır. Görsel-metinsel entegrasyonun dermatolojide kritik olduğu, ruh sağlığında ise empatik yanıt üretiminin öne çıktığı belirtilmiştir. Bu çalışmalar, veri gizliliği, yorumlanabilirlik ve klinik doğrulama eksikliklerinin halen önemli engeller olduğunu ortaya koymaktadır. Huang ve ark. [20], klinik notlar üzerinde eğitilen ClinicalBERT modelinin hasta yeniden yatış tahmininde yüksek başarı elde ettiğini ve dikkat ağırlıklarının doktorların karar mekanizmasını anlamasına katkı sunduğunu göstermiştir. Li ve ark. [21], 100.000 hasta-doktor diyaloguyla eğitilen ChatDoctor modelini tanıtarak bilgi erişimini çevrim içi destekli hale getirmiş, Han ve ark. [186] ise 160.000'den fazla tıbbi diyalogla eğitilen MedAlpaca'yı geliştirerek açık kaynaklı tıbbi modellerin potansiyelini göstermiştir. Labrak ve ark. [22], Mistral tabanlı BioMistral ailesini tanıtmış ve tıp alanında mevcut açık modellerden daha üstün performans elde etmiştir. Fan ve ark. [108], çoklu LLM ajanlarının sanal hastane ortamında etkileşimini test eden yapay zeka (Artificial Intelligence, AI) Hospital sistemini sunmuş; uzun diyaloglarda klinik akıl yürütmenin tutarlılığında açıklar

bulunduğunu göstermiştir. Li ve ark. [187], Agent Hospital adlı benzer bir simülasyon ortamında LLM tabanlı doktor ajanlarının binlerce sanal vaka üzerinde deneyim kazandığını ortaya koymuş, ancak bu tür uygulamaların klinik geçerlilik açısından sıkı denetim gerektirdiğini vurgulamıştır. Eğitim alanında LLM'lerin potansiyeli, kişiselleştirilmiş öğrenme, ödev desteği ve öğretmen yardımcılığı gibi işlevlerle büyüktür; ancak pedagojik geçerlilik, eşitlik ve etik konular öncelikli endişelerdir. Kasneci ve ark. [188], LLM'lerin eğitimde kullanımını fırsatlar ve riskler çerçevesinde analiz etmiş; özellikle eleştirel düşünme, akademik dürüstlük ve pedagojik yönlendirme konularına dikkat çekmiştir. Mustafa Seyrek [189], Türk öğretmenlerin yapay zekaya ilişkin algılarını incelemiş ve öğretmenlerin soru hazırlama, etkinlik planlama ve öğrenci değerlendirme süreçlerinde yapay zeka araçlarını aktif biçimde kullandığını, ancak yaratıcılık kaybı ve eşitsizlik endişelerinin yaygın olduğunu saptamıştır. Eğitimde model uyarlamaları, yaşa uygun içerik üretimi, pedagojik denge ve doğrulama mekanizmaları gerektirir.

Dile özgü alan uyarlamaları, özellikle düşük kaynaklı diller için temel bir araştırma hattı haline gelmiştir. Kesgin ve ark. [37], tamamen Türkçe verilerle eğitilen cosmosGPT modelini tanıtmış ve küçük ölçekli tek dilli modellerin büyük çok dilli modellerle rekabet edebileceğini göstermiştir. Toraman ve ark. [26], Türkçe açık kaynak modellerin sürekli eğitim, yönerge ince ayarı ve görev tabanlı ince ayar stratejilerini karşılaştırarak kelime haznesi genişletmenin etkisiz, veri kalitesinin ise belirleyici olduğunu ortaya koymuştur. Açıkgöz ve ark. [41], Türkçe LLM performansını hem İngilizce'den uyarlama hem de sıfırdan ön-eğitim yaklaşımlarıyla incelemiş ve TR-LLM karşılaştırma panosu geliştirerek Türkçe'nin özgün değerlendirme çerçevesini tanımlamıştır.

Çok dilli ve kültürlerarası uyarlama çalışmaları, dilsel çeşitliliği kapsayan modelleme yaklaşımlarına yönelmiştir. Yang ve ark. [190], Macarca için hem tekdilli hem üç dilli GPT-3 modelleri eğitmiş ve orta kaynaklı dillerde çok dilli eğitimin uygulanabilirliğini göstermiştir. Larcher ve ark. [191], Portekizce için sıfırdan model eğitimi yaparak dil-özgü yaklaşımların dilbilimsel içgörüler kazandırdığını göstermiştir. Kohli ve ark. [192], düşük kaynaklı Odia dili için LLaMA-7B tabanlı ince ayar süreci yürütmüş ve yönerge verisinin kritik önemini ortaya koymuştur. Garcia ve ark. [193], Portekizce görevlerde Bode modelini; Santilli ve ark. [194], İtalyanca talimat izleyen Camoscio modelini; Almeida ve ark. [195] ise Portekizce eğitilmiş Sabiá-2 ailesini tanıtmıştır. Correa ve ark. [196], Brezilya Portekizcesi için geliştirilen TeenyTinyLlama modeliyle düşük kaynaklı dillerde erişilebilir model geliştirme stratejilerini örneklemiştir.

Alanlar arası bilgi aktarımı, modelin hem genel hem uzmanlaşmış görevlerde

dengeli performans göstermesi açısından önemlidir. Zhang ve ark. [197], veri harmanlama stratejilerinin modelin genel bilgiyle alan uzmanlığı arasındaki dengeyi koruyabildiğini göstermiştir. Ling ve ark. [198], alan uzmanlaşmasını sistematik biçimde sınıflandırmış ve görev çeşitliliği ile veri kapsamının uzmanlaşmanın temel belirleyicileri olduğunu vurgulamıştır. Zheng ve ark. [199], DragFT adlı çok yönlü ince ayar çerçevesiyle alan odaklı makine çevirisinde GPT-3.5 ve GPT-4'ü aşan sonuçlar elde etmiştir. Bu yaklaşımlar, alan bilgisinin yalnızca aktarım değil, aynı zamanda bilgi koruma stratejileriyle bütünleştirilmesi gerektiğini göstermektedir.

Alan odaklı değerlendirme, genel karşılaştırıcıların ötesine geçerek sağlık gibi alanlar için özel ölçütler gerektirir. Inserte ve ark. [200], finansal duygu analizinde küçük ölçekli modellerin özel verilerle büyük modellerle yarışabileceğini göstermiştir. Sağlık uygulamalarında yalnızca doğruluk değil; klinik güvenlik, kılavuz uyumu, belirsizlik yönetimi ve hasta mahremiyetine uygunluk gibi boyutların da ölçüme dahil edilmesi gerekir. Bu durum, alan odaklı LLM'lerin yalnızca performans değil, aynı zamanda uygunluk ve sorumlu kullanım ölçütlerine göre de test edilmesi gerektiğini göstermektedir.

Etik ve uygulamalı boyutlar, alana özgü modellerin sürdürülebilir kullanımının merkezinde yer alır. Sağlık alanında hasta güvenliği ve hasta mahremiyeti, bu sistemlerin sorumlu biçimde kullanılabilmesi için temel gereksinimlerdir. Maties [201], sağlıkta üretken yapay zeka uygulamalarının güvenlik ve uyumluluk gereksinimleri açısından halen olgunlaşmadığını bildirmiştir. Bu nedenle federatif öğrenme, farklılaştırılmış gizlilik ve yerel dağıtım gibi yaklaşımlar giderek daha fazla önem kazanmaktadır. Geniş veri ihtiyacı ile gizlilik kısıtları arasındaki denge, aktif bir araştırma konusudur.

3.7 Dil Modeli Uyarlaması ve Aktarımı

Büyük dil modellerinin yeni dillere, alanlara ve görevlere uyarlanması, İngilizce ve yüksek kaynaklı dillerin ötesinde yapay zeka yeteneklerinin demokratikleşmesi açısından temel bir araştırma alanıdır. Özellikle düşük kaynaklı dillerde diller arası aktarım, kelime dağarcığı uyarlaması, sürekli ön-eğitim ve felaket unutmanın azaltılması gibi konular, yapay zekanın evrensel erişilebilirliğini sağlama yolunda kritik rol oynamaktadır.

Diller arası aktarım, modellerin farklı diller arasında bilgi paylaşabilme kapasitesini ifade eder. Artetxe ve ark. [202], çok dilli modellerin başarısının yalnızca paylaşılan kelime haznelerine dayanmadığını, tekdilli temsillerin de aktarılabilir soyutlama

katmanları geliştirebildiğini göstermiştir. Transformer tabanlı bir modeli bir dilde eğitip, yalnızca gömme (embedding) katmanını yeni bir dil için yeniden öğrenerek çok dilli BERT ile karşılaştırılabilir sonuçlar elde etmişlerdir. Qin ve ark. [203], çok dilli LLM'leri hizalama stratejilerine göre sistematik biçimde sınıflandırmış, kod değişimi, kültürel adaptasyon ve çok dilli akıl yürütme gibi eğilimleri özetlemiştir. Zhao ve ark. [39], LLaMA modeli için yapılan deneylerde, düşük kaynaklı dillerde yalnızca %1 oranında ön-eğitim verisiyle bile güçlü aktarım performansına ulaşılabileceğini göstermiştir.

Dillerin yapısal farklılıkları, kelime dağarcığı ve tokenizasyon düzeyinde özel uyarlamalar gerektirir. Cui ve ark. [204], LLaMA modeline 20.000 yeni Çince token ekleyerek anlam çözümlemesi ve üretim başarısını artırmıştır. Csaki ve ark. [205], İngilizce modelleri Macarca ve Tayca'ya uyarlayarak sıfırdan eğitime kıyasla daha iyi sonuçlar elde etmiş; dikkatli veri karışımıyla felaket unutmayı azaltmıştır. Dobler ve ark. [206], yeni dillere token eklerken gömmeleri anlamsal benzerliğe göre başlatan FOCUS yöntemini tanıtmış, bu sayede rastgele başlatmalara kıyasla çok daha kararlı sonuçlar elde edilmiştir. Ebrahimi ve ark. [207], 1600 dilde Yeni Ahit verisiyle eğitilen modellerde düşük veri koşullarında bile doğru tasarım sayesinde güçlü aktarım gözlemlemiştir.

Sürekli ön-eğitim stratejileri, modelin yeni verilerle evrim geçirmesini sağlarken unutma riskini de beraberinde getirir. Yukarıda tartışılan sürekli öğrenme yaklaşımları (bkz. Bölüm 3.5), bu bağlamda da geçerliliğini korumaktadır. Li ve ark. [180], sürekli ön-eğitim sırasında bilgi tekrarı ve unutmanın model performansını düşürdüğünü ortaya koymuştur.

Yönerge ince ayarı (instruction tuning) ve insan geri bildirimiyle güçlendirilmiş öğrenme (Reinforcement Learning from Human Feedback, RLHF), model davranışını hedef kullanıcı niyetleriyle hizalama açısından kritik önemdedir. Wei ve ark. [208], FLAN modelini 60'tan fazla görevle eğiterek GPT-3'ü birçok ölçütte geride bırakmıştır. Ouyang ve ark. [209], 1,3B parametreli InstructGPT modelinin 175B GPT-3'e kıyasla daha tutarlı yanıtlar ürettiğini göstermiştir. Zhou ve ark. [210], yalnızca 1.000 örnekle eğitilen LIMA modelinin GPT-4 düzeyine yakın sonuçlar verdiğini belirlemiş ve büyük bilgi birikiminin ön-eğitimde kazanıldığını vurgulamıştır. Köksal ve ark. [36], düşük kaynaklı dillerde insan etiketine gerek duymadan yönerge verisi üreten MURI yöntemini tanıtmış ve 200 dilde 2 milyon örnekle yüksek başarı elde etmiştir.

Düşük kaynaklı dillerin uyarlanması, yapay zekada dilsel kapsayıcılığın sağlanmasında kilit bir adımdır. Kuulmets ve ark. [211], Estonca için küçük ölçekli

ek ön-eğitimin çapraz dil yönerge ince ayarıyla büyük kazanımlar sağladığını göstermiştir. Holmström ve ark. [212], İsveççe'de çeviri tabanlı yönerge ince ayarının sıfır-atış performansını artırdığını bildirmiştir. Azime Walia [213], Amharca için açık kaynak boru hatları geliştirmiş, Owen ve ark. [214] ise Endonezce ve 11 bölgesel dilde eğitilen Komodo modelinin GPT-3.5'i birçok görevde geride bıraktığını göstermiştir.

Çok dilli mimariler, farklı dillerdeki temsilleri aynı çatı altında bütünleştirmek için çeşitli stratejiler önermektedir. Blevins ve ark. [215], "çok dillilik laneti"ni (curse of multilinguality) aşmak için uzman modellerin birlikte çalıştığı X-ELM yapısını tanıtmış ve yeni dillerin kolayca eklenebilmesini sağlamıştır. Martins ve ark. [216], Avrupa Birliği'nin tüm dillerini kapsayan EuroLLM projesini tanıtarak çok dilli tokenizasyonun verimlilik üzerindeki etkisini tartışmıştır. Üstün ve ark. [217], 101 dili destekleyen Aya modelini geliştirerek 99 dilde güvenlik, yanlılık ve performans ölçütlerini kapsayan geniş bir değerlendirme sunmuştur.

Aktarım öğrenmesinde verimlilik, hem hesaplama hem veri kalitesi açısından belirleyici bir faktördür. Wang ve ark. [218], pahalı insan tercihi etiketlemeleri olmadan yüksek başarı elde eden C-RLFT yöntemini tanıtmış; Mukherjee ve ark. [219], GPT-4'ün düşünme adımlarını taklit eden Orca modelini geliştirerek Vicuna-13B'yi aşmıştır. Bu yöntemler, açıklama izlerine (explanation traces) dayalı öğretim verilerinin modelin akıl yürütme kapasitesini artırabileceğini göstermektedir.

Diller arası değerlendirme ve ölçüt geliştirme çalışmaları, kültürel önyargıları ve ölçüm farklılıklarını gidermek açısından önemlidir. Husain ve ark. [220], Latin olmayan alfabelerde romanizasyonun hizalamayı iyileştirdiğini göstermiştir. Cengiz ve ark. [38], 17 Türkçe karşılaştırıcı veri kümesinin %70'inin kalite standartlarını karşılamadığını belirlemiş ve kültürel bilgi eksikliğinin model değerlendirmelerini olumsuz etkilediğini vurgulamıştır. Luo ve ark. [179], 1B–7B parametreli modellerde sürekli yönerge ince ayarı sırasında felaket unutmanın arttığını ve BLOOMZ'un mT0'a göre daha dayanıklı olduğunu göstermiştir (felaket unutma kavramı için bkz. Bölüm 3.5).

Model birleştirme (model merging) ve topluluk temelli yaklaşımlar, bilgi paylaşımını ölçeklenebilir hale getirmektedir. Yukarıda değinilen KAFA-Merge yöntemi [43], bu alanda Türkçe özelinde önemli bir katkı sunmuştur. Açık lisanslı Falcon modelleri [221], 7B–180B parametre aralığında 3,5 trilyon token üzerinde eğitilerek PaLM ve LLaMA 2'yi geride bırakmış; açık bilim ekosisteminin hızlanmasına katkı sağlamıştır. Benzer şekilde, 01.AI ekibi [222] tarafından

geliştirilen Yi modelleri, 3,1 trilyon token üzerinde eğitilmiş ve veri kalitesinin model boyutundan daha kritik olduğunu ortaya koymuştur.

Çok dilli akıl yürütme ve güvenlik araştırmaları da giderek önem kazanmaktadır. Zhang ve ark. [137], çeviri bağımlı ve bağımsız görevler arasındaki farkı analiz ederek LLM'lerin dil özgül görevlerde zorlandığını belirtmiştir. Wang ve ark. [138], XSafety karşılaştırıcısı ile 10 dilde 14 güvenlik boyutunu değerlendirmiş ve İngilizce dışındaki sorgularda hatalı yanıt oranlarının iki katına çıktığını göstermiştir. Güvenlik odaklı istemlerle bu oran %19,1'den %9,7'ye düşürülmüştür.

3.8 Çok Modlu ve Yeni Eğilimler

Büyük dil modellerinin evrimi yalnızca metin tabanlı biçimlerle sınırlı kalmayıp, görsel, işitsel ve diğer duyusal girdilerin entegrasyonuyla daha kapsamlı yapay zeka sistemlerinin gelişmesini sağlamıştır. Bu yönelim, insan bilişinin çoklu duyusal doğasını taklit eden sistemlerin ortaya çıkmasına yol açmakta ve geleceğin genel yapay zekaya (Artificial General Intelligence, AGI) doğru atılmış önemli bir adımı temsil etmektedir.

Team Gemini [223], görüntü, ses, video ve metin anlayışı alanlarında olağanüstü yetenekler sergileyen Gemini model ailesini tanıtmıştır. Ultra, Pro ve Nano sürümlerinden oluşan bu modeller, karmaşık akıl yürütmeden cihaz üzerinde çalışmaya kadar geniş bir uygulama yelpazesi sunmaktadır. Gemini Ultra, 32 karşılaştırıcıdan 30'unda alanın en yüksek performansına ulaşmış ve MMLU testinde insan uzman düzeyine erişen ilk model olmuştur. Ayrıca, 20 çok modlu değerlendirmede de mevcut rekorları aşarak birleşik çok modlu mimarilerin çapraz-modal akıl yürütme potansiyelini ortaya koymuştur. Jang ve ark. [224], Görsel-Kompozit Sorgu ile Görsel Getirme (Composed Image Retrieval, CIR) görevinde metin ve görsel temsilleri doğrudan birleştiren **Slerp** yöntemiyle sıfır-atış başarı elde etmiş ve Metin-Anchored İnce Ayarı (Text-Anchored Tuning, TAT) ile görüntü kodlayıcısını optimize ederek metin-görsel boşluğunu azaltmıştır. Yue ve ark. [113], 39 dili destekleyen açık kaynaklı çok dilli çok modlu model PANGEA'yı tanıtmış, kültürel uygunluğu da gözeten çok dilli görsel-sözel anlama sistemlerinin mümkün olduğunu göstermiştir.

Konuşma ve ses işleme, doğal insan-bilgisayar etkileşiminin temel bileşenlerinden biridir. Ancak, bu alan görsel modalitelere kıyasla daha az incelenmiştir. Konuşma yeteneğinin modellerle bütünleştirilmesi, daha doğal kullanıcı arayüzleri, görme engelliler veya okuryazarlık kısıtlı kullanıcılar için erişilebilirlik sağlamak açısından önemlidir. Bu alandaki temel zorluklar, değişken uzunluktaki ses

dizilerini işleme, gürültülü ortamlarda kararlılık, aksan çeşitliliği ve vurgulama, tonlama, ritim gibi prozodik özelliklerin yakalanmasıdır. Yeni yaklaşımlar, transkripsiyon bağımlılığını ortadan kaldırarak doğrudan ham ses sinyallerinden öğrenebilen uçtan uca modeller geliştirmektedir. Bu yöntemler, paralinguistik bilgileri koruyarak anlam çıkarımında daha derin bir temsil elde edilmesini sağlar.

Çok modlu akıl yürütme yalnızca modalite entegrasyonundan ibaret değildir; farklı veri türleri arasında bilgi aktarımı ve anlam çıkarımı yapabilme kapasitesi gerektirir. Yue ve ark. [113], uygun veri ve mimari seçimiyle çok dilli görsel soru yanıtlama, açıklama üretimi ve mantık yürütme görevlerinin mümkün olduğunu göstermiştir. Bununla birlikte, çok modlu sistemlerin küresel ölçekte uygulanabilirliği kültürel bağlamların doğru yorumlanmasını da zorunlu kılar; çünkü renkler, semboller ve jestler kültürler arasında farklı anlamlar taşıyabilir.

Gelişmiş model mimarileri, çok modlulukta soyutlama ve görev uyarlanabilirliğini artırmak için yeni yollar önermektedir. Hong ve ark. [225], insan iş akışlarını çok ajanlı LLM sistemlerine entegre eden MetaGPT çerçevesini sunmuştur. Bu yapı, Standart İşletim Prosedürlerini (Standard Operating Procedure, SOP) istem dizilerine kodlayarak yazılım mühendisliği görevlerinde yüksek tutarlılık sağlamıştır. Sabour ve ark. [226], kapsül ağları (capsule networks) ve dinamik yönlendirme (routing-by-agreement) kavramlarını tanıtarak, dil modellerinde bileşik yapıları anlamlandırmak için yeni bir paradigma önermiştir. Locatello ve ark. [146], Slot Attention bileşeniyle çoklu nesne temsillerini ayrıştırarak uzun bağlamlı dil modellemede daha etkin bileşik temsiller üretmiştir.

Verimlilik ve optimizasyon alanında da dikkate değer yenilikler ortaya çıkmıştır. Lin ve ark. [81], yalnızca anlam açısından kritik token'ları öğrenen seçici dil modelleme (selective language modeling) yöntemini önermiş ve veri verimliliğini artırarak MATH veri kümesinde %30 doğruluk artışı sağlamıştır. Zhao ve ark. [227], dış ödül sinyali olmaksızın kendi güven puanlarını (self-certainty) ödül olarak kullanan INTUITOR yöntemini tanıtmış, bu sayede tam denetimsiz biçimde GRPO düzeyinde performans elde etmiştir. Jolicoeur-Martineau [228], yalnızca 7 milyon parametreli Tiny Recursive Model (TRM) ile hiyerarşik akıl yürütmede büyük modellerle yarışan sonuçlar elde etmiş ve küçük ama yinelemeli yapılarla verimliliğin korunabileceğini göstermiştir.

Uzun bağlam işleme, büyük dil modellerinin en zorlu problemlerinden biridir. Standart dikkat mekanizmalarının dizinin uzunluğuyla karesel olarak ölçeklenmesi, binlerce token'lık girdilerde hesaplama maliyetini artırmaktadır. Lieber ve ark. [154], Transformer ve Mamba durum uzayı modellerini birleştirerek 256K token'a

kadar bağlam işleyebilen Jamba modelini tanıtmış, uzun bağımlılıkları verimli biçimde işleyip yüksek doğruluk sağlamıştır. Ancak, uzun bağlam yalnızca kapasite değil, ilgisiz bilgiyi filtreleme ve tutarlı akıl yürütme becerisi de gerektirir. Bu nedenle gelecekte yeni hedef fonksiyonlar ve veri stratejileriyle bellek verimliliği geliştirilecektir.

Yeni eğitim paradigmaları, tokenizasyonun ötesinde öğrenmeyi yeniden tanımlamaktadır. Videau ve ark. [56], sabit tokenizasyonu aşmak için bayt düzeyinden sözcük ve ifadeye kadar hiyerarşik temsiller öğrenen çok ölçekli özyinelemeli U-Net mimarisini önermiştir. Wang ve ark. [229], LLM ajanlarının eylemlerini yürütülebilir Python kodu biçiminde temsil eden CodeAct çerçevesini geliştirerek çok adımlı etkileşimlerde başarı oranını %20'ye kadar artırmıştır. Bu yaklaşım, program yürütümü ve dil anlama arasındaki sınırların giderek belirsizleştiğini göstermektedir.

Değerlendirme ve analiz araçları da hızla evrilmektedir. Contextual Retrieval yaklaşımı [230], bilgi erişimde bağlam kaybını azaltmak için açıklayıcı metinlerle zenginleştirilmiş gömme yapıları kullanarak doğruluğu %67 oranında artırmıştır. Claude Sonnet 4.5 sürümü [231], akıl yürütme, araç kullanımı ve yönerge izleme alanlarında önemli ilerlemeler kaydederek birkaç ay içinde önceki sürümleri aşmıştır. Bu sonuç, model evriminin artık yıllar değil, aylarla ölçüldüğünü göstermektedir.

Uygulamalı sistemler, temel modellerin klinik, mobil veya cihaz içi uygulamalara entegrasyonunu hızlandırmıştır. Apple [232], cihaz üzerinde ve sunucu tarafında çalışan gizlilik odaklı temel modelleri tanıtarak hibrit bir dağıtım paradigması önermiştir. Cihaz üzerindeki modeller çevrimdışı çalışabilirken, sunucu tabanlı modeller yüksek hesaplama gücü gerektiren görevlerde kullanılmaktadır. Leong ve ark. [233], LLaMA3-8B tabanlı MediGen modelini geliştirerek doktor-hasta diyaloglarından tıbbi dokümantasyon üretimini otomatikleştirmiş; aynı ekip [234], MediNotes sistemini tanıtarak sesli ve yazılı girdilerden gerçek zamanlı Subjective-Objective-Assessment-Plan (SOAP) notu üreten bir yapı geliştirmiştir. QLoRA ve PEFT yöntemleriyle optimize edilen bu sistem ACI-BENCH veri kümesinde %79,2 doğruluk elde etmiştir.

Model anlama ve yorumlanabilirlik, bilimsel güvenin temelidir. Dikkat görselleştirme, nedensel müdahale, iç temsil probingi gibi yöntemler modelin iç dinamiklerini anlamayı sağlar. Bununla birlikte, yüksek riskli alanlarda yorumlanabilirlik yalnızca teknik bir tercih değil, toplumsal bir zorunluluk haline gelmiştir. Açıklanabilirlik, kullanıcı güvenini tesis etmek ve insan denetimini

sürdürmek için vazgeçilmezdir.

Etik boyutlar, bu teknolojilerin yönünü belirleyen en kritik faktörlerden biridir. Baeza-Yates [143], ayrımcılık, anlamsal eksiklik, kültürel önyargılar ve bilişsel yanlılık gibi temel sorunlara dikkat çekmiştir. Bender ve ark. [121], sürekli büyüyen modellerin çevresel ve toplumsal maliyetlerine değinmiş, verinin rastgele toplanması yerine özenli kürasyonun önemini vurgulamıştır. Bu yaklaşımlar, "daha büyük her zaman daha iyi değildir" ilkesini sorumlu yapay zeka geliştirme anlayışının merkezine yerleştirmiştir.

Politika ve düzenleme boyutunda, America's AI Action Plan [235], güvenlik, etik kullanım, istihdam ve yenilik arasında denge kuran ulusal stratejiyi yayımlamıştır. Aynı kapsamda oluşturulan politika çerçevesi [236], gizlilik, adalet, demokratik değerler ve insan özerkliğini koruma ilkelerini temel almıştır.

Açık bilim ve ticarî gizlilik arasındaki gerilim, günümüz LLM araştırmalarının belirleyici özelliklerinden biridir. LLaMA, BLOOM ve Falcon gibi açık kaynaklı modeller araştırmayı demokratikleştirirken; GPT-4, Claude ve Gemini gibi kapalı modeller sınırlı şeffaflık sunmaktadır. Açık yaklaşımlar katılım ve denetim kolaylığı sağlarken kötüye kullanımı artırabilir; kapalı yaklaşımlar ise güvenliği korurken erişimi sınırlar. Gelecekte bu iki uç arasında güvenlik odaklı hibrit paylaşım stratejileri öne çıkacaktır.

3.9 Son Gelişmeler ve Geleceğe Yönelik Eğilimler (2024–2025)

Türkçe ve düşük kaynaklı diller için büyük dil modelleri alanı, 2024–2025 döneminde dikkate değer bir ivme kazanmış; çok dilli NLP alanında paradigma değişimleri, verimlilik odaklı mimariler ve etik yönetişim yaklaşımlarıyla şekillenmiştir. Bu dönemdeki gelişmeler, yalnızca performans artışını değil, aynı zamanda erişilebilirlik, açıklanabilirlik ve kültürel kapsayıcılığı da merkeze almıştır.

Son yıllarda ortaya çıkan *Küçük Dil Modelleri (Small Language Models, SLM)* eğilimi, "daha büyük her zaman daha iyidir" anlayışını sorgulamıştır. Phi-3, Phi-4 ve Gemma 2 gibi modeller, yüksek kaliteli veri seçimi ve gelişmiş eğitim stratejileriyle yüz milyarlarca parametreye sahip dev modellere rakip hale gelmiştir. Türkçe NLP açısından bu dönüşüm kritik bir fırsat yaratmıştır: büyük çok dilli modellerin yüksek hesaplama maliyeti, tarihsel olarak düşük kaynaklı dillerin aleyhine olurken, SLM yaklaşımı üniversitelerin, araştırma merkezlerinin ve yerel girişimlerin kendi dil modellerini geliştirmesini mümkün kılmaktadır. 1-7

milyar parametre aralığındaki Türkçe modellerin, nitelikli Türkçe derlem verisiyle eğitildiklerinde büyük ölçekli çok dilli sistemleri aşabildiği gözlemlenmiştir. Küçük modellerin çevrimdışı çalışabilirliği, düşük enerji tüketimi ve hızlı iterasyon imkanı, bu alanın demokratikleşmesine önemli katkı sağlamaktadır.

Çok dillilikte yaşanan en temel problemlerden biri olan "*curse of multilinguality*" sorununa karşı 2024–2025 döneminde öne çıkan çözüm, MoE mimarileri olmuştur. Bu mimariler, her girdi için yalnızca belirli uzman alt ağları etkinleştirerek toplam kapasiteyi artırırken hesaplama maliyetini düşürür. Türkçe ve benzeri eklemeli yapıya sahip diller için MoE tabanlı tasarımlar, morfolojik uzman modüller aracılığıyla biçimbilimsel yapıların daha doğru temsil edilmesini sağlar. Ayrıca, modüler yapı, yeni Türk lehçelerinin veya dillerinin modele sonradan eklenebilmesini kolaylaştırarak sürekli genişleyebilen çok dilli sistemlerin önünü açmaktadır.

Bağlam pencerelerinin 4K'dan 128K hatta 200K token'a kadar genişlemesi, Türkçe gibi serbest sözdizimine sahip diller için dönüştürücü bir yenilik olmuştur. Uzun menzilli bağımlılıkların yakalanabilmesi; akademik veya teknik belgelerin parçalara bölünmeden işlenmesi, diyalog geçmişinin korunması ve uzun metinlerde söylem bütünlüğünün sağlanması açısından büyük avantaj sunmaktadır. Ancak, bu potansiyelin tam olarak kullanılabilmesi, geniş bağlamı verimli biçimde kullanabilen eğitim stratejilerinin geliştirilmesine bağlıdır.

Metin tabanlı modellerin ötesine geçilerek görsel ve işitsel modalitelerin entegrasyonu da Türkçe için yeni bir araştırma yönü açmıştır. Görsel-dil modelleri, Türkçe görsel açıklama üretimi, erişilebilirlik çözümleri ve kültürel mirasın dijitalleştirilmesi gibi alanlarda etkili olurken; konuşma entegrasyonu, sesli asistanlar, dil öğrenme uygulamaları ve telaffuz geri bildirimi sistemleri için temel oluşturmaktadır. Ancak, İngilizce merkezli çok modlu veri setlerinin Türkçe karşılıklarının azlığı, kültürel anlam çeşitliliğini yansıtan yerel veri kümelerinin oluşturulmasını zorunlu kılmaktadır.

Verimli ince ayar yöntemleri de bu dönemde önemli yeniliklerle genişlemiştir. **LoRA**'nın başarısını takiben, DoRA (Weight-Decomposed LoRA) ağırlıkları büyüklük ve yön bileşenlerine ayırarak farklı öğrenme oranlarıyla güncelleme yaparken, AdaLoRA (Adaptive LoRA) parametre bütçesini modül önemine göre dinamik biçimde dağıtarak Türkçe'nin morfolojik yapısına uyarlanabilirliği artırmıştır. GaLore ve ReLoRA gibi yöntemler, gradyanları düşük-rank uzaylara projelendirerek bellek verimliliği sağlamıştır. Bu yöntemler, Türkçe modellerin alan uyarlaması, sürekli öğrenme ve görev-özgü ince ayarlama süreçlerinde düşük

maliyetle yüksek başarı elde etmesine olanak tanımaktadır.

Düşük kaynaklı dillerde sentetik veri üretimi, veri yetersizliğini gidermek için etkili bir strateji olarak benimsenmiştir. Gelişmiş LLM'ler, Türkçe soru-cevap çiftleri, çeviri örnekleri, yönerge tabanlı veri kümeleri ve kültürel içerik üretebilmektedir. Ancak, bu verilerin güvenilirliğini sağlamak için insan denetimi, çoklu model üretimi ve kalite tabanlı filtreleme mekanizmaları gereklidir. Bu yaklaşım, özellikle tıbbi alanda Türkçe veri eksikliğini azaltarak model çeşitliliğini artırmaktadır.

Yapay zeka sistemlerinin kültürel değerlere uygun davranış sergilemesini hedefleyen *Anayasal Yapay Zeka (Constitutional AI)* yaklaşımı, 2025 itibarıyla önemli bir araştırma alanı haline gelmiştir. Değer uyumlaştırması (value alignment), modellerin davranışlarını belirli etik ilkeler çerçevesinde yönlendirir. Türkçe için bu, yerel kültürel normların, etik metinlerin ve kurumsal çerçevenin dikkate alınması anlamına gelmektedir. Türk toplumunun iletişim biçimleri ve değer sistemleri üzerinde eğitilen modeller, kültürel olarak daha uygun yanıtlar üretebilir.

Ajan tabanlı sistemler, araç kullanımı ve çok adımlı planlama yeteneğiyle Türkçe uygulamalara da yayılmaktadır. Bu sistemler müşteri desteği, eğitim, kodlama asistanlığı ve belge analizi gibi alanlarda Türkçe etkileşim sunmaktadır. Ancak, yönerge verisi eksikliği ve kültürel bağlama uygun yanıt üretimi, ajansal sistemlerin Türkçe uyarlamasında halen açık bir araştırma alanıdır.

Model birleştirme (*model merging*) teknikleri de 2024–2025 döneminde büyük ilerleme kaydetmiştir. DARE (Drop and Rescale) ve TIES (Trim, Elect Sign & Merge) gibi yöntemler, farklı modellerin ağırlıklarını etkin biçimde harmanlayarak çakışmaları azaltmaktadır. Bu yaklaşımlar, Türkçe dil yetenekleri ile alan uzmanlığını birleştiren hibrit modellerin geliştirilmesini, Türk lehçeleri arasında bilgi aktarımını ve çok yetenekli birleşik modellerin oluşturulmasını kolaylaştırmaktadır.

Yeni değerlendirme yöntemleri ve karşılaştırıcılar, modellerin doygunluğa ulaşmasıyla yeniden önem kazanmıştır. Türkçe için halen eksik olan alanlar arasında çok turlu diyalog, kültürel bilgi, akıl yürütme, kod üretimi ve teknik metin yazımı yer almaktadır. Bu doğrultuda dinamik ve çok boyutlu karşılaştırıcılar geliştirilmektedir. Ayrıca, kültürel uygunluk testleri, yerel dillerdeki toplumsal normlarla hizalanmayı ölçmek için kullanılmaya başlanmıştır.

3.10 Bölüm Özeti

Bu bölüm, Türkçe ve düşük kaynaklı diller bağlamında literatürü tokenizasyon, veri kalitesi, değerlendirme ve alan uyarlaması eksenlerinde derlemiş; güncel eğilimleri ve metodolojik boşlukları görünür kılmıştır. Ayrıca çok dilli karşılaştırıcıların sınırlılıkları, etik ve lisans tartışmaları ile yeni değerlendirme paradigmalarının yükselişi özetlenmiştir. Böylece tezin önerdiği bütüncül yaklaşımın literatürdeki yerini ve gerekçesini temellendiren bir arka plan sunulmuştur.

4
VERİ SETLERİ VE KAYNAKLAR

Bu bölüm, tez kapsamında geliştirilen ve kullanılan veri setlerini tek bir çerçevede toplar. Türkçe, morfolojik açıdan zengin ve görece düşük kaynaklı bir dil olduğundan, hem model eğitimi hem de nesnel değerlendirme için güvenilir, izlenebilir ve yeniden üretilebilir veri kaynaklarına ihtiyaç duyar. Bu tezde veri ekosistemi iki ana amaca hizmet edecek şekilde kurgulanmıştır: (i) Türkçe LLM'lerin genel yeteneklerini kültürel olarak uygun bir ölçekte değerlendirmek, (ii) sağlık gibi uzmanlık gerektiren alanlarda güvenilir uyarlama geliştirmeyi mümkün kılmak. Buna ek olarak, duygu analizi ve kültürel metinler gibi tamamlayıcı kaynaklar; hem değerlendirme kapsamını genişletmek hem de farklı veri türleri üzerinde model davranışını gözlemlemek için kullanılmıştır.

Bu bölümde yalnızca veri setlerinin ne olduğu değil, nasıl üretildiği, hangi amaçlarla kullanıldığı ve hangi lisans/erişim koşullarında yayımlandığı da özetlenir. Böylece izleyen yöntem ve deney bölümlerinde kullanılan veri çıktı paketlerinin kapsamı ile sınırlılıkları açık hale gelir. Embedding tarafında, öğretmen vektörlerinin önceden hesaplandığı büyük ölçekli çok dilli Wikipedia korpusu da veri ekosisteminin yeni bir bileşeni olarak tanımlanır.

4.1 TR-MMLU Ekosistemi: Türkçe Çok Alanlı Değerlendirme

TR-MMLU, Türkçe büyük dil modellerinin çok alanlı bilgi, muhakeme ve kavramsal anlama performansını ölçmek üzere tasarlanmış bir değerlendirme çerçevesidir [237], [16]. Bu tezde TR-MMLU iki katmanlı bir yapı olarak ele alınır: geniş bir *soru havuzu* ve bu havuzdan seçilmiş, standartlaştırılmış bir *karşılaştırıcı* alt kümesi. Zenodo üzerinde yayımlanan kaynak, ulusal sınavlar ve açık öğretim içeriklerinden türetilmiş yaklaşık **280.000** çoktan seçmeli sorudan oluşan kapsamlı bir soru havuzunu raporlar; bu havuz **67 bölüm** ile **800+ konu** çeşitliliğini kapsayacak biçimde yapılandırılmaktadır [238]. Veri şeması, çoktan seçmeli formatı destekleyecek şekilde `bölüm`, `konu`, `soru`, `cevap`, `aciklama`

ve `secenekler` alanlarını içerir [238]. Zenodo kaydında ayrıca `cevap` sütununa ilişkin bir hatanın sonraki sürümlerde düzeltildiğine dair not bulunmaktadır; bu nedenle yeniden üretilebilirlik açısından kullanılan sürüm bilgisinin raporlanması önemlidir [238].

Bu geniş havuzdan, değerlendirme standardı olarak kullanılan TR-MMLU karşılaştırıcısı oluşturulmuştur. TR-MMLU karşılaştırıcısı, **62 bölüm/kategori** altında **6.200** çoktan seçmeli sorudan oluşur ve her soru beş seçenekli (A–E) tek doğru cevap formatındadır [237]. Bu yapı, modellerin yalnızca dilsel akıcılık değil, bilgiye dayalı çıkarım ve disiplinler arası genelleme yeteneğini de ölçmeyi hedefler. Tezde raporlanan model karşılaştırmaları ve liderlik tablosu çıktıları, TR-MMLU'yu bir veri setinin ötesine taşıyarak, değerlendirme çıktılarının şeffaf biçimde izlenebildiği bir ekosistem haline getirmektedir [4]. Liderlik tablosu 05 Şubat 2026 itibarıyla 85 model değerlendirmesine ulaşmış olup, bu kapsam TR-MMLU'yu Türkiye'de kamuya açık Türkçe MMLU karşılaştırıcıları içinde en yüksek model kapsamına sahip karşılaştırıcı olarak konumlandırmaktadır.

4.2 Alan-Özel Veri Setleri: Sağlık

Türkçe Tıbbi Soru–Cevap Veri Seti, DoktorSitesi.com kaynaklı hasta–doktor etkileşimlerinden türetilmiş, Türkçe sağlık dili için büyük ölçekli bir **soru–cevap** koleksiyonudur [239]. Zenodo kaydında veri setinin **167.732** soru–cevap çifti içerdiği belirtilmektedir [239]. Veri seti aynı zamanda Hugging Face üzerinde **384.064** satır olarak yayımlanmış olup beş bölüme ayrılmıştır: eğitim (~150K), doğrulama, test, doktor-bazlı test ve uzmanlık-bazlı test [240].

4.2.1 Veri Şeması

Her kayıt dört temel alan içerir: yanıtlayıcı hekimin unvanını belirten `doctor_title` (Uzm. Dr., Prof. Dr., Doç. Dr. vb.; toplamda **36** farklı unvan sınıfı), hekimin uzmanlık alanını belirten `doctor_speciality` (Dermatoloji, Kadın Hastalıkları ve Doğum, Üroloji vb.; toplamda **116** farklı uzmanlık sınıfı), hastanın sağlık sorusunun tam metnini içeren `question_content` ve hekimin yanıtının tam metnini içeren `question_answer`.

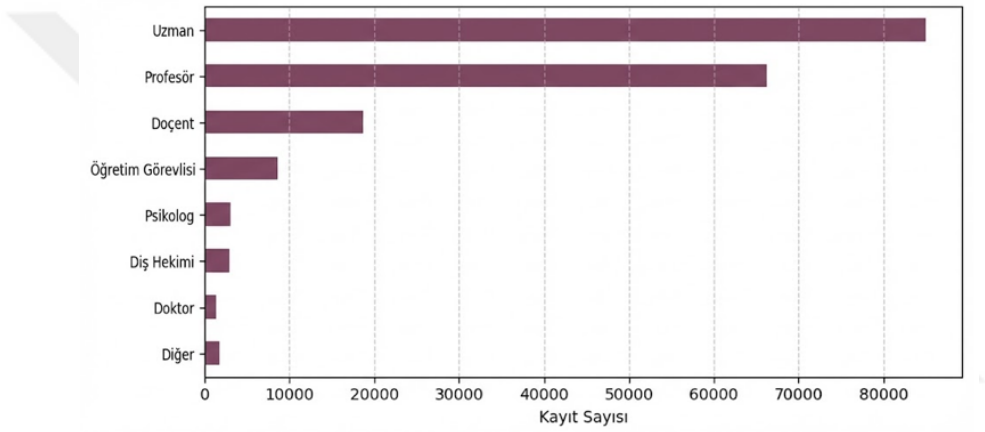
4.2.2 Veri Toplama ve Temizlik

Veri, DoktorSitesi.com platformundaki kamuya açık hasta–hekim etkileşimlerinden otomatik olarak toplanmıştır. Toplama sürecinde HTML biçimlendirme kalıntıları, yinelenen kayıtlar ve yapısal bozukluklar temizlenmiştir. Kişisel tanımlayıcı

bilgilerin (ad, iletişim bilgisi vb.) çıkarılmasına özen gösterilmiş; yalnızca tıbbi içerik ve hekim nitelikleri korunmuştur. Veri seti herkese açık etkileşimlerden oluşması nedeniyle açık lisansla yayımlanmıştır; güncel lisans bilgisi Zenodo ve Hugging Face veri kartlarında yer almaktadır [239].

4.2.3 Unvan ve Uzmanlık Dağılımı

Veri setindeki hekim unvanlarının dağılımı, yanıtların büyük çoğunluğunun uzman hekimler (*Uzm. Dr.*) tarafından sağlandığını, profesör ve doçent unvanlı hekimlerin ise daha düşük oranda temsil edildiğini göstermektedir (Şekil 4.1). Bu dağılım, tez kapsamında kavramsal olarak tartışılan DQ-ALR çerçevesinde (Bölüm 5) veri heterojenliğinin ve kaynak güvenilirliğinin önemini somutlaştırır.

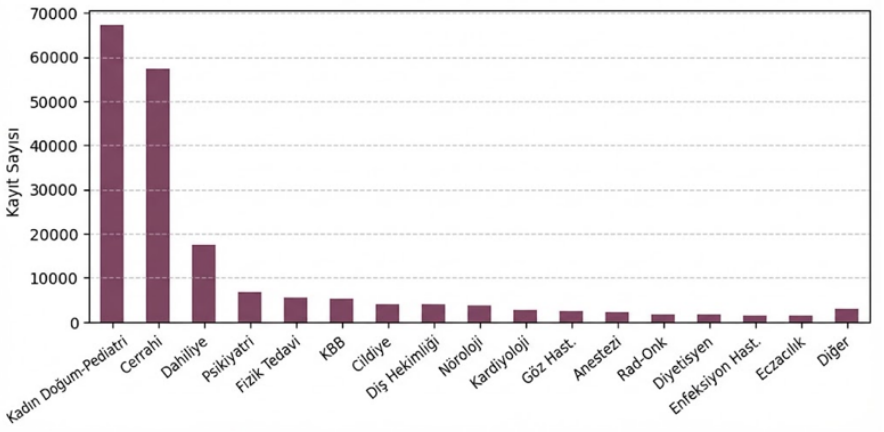


Şekil 4.1 DoktorSitesi veri setindeki sağlayıcı unvan dağılımı.

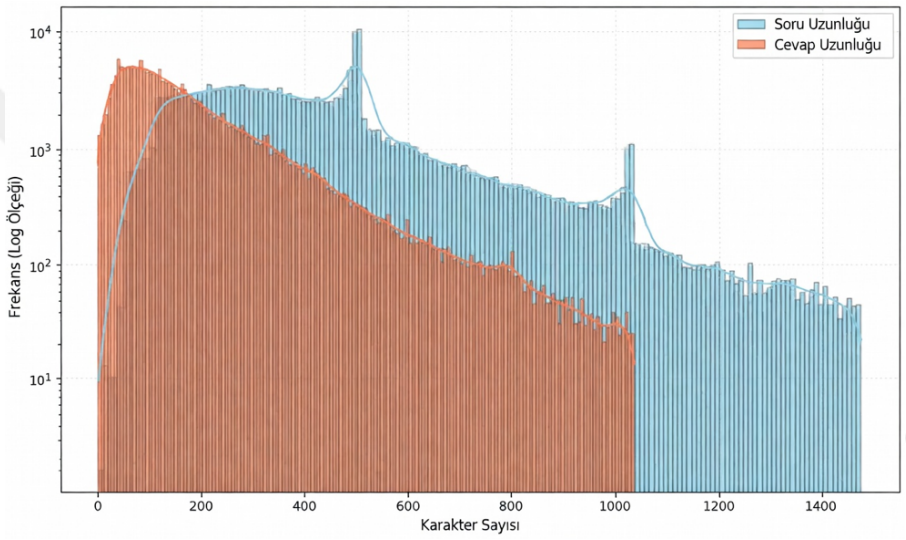
Uzmanlık alanı dağılımı ise veri setinin geniş bir tıbbi yelpaze kapsadığını; ancak bazı alanların (örneğin dermatoloji, kadın hastalıkları) belirgin biçimde daha yoğun temsil edildiğini ortaya koyar (Şekil 4.2). Bu heterojenlik, alan-özel model uyarlamasında çeşitlilik ve dengesizlik arasındaki ödünleşimi somutlaştırır.

4.2.4 Soru ve Yanıt Uzunluk Dağılımı

Soru ve yanıt metinlerinin uzunluk dağılımı, hasta sorularının tipik olarak birkaç cümleden oluştuğunu, hekim yanıtlarının ise daha değişken ve ortalama olarak daha uzun olduğunu gösterir (Şekil 4.3). Bu dağılım, örneklerin bilgi yoğunluğu ve metin tutarlılığı açısından heterojenliğini nicel olarak somutlaştırır.



Şekil 4.2 DoktorSitesi veri setindeki uzmanlık alanı dağılımı.



Şekil 4.3 DoktorSitesi veri setindeki soru ve yanıt uzunluk dağılımı.

4.2.5 Kullanım Amacı

Bu veri seti, tez kapsamında DoctorLlama ve DoctorGemma gibi alan-özel modellerin LoRA ile ince ayarlanmasında doğrudan kullanılmıştır [241], [6]. Hekim unvanı ve uzmanlık alanı gibi meta-bilgiler, alan-özel uyarlama sürecinde bağlam bilgisi olarak değerlendirilmiştir.

4.3 Kontrollü Tokenizasyon Deneyi İçin Birleşik Veri Seti

Tokenizasyonun alt görevlerdeki etkisini, ön-eğitim geçmişi ve veri karışımı gibi dış etkenlerden ayrıştırabilmek için tezde yalnızca tokenizer değişkeninin ayrıştırıldığı kontrollü bir deney hattı kurulmuştur (Bölüm 5.8). Bu hattın veri ayağında, öğretmen embedding'lerinin önceden hesaplandığı bir Türkçe metin korpusu alt kümesi temel alınmış [242] ve bunun üzerinden dört tokenizer akışını aynı örnek üzerinde birlikte saklayan birleşik bir veri seti üretilmiştir

[243]. Birleşik veri seti, aynı örnek için dört farklı tokenizer akışına ait token kimlik dizilerini ayrı sütunlarda saklar: `mft_input_ids`, `tabi_input_ids`, `cosmos_input_ids` ve `newmindaiMursit_input_ids`. Böylece dört kontrol modelinin eğitimi, aynı metin içeriği ve aynı öğretmen hedefleri üzerinde yürütülebilmiş; ayrıca bağlam penceresi sınırı aşıldığında örneklerin tüm akışlardan birlikte elenmesiyle kırpma kaynaklı içerik farklarının kontrollü biçimde engellenmesi sağlanmıştır.

4.4 Tamamlayıcı Veri Setleri: Duygu, Teyit ve Kültürel Metinler

Duygu analizi için geliştirilen koleksiyon, sosyal medya ve e-ticaret platformlarındaki kullanıcı yorumlarını bir araya getirerek pozitif/negatif/nötr duygusal ton sınıflandırmasına yönelik büyük ölçekli bir kaynak sunar. Tez kapsamında veri; farklı platformlardan toplanan milyonlarca yorumu kapsayacak biçimde genişletilmiş, dilsel temizlik uygulanmış ve otomatik etiketleme çıktıları sınırlı bir örneklem üzerinde manuel kontrolle doğrulanmıştır. Bu veri seti, Türkçede duygu sınıflandırması ve kültürel bağlamda tutum analizi gibi görevler için yardımcı bir eğitim/değerlendirme zemini sağlar.

Teyit (fact-checking) veri seti, doğrulama platformlarının yayımladığı iddia–değerlendirme içeriklerinden türetilmiş örnekleri içerir. Tez kapsamında; iddia metni, doğrulama açıklaması ve doğruluk etiketi (doğru/yanlış/yanıltıcı/doğrulanamaz gibi) bilgilerini koruyacak şekilde yapılandırılmıştır. Bu kaynak, dezenformasyon tespiti, güvenilirlik sınıflandırması ve üretken modellerde güvenlik/etik yanıt davranışlarının test edilmesi gibi senaryolarda kullanılmaktadır.

Kültürel metin koleksiyonları iki tamamlayıcı kaynaktan oluşur. Birincisi, masal ve anlatı geleneğini temsil eden ve on binlerce cümleyi kapsayan derlem; deyim, atasözü ve halk hikâyesi gibi kültürel referansları içeren metinleri bir araya getirir. İkincisi ise, tarihsel olayları kısa açıklamalarla sunan ve on binler ölçeğinde kayıt içeren zaman-odaklı bir koleksiyondur. Bu iki kaynak, üretken modellerin kültürel bağlam farkındalığını artırmaya ve farklı bilgi türlerinin model davranışı üzerindeki etkisini gözlemlemeye yöneliktir.

4.5 Embedding Model Eğitim Kaynakları

Tezde geliştirilen uzun bağlamlı Türkçe cümle gömme modeli (embeddingmagibu-200m), iki ana veri kaynağına dayanır: tokenizer eğitimi

için geniş ölçekli Türkçe bir korpus ve çevrimdışı damıtma için önceden hesaplanmış öğretmen embedding'leri içeren bir örnek kümesi.

Cosmos Turkish Corpus v1.0, Cosmos AI Research Group tarafından yayımlanan ve Türkçe sürekli ön-eğitim senaryolarında kullanılmak üzere derlenmiş bir ön-eğitim korpusudur [244]. Veri kartında; korpusun yaklaşık **15 milyar token** ölçeğinde olduğu, **9.075.453** örnek içerdiği ve `url` ile `text` alanlarından oluştuğu belirtilmektedir. URL tabanlı deduplikasyon uygulanmış olmakla birlikte, kullanım senaryosuna göre ek filtreleme gerekebileceği vurgulanmaktadır. Bu tezde Cosmos korpusu, tokenizer eğitimi ve embedding modelinin veri çeşitliliğini artıran temel metin kaynağı olarak kullanılmıştır (Bölüm 5).

embeddingmagibu-200m için SentencePiece-BPE tokenizer, bu korpus üzerinde 128K sözlük boyutuyla eğitilmiş ve daha sonra EmbeddingGemma-300M omurgasının tokenizer uyarlamasında kullanılmıştır [245], [246].

Cosmos Corpus with Precomputed Embeddings veri seti, öğretmen modelin her adımda yeniden çalıştırılmasının maliyetini azaltmak amacıyla embedding vektörlerinin önceden hesaplanarak yayımlanmış hâlidir [242]. Hugging Face veri kartında bu kaynağın **280.000** örnekten oluştuğu ve `url`, `text` ile iki ayrı embedding alanını (`teacher_embedding_final`, `teacher_embedding_pre_dense`) içerdiği raporlanmaktadır. Bu yaklaşım, çevrimdışı damıtma eğitimini hızlandırırken, deneylerin yeniden üretilebilirliğini de güçlendirir [247].

Wikipedia-40 Languages (embedding öncesi) veri seti, 40 dilde Wikipedia makalelerinden örnekler toplanarak çok dilli ve dengeli bir eğitim gövdesi oluşturmayı amaçlar [248]. Türkçe ve İngilizce için 100K/30K/10K, diğer 38 dil için 10K/3K/1K (train/validation/test) kota hedeflenmiş; Wikimedia akış üzerinden toplanan örnek sayısı hedefin altında kaldığında dil başına oranlı düzeltme uygulanmıştır. Her örnek `lang`, `title`, `text` ve `url` alanlarını içerir. Bu korpus, çok dilli kapsama ve dil dengesi açısından öğretmen embedding üretimi için temel kaynak olarak kullanılmıştır.

Wikipedia-40 Languages with Embeddings veri seti, yukarıdaki korpusun EmbeddingGemma-300M öğretmen modeli ile embedding'lenmiş sürümüdür [249]. Üretim süreci, büyük ölçek nedeniyle parça parça push edilen bir akışla yürütülmüş ve sonuç tek bir eğitim bölümü altında yaklaşık 580K satır olarak yayımlanmıştır. Şema `lang`, `title`, `text`, `url` alanlarına ek olarak iki embedding sütunu içerir: `teacher_embedding_final` ve `teacher_embedding_pre_dense`. Embedding boyutu 768 olup, veri Parquet

formatında optimize edilmiştir. Bu veri seti, çevrimdışı damıtma hedeflerinin ölçeklenebilir biçimde üretilmesini sağlayarak embeddingmagibu-200m eğitim hattını desteklemiştir.

4.6 Embedding Değerlendirme Veri Setleri: STSbTR, TR-MTEB ve TabiBench

Embeddingmagibu-200m'in değerlendirmesi bu revizyonda iki ana protokol üzerinden yürütülmüştür: (i) tek görevli semantik benzerlik ölçümü (STSbTR), (ii) çok görevli embedding karşılaştırıcısı (TR-MTEB). Tezde raporlanan resmi performans değerleri bu iki değerlendirme protokolünün çıktıları esas alınarak sunulmuştur.

STSbTR, İngilizce STS veri setinin makine çevirisi ile oluşturulmuş Türkçe bir anlamsal benzerlik veri setidir [123], [250]. Veri kartı bölüm bilgilerine göre eğitim/doğrulama/test bölümleri sırasıyla **5.749**, **1.500** ve **1.379** örnek içermektedir. Bu revizyonda STS raporlaması hem eğitim hem test split'lerini birlikte içerecek şekilde yapılmıştır (Bölüm 6).

TR-MTEB, Türkçe cümle gömme modelleri için tasarlanmış çok görevli bir değerlendirme setidir [90], [91]. TR-MTEB değerlendirmesi **26 görev** üzerinde raporlanmıştır. Bu görevler; getirim, sınıflandırma, kümeleme, STS, NLI ve bitext mining ailelerini tek bir protokol altında birleştirerek modelin görevler arası genellenebilirliğini daha şeffaf biçimde karşılaştırmayı mümkün kılar. Tüm görev düzeyi ayrıntılar ve güncel karşılaştırma çıktıları için okuyucular ayrıca kamuya açık TR-MTEB sonuç gezginine yönlendirilmektedir [251].

TabiBench, TabiBERT çalışmasıyla birlikte yayımlanan tamamlayıcı bir değerlendirme setidir [252]. Bu revizyonda TabiBench sonuçları ana performans iddiaları için kullanılmamış; ikincil/supplementary karşılaştırma olarak eklerde korunmuştur (Bölüm 6, Ek A).

4.7 Veri Yönetişimi, Etik ve Lisanslama

Tez boyunca veri toplama ve yayımlama süreçleri; kamuya açık kaynak kullanımı, kişisel verilerin korunması, izlenebilirlik ve yeniden üretilebilirlik ilkelerine göre yürütülmüştür. Sağlık gibi hassas alanlarda, veri kaynakları yalnızca kamusal etkileşimler ve açık arşivlerden seçilmiş; kişisel tanımlayıcı bilgiler çıkarılmış ve verinin kullanıma uygunluğu temkinli bir filtreleme yaklaşımıyla değerlendirilmiştir. TR-MMLU soru havuzu ve Türkçe tıbbi soru–cevap veri

seti, Zenodo üzerinde Digital Object Identifier (DOI) ile yayımlanmıştır [238], [239]. Tüm veri kaynaklarının lisans ve erişim koşulları, ilgili platformlardaki (Zenodo, Hugging Face) veri kartları üzerinden izlenebilir ve doğrulanabilir biçimde sağlanmıştır.

4.8 Bölüm Özeti

Bu bölümde, tezdeki yöntem ve deneylerin dayandığı veri ekosistemi; TR-MMLU değerlendirme altyapısı, sağlık alanına yönelik alan-özel kaynaklar, tamamlayıcı toplumsal/kültürel veri setleri ve embedding eğitimi için kullanılan korpuslar bağlamında sunulmuştur. Cosmos korpusu ve precomputed embedding alt kümeleriyle birlikte, çok dilli Wikipedia-40 korpusu ve embedding'li sürümü de çevrimdışı damıtma hattının veri ayağını tamamlayıcı şekilde tanımlanmıştır. Bu bütüncül görünüm, Türkçe LLM geliştirme sürecinde değerlendirme standardizasyonu ile alan uyarlamasını tek bir hat üzerinde birleştirmeyi ve veri çıktı paketlerinin izlenebilirliğini güçlendirmeyi mümkün kılmaktadır.

5
YÖNTEM VE ÖNERİLEN YAKLAŞIMLAR

Bu bölüm, tez kapsamında geliştirilen yöntemsel çerçeveyi uçtan uca tanımlar. Tezin temel iddiası, Türkçe gibi biçimbilimsel olarak zengin ve görece düşük kaynaklı dillerde LLM'lerin başarımını artırmanın yalnızca daha büyük model veya daha fazla veriyle değil; dilbilimsel bütünlüğü koruyan tokenizasyon, veri kalitesini doğrudan optimizasyona yansıtan öğrenme stratejileri ve alan-özel uyarlamayı maliyet-etkin biçimde mümkün kılan adaptasyon yaklaşımlarıyla birlikte ele alınmasıyla mümkün olduğudur. Bu yaklaşım, TR-MMLU ekosistemini temel değerlendirme standardı olarak kullanırken, sağlık gibi uzmanlık gerektiren alanlarda güvenilirlik ve aktarılabilirlik hedeflerini aynı çatı altında birleştirir [237], [6].

Yöntemsel çerçeve, veri hazırlama ve temsil öğrenmeden başlayıp değerlendirme protokollerine kadar uzanan bir hattı kapsar. Böylece tokenizasyon ve veri kalitesi gibi "girdi" düzeyindeki kararların, alan-özel uyarlama ve embedding damıtma gibi "model içi" süreçlere; oradan da TR-MMLU, STSbTR ve TR-MTEB gibi "çıktı" düzeyindeki ölçümlere nasıl yansıdığı adım adım izlenebilir hale gelir. Bölüm boyunca amaç, her bileşenin tek başına katkısını göstermekten ziyade, bu bileşenlerin birlikte çalıştığında ortaya çıkan sistematik etkiyi açık ve tekrarlanabilir biçimde görünür kılmaktır.

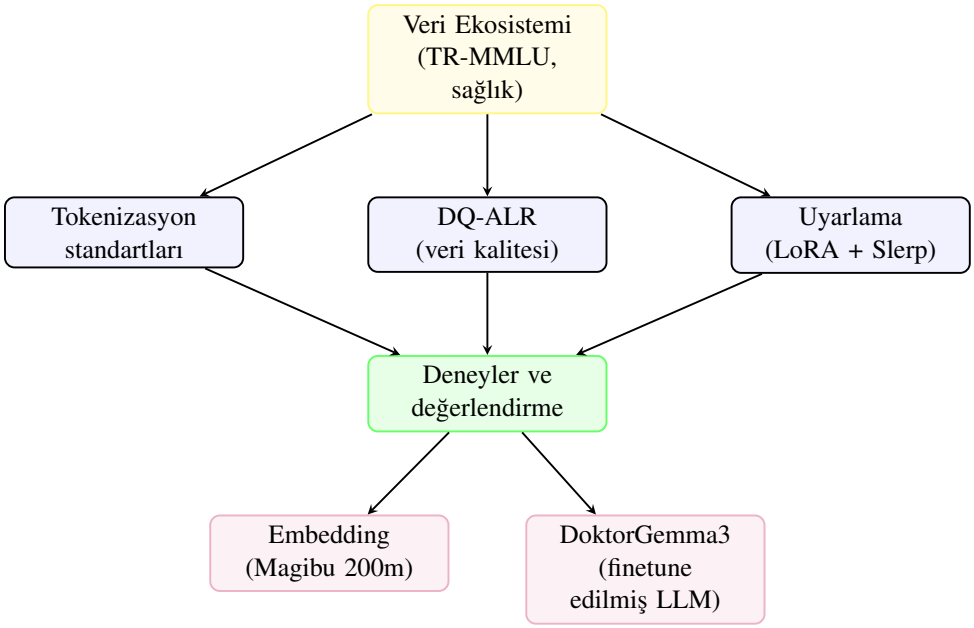
Bu nedenle bölüm, yalnızca yöntem tanımı sunmakla kalmaz; her yöntemin hangi veri varsayımlarıyla çalıştığını, hangi ölçütlerle doğrulandığını ve hangi deneysel kararları gerektirdiğini de açıklar. Böylece metodoloji, model tasarımını veri ve değerlendirme boyutlarıyla birlikte ele alan bütüncül bir hat olarak konumlandırılır.

5.1 Hipotezlerle İlişki

Bu bölüm, Giriş'te tanımlanan hipotezlerin (**H1-H5**) yöntemsel karşılığını kurar (bkz. Bölüm 1.3). Tokenizasyon standartları ve MFT tasarımı, morfolojik farkındalıklı tokenizasyonun temsili güçlendireceği (**H1**) ve %TR / %Pure

gibi metriklerin tokenizasyon kalitesi için öngörücü sinyaller sunduğu (**H2**) iddialarını test edilebilir hale getirecek ölçüm ve tasarım kararlarını tanımlar. DQ-ALR, veri kalitesi sinyalinin öğrenme dinamiklerinde dikkate alınmasının model davranışını etkileyebileceği (**H3**) savını kavramsal olarak tartışır. LoRA ve Slerp ise alan-özel uyarlamada maliyet-etkin kazanım sağlanırken genel yeteneğin korunabileceği (**H4**) iddiasını sınamak üzere kullanılan adaptasyon ailesini tanımlar; TR-MMLU'nun yerel ve kültürel olarak özgün bir karşılaştırıcı olarak kullanımı da Türkçe modeller arasındaki farkların görünür kılınmasına ilişkin **H5** tartışmasını deneysel olarak görünür kılacak protokol kararlarını destekler.

Şekil 5.1 tezin yöntem bileşenleri arasındaki akışı özetlemektedir. Öncelikle, tokenizasyonun Türkçe üzerindeki dilbilimsel etkisi nicel olarak ölçülür ve %TR ile %Pure metrikleri üzerinden standartlaştırılır [3]. Bu değerlendirme, MFT tasarımının hem hedefini hem de doğrulama ölçütlerini belirler [6]. İkinci olarak, heterojen kaynaklardan gelen örneklerin güvenilirliğinin değişken olduğu kabul edilerek, DQ-ALR yaklaşımı kavramsal bir çerçeve olarak önerilir; bu tezdeki tartışmada özellikle tıbbi metinlerde sağlayıcı uzmanlığı gibi sinyallerin potansiyel kullanımı ele alınır [239]. Üçüncü olarak, alan-özel uyarlamada LoRA gibi PEFT yöntemleri ile uzmanlık kazanımı düşük maliyetle elde edilir; ayrıca Slerp gibi ağırlık uzayı birleştirme stratejileri ile uzmanlık performansı korunurken genel dil yeteneğinin geri kazanılması hedeflenir [30]. Son olarak, cümle gömme tarafında embeddingmagibu-200m modeli; tokenizer eğitimi, ağırlık koruyan klonlama ve çevrimdışı damıtma ile yeniden eğitim gerektirmeden Türkçeye uyarlanır [245].



Şekil 5.1 Tezde önerilen yöntemsel çerçeve.

5.2 Tokenizasyon Standartları ve Değerlendirme Yaklaşımı

Tokenizasyon, bir dil modelinin gördüğü sembol dizisini ve dolayısıyla hem hesaplama maliyetini hem de öğrenilen temsillerin dilbilimsel tutarlılığını belirler. Modern LLM'lerde yaygın olan alt-birim temelli yöntemler (BPE; WordPiece; SentencePiece), İngilizce gibi görece yalın morfolojiye sahip dillerde iyi bir verimlilik–kapsam dengesi kurarken; Türkçe gibi eklemeli dillerde aynı kelime kökünden üretilen geniş biçimsel çeşitlilik nedeniyle anlamlı morfem sınırlarını sıklıkla ihlal edebilir [9], [253], [254]. Bu tür parçalanmalar, hem bağlam penceresinin daha hızlı tüketilmesine hem de kök–ek ilişkilerinin temsilde izlenmesini zorlaştıran bir "semantik kırılma"ya yol açar [3].

Bu tez, tokenizasyonu yalnızca "sık görülen parçaları öğrenme" problemi olarak değil; aynı zamanda "dilsel bütünlüğü koruyarak temsil verimliliğini artırma" problemi olarak ele alır. Bu çerçevede iki ölçüt öne çıkar: hedef dildeki morfolojik birimlerle uyumu yakalayan %TR ve tokenların anlamsal olarak bölünemez birimlere ne ölçüde karşılık geldiğini yakalayan %Pure. Her iki metrik de tokenizasyonun dilbilimsel yönünü ölçmek için tasarlanmış olup, TR-MMLU gibi kültürel ve dilsel olarak yerel bir karşılaştırıcı üzerinde karşılaştırılabilirlik sağlar [237], [3].

%TR (Language-Specific Token Percentage), tokenizerın ürettiği tokenların ne kadarının Türkçe kök/ek sözlüğü ve işlevsel sözlükle eşleştiğini ölçer. Bu tezde %TR hesabı, token dizisinin her bir öğesinin (i) kök sözlüğü, (ii) ek sözlüğü veya (iii) fonksiyonel birim listesi ile eşleşmesine dayandırılmıştır. Formülasyon aşağıdaki gibidir:

$$\%TR = \frac{N_{\text{valid-TR}}}{N_{\text{total}}} \times 100 \tag{5.1}$$

Burada $N_{\text{valid-TR}}$ Türkçede "geçerli" kabul edilen token sayısını, N_{total} toplam token sayısını temsil eder. Formül 5.1, tokenların dil-özgü sözlükle eşleşme oranını nicel olarak verir [3].

%Pure (Token Purity) ise üretilen tokenların morfem sınırlarına uygunluğunu ve anlamsal bütünlüğünü ölçer. Tokenin, bağımsız bir kök veya dilbilgisel olarak geçerli bir ek birimi olarak yorumlanabilmesi beklenir; anlamsız alt-parçalara bölünme arttıkça %Pure düşer. Bu tezde kullanılan tanım, morfolojik çözümleme çıktılarından türetilen "bölünemez birim" sayısına dayalıdır:

$$\%Pure = \frac{N_{\text{pure}}}{N_{\text{total}}} \times 100 \tag{5.2}$$

Formül 5.2, morfem bütünlüğünü yakalayan %Pure ölçütünü tanımlar. Bu iki

metrik birlikte okunduğunda, yalnızca sözlük uyumu (%TR) değil, aynı zamanda morfem bütünlüğü (%Pure) izlenebilir hale gelir [3].

Ölçümün döngüselliğini (tokenizerın "kendi sözlüğüne göre" başarılı görünmesi) önlemek için, %TR ve %Pure hesaplarında değerlendirilen tokenizerın kendi sözlüğü referans alınmaz; bunun yerine tokenizerdan bağımsız biçimde derlenmiş kök/ek kaynakları ve ayrı bir doğrulama hattı kullanılır. Böylece metrikler, farklı tokenizasyon yaklaşımlarını aynı dilbilimsel referans çerçevesinde karşılaştırmayı hedefler [3].

Metriklerin daha açık bir notasyonla ifade edilebilmesi için, bir veri kümesi $D = \{x_i\}_{i=1}^{n}$ ve bir tokenizasyon fonksiyonu $\tau(\cdot)$ tanımlanabilir. Bu durumda toplam token sayısı $N_{\text{total}} = \sum_{i=1}^{n} |\tau(x_i)|$ olur. Aynı veri kümesi için tokenizerın benzersiz token üretimi, $U = |\bigcup_{i=1}^{n} \tau(x_i)|$ ile, tokenizasyonun bağlam penceresini ne kadar hızlı tükettiğini ifade eden "token ayak izi" ise $F(\tau, D) = \sum_{i=1}^{n} |\tau(x_i)|$ ile izlenebilir. Bu notasyon, %TR ve %Pure gibi dilbilimsel metriklerin, token sayısı ve sözlük kullanımı gibi hesaplamalı metriklerle aynı çerçevede konuşulabilmesini sağlar.

Tokenizasyon standartları, tezde üç pratik ölçüm ailesiyle birlikte kullanılır. Birincisi, metnin model açısından taşıdığı "token ayak izi"dir; bu ayak izinin büyümesi, sabit bağlam penceresinde daha az içerik görülebilmesi ve self-attention maliyetinin artması anlamına gelir. İkincisi, sözlük boyutu ile temsil verimliliği arasındaki dengedir; büyük sözlükler parçalanmayı azaltabilirken embedding tablosu parametrelerini yükseltir, küçük sözlükler ise token sayısını artırabilir. Üçüncüsü, ölçümün kendisinin tekrarlanabilirliğidir; aynı veri kümesi üzerinde aynı normalizasyon kurallarıyla hesaplanmayan metriklerin karşılaştırılabilirliği düşer. Bu nedenle tezde tokenizasyon değerlendirmesi, TR-MMLU metinleri üzerinde sabit bir normalize-edilmiş metin hattı ve aynı sözlük setleri kullanılarak yürütülür [237], [3].

Tokenizasyon standartlarının tezdeki rolü iki yönlüdür. Birinci yön, farklı modellerin tokenizasyon kaynaklı verimlilik eşitsizliklerini görünür kılmak ve Türkçeye daha iyi hizalanmış tokenizasyonun TR-MMLU başarımıyla ilişkisini nicel olarak göstermektir. İkinci yön ise, MFT'nin tasarım hedeflerini ölçülebilir hale getirerek, geliştirilen tokenizasyonun "neden daha iyi" olduğuna dair gerekçeyi sadece nitel örneklerle değil, nicel metriklerle de temellendirmektir. Tokenizasyon standartları çalışmasında, TR-MMLU üzerinde %TR ile başarım arasında güçlü pozitif korelasyon raporlanmaktadır (ör. r=0,90) [3]. Bu bulgu, tezde kullanılan stratejinin, "dil-özgü token kapsaması yükseldikçe Türkçe ölçme görevlerinde

başarımın artması" biçiminde özetlenebilecek ana sezgisini destekler.

5.3 MFT Tokenizasyonu

MFT, istatistiksel alt-birim tokenizasyonun güçlü yanlarını korurken Türkçenin morfolojik ve fonolojik özelliklerini doğrudan tokenizasyon kararına dahil eden hibrit bir yaklaşımdır [6]. Temel hedef, kelime kökünü ve ek dizisini mümkün olduğunca tutarlı biçimde ayırmak; ses olayları nedeniyle oluşan yüzey varyantlarını tek bir kimlik altında birleştirmek; sözlük dışı veya belirsiz vakalarda ise BPE tabanlı bir geri dönüş (fallback) ile kapsama sağlamaktır. Bu yaklaşım, tokenizasyonun dilbilimsel bütünlüğünü artırırken, gereksiz sözlük şişmesini de kontrol altında tutmayı amaçlar.

MFT sözlüğü, farklı bilgi türlerini ayrı kanallarda taşıyacak biçimde yapılandırılmıştır: (i) fonolojik varyantların normal kimliklere eşlenmesiyle derlenen yaklaşık 22K ölçeğinde **kök envanteri**, (ii) 177 farklı yüzey ek biçimini 72 normal ek kimliği altında birleştiren **ek envanteri**, (iii) sözlük dışı durumlar için 12.696 birimlik **BPE tabanlı tamamlayıcı altbirimler** ve (iv) boşluk/tab/noktalama gibi yapısal bilgiyi kayıpsız taşımayı hedefleyen **özel tokenler**. Yazımsal bilgi, sözlüğü büyütmeden korunur: büyük harfli başlangıç gibi durumlar ayrıca bir işaretleyiciyle temsil edilir (örn. `<uppercase>`); boşluklar gibi biçimsel ayrımlar da açık tokenlerle taşınarak encode–decode çevriminde izlenebilirlik hedeflenir [6].

5.3.1 Sözlük Bileşimi ve Kimlik Aralıkları

MFT'nin sözlük tasarımı, "yüzey biçimi sayısı" ile "öğrenilecek kimlik sayısı"nı bilinçli biçimde ayırır: aynı kök veya ek için birden fazla yüzey varyantı (alomorf) bulunabilir, ancak bunlar tek bir normal kimlik altında birleştirilebilir. Bu sayede hem fonolojik çeşitlilik kapsanır hem de embedding tablosu gereksiz yere şişmez. Tablo 5.1 bu bileşimi ve kimlik aralıklarını özetlemektedir [6].

Tablo 5.1 MFT sözlük bileşimi ve kimlik aralıkları.

Bileşen	**Girdi sayısı**	**Kimlik sayısı**	**Kimlik aralığı**
Kökler	22.231	20.000	0-19.999
Ekler (yüzey biçimleri)	177	72	20.000-20.071
BPE altbirimleri	12.696	12.696	20.072-32.767
Toplam	–	**32.768**	–

MFT'nin ilk bileşeni fonolojik normalizasyondur. Türkçede aynı ekin farklı yüzey

biçimleri (ör. *-lar*/*-ler*, *-dan*/*-den*, *-tan*/*-ten*) ünlü uyumu ve sertleşme/yumuşama gibi kurallar nedeniyle ortaya çıkar. Benzer biçimde, köklerde ünsüz yumuşaması veya son ünsüz düşmesi gibi süreçler, kökün yüzey formunu değiştirir (ör. *kitap* → *kitabı*). MFT bu varyantları, "aynı morfemin farklı alomorfları" olarak ele alır ve ortak kimliğe eşler. Böylece, sözlükte aynı morfemin farklı yüzey biçimleri için ayrı tokenlar tutma ihtiyacı azalır; hem kapsama hem de parametre verimliliği açısından daha tutarlı bir yapı oluşur.

İkinci bileşen sözlük güdümlü morfolojik ayrıştırmadır. Tokenizasyon, serbest metni önce "segment" düzeyinde ele alır; boşluk, noktalama ve özel işaretlerin ayrıştırılması, morfolojik kararlarla karışmayacak şekilde normalizasyon aşamasında yapılır. Ardından, her kelime segmenti için kök tespiti iki seviyeli bir stratejiyle yürütülür: kelime doğrudan kök sözlüğünde bulunuyorsa tek adımda kök kimliği üretilir; bulunmuyorsa en uzun eşleşen kök adayını bulmak için kelime üzerinde soldan sağa ilerleyen bir arama yapılır. Bu aşamanın amacı, frekans temelli parçalama yerine dilbilgisel olarak geçerli bir kök–ek dizisi üretmektir.

Kök bulunduğunda ek ayrıştırması, kalan parçanın ek sözlüğüyle iteratif olarak eşleştirilmesiyle devam eder. Bu süreçte fonolojik normalizasyonun rolü kritiktir; eklerin yüzey biçimleri bağlama göre değiştiği için, eşleştirme "yüzey form" üzerinden değil "normal morfem" üzerinden yürütülür. Kök–ek eşleşmesi başarısız olduğunda veya belirsizlik oluştuğunda, kalan parça istatistiksel alt-birim mekanizmasına aktarılır. Böylece MFT, tamamen kural tabanlı yaklaşımların "sözlük dışı" zafiyetini azaltırken, saf BPE/WordPiece türü yaklaşımların "anlamsız parçalanma" eğilimini de sınırlar [6].

Üçüncü bileşen, yazımsal bilgiyi sözlük şişirmeden temsil etmektir. Büyük/küçük harf farkları Türkçede özel isimler ve cümle başı gibi durumlar için anlamlıdır; ancak her varyantı ayrı token yapmak sözlüğü büyütür. Bu nedenle MFT, harf bilgisini ayrı bir işaretleyici üzerinden taşır ve yüzey formdaki kapitalizasyonu ayrı bir token ile temsil eder (örn. `<uppercase>`) [6].Boşluk gibi yapısal ayrımlar da açık biçimde temsil edilerek encode–decode çevriminde kayıpsızlık hedeflenir.

Bu tasarımın pratikte çalışabilmesi, iki sözlük bileşeninin kalitesine bağlıdır: kök sözlüğü ve ek sözlüğü. Hibrit tokenizasyon yaklaşımında ek sözlüğü, Türkçede üretken biçimde kullanılan ekleri ve bazı işlevsel parçaları kapsayacak şekilde, fonolojik varyantları tek bir normal kimlik altında birleştirecek biçimde hazırlanır; böylece örneğin ünlü uyumu kaynaklı alomorflar ayrı tokenlar olarak şişmez [6]. Sözlük dışı kapsam için kullanılan BPE tamamlayıcı bir mekanizmadır; amaç, sözlük tabanlı kararların başarısız olduğu yerde kapsama sağlamak, ancak

mümkün olduğunca morfem sınırlarını "kırmayan" bir parçalamayı tercih etmektir. Bu nedenle fallback yalnızca kök–ek kapanışı yapılamadığında devreye girer ve tokenizasyonun ana yolu haline getirilmez.

Algoritma 1, MFT yaklaşımının kelime düzeyindeki karar akışını göstermektedir. Süreç, girdi metnin taranmasıyla başlar ve metin, kelime segmentleri ile özel segmentlere ayrılır. Boşluklar, noktalama işaretleri, sayılar ve kontrol karakterleri gibi özel segmentler herhangi bir biçimbilimsel işleme tabi tutulmadan doğrudan token dizisine eklenir. Bu sayede metnin yüzeysel yapısı kayıpsız biçimde korunur.

Kelime segmentleri için ilk olarak bir ön işleme aşaması uygulanır. Bu aşamada büyük harfle başlayan sözcükler tespit edilerek bu bilgi ayrı bir `<uppercase>` tokeni ile kodlanır. Ardından Türkçeye özgü küçük harfe dönüştürme işlemi gerçekleştirilir; bu dönüşüm sırasında $\dot{I} \rightarrow i$ ve $I \rightarrow \imath$ gibi dilsel olarak kritik eşlemeler korunur.

Ön işleme tamamlandıktan sonra kelime, kök sözlüğü üzerinden biçimbilimsel çözümlemeye tabi tutulur. İlk olarak doğrudan bir kök eşleşmesi aranır; eşleşme bulunamazsa kelimenin başından itibaren en uzun olası kök adayını bulmaya yönelik açgözlü bir arama gerçekleştirilir. Geçerli bir kök bulunduğunda, kelimenin geri kalan kısmı için ek dizisi ayrıştırılmaya çalışılır. Eğer ek dizisi çıkarılamazsa ya da kalan parça anlamlı bir kök veya ek kombinasyonu oluşturacak şekilde kapanmıyorsa, sistem yedek mekanizma olarak BPE modeline başvurur. Bu yaklaşım, "önce dilbilimsel karar, ardından istatistiksel kapsama" ilkesini sistematik bir çerçeveye oturtur [6].

Algoritma 1'deki bu akış, uygulamada önce morfolojik ayrıştırmayı, yalnızca gerekli olduğunda BPE yedek mekanizmasını devreye alacak şekilde işler.

5.3.2 Kodlama (Encode) Kararı

MFT'nin kodlama aşaması, kelimeyi tek bir ham BPE akışına bırakmak yerine kökler, ekler ve tamamlayıcı BPE alt birimleri arasında kontrollü bir seçim problemi olarak ele alır. Tüm sözlük bileşenleri Trie veri yapısı içinde tutulur ve her karakter konumunda tüm olası ön-ek eşleşmeleri verimli biçimde elde edilir. Bu yaklaşım, kelime uzunluğu ile pratikte doğrusal maliyetli bir tarama sağlamayı hedefler [6].

Kodlama sürecinde yazımsal işaretleme kritik bir rol oynar. Büyük harf bilgisi ayrı bir tokenle kodlanarak kök ve ek temsillerinin yüzey biçimlerden bağımsızlaştırılması sağlanır. Kelime sınırları, kök tokenlerinde baştaki boşluğun açık biçimde temsil edilmesi yoluyla korunur ve böylece token dizisinden yüzey

Algorithm 1 Morfoloji-Öncelikli Tokenizasyon (Encode) İş Hattı

Girdi: Ham metin dizgesi S
Çıktı: Token Kimlikleri (Token ID) dizisi T
$S \leftarrow$ `on_isleme`(S) ▷ Boşluk ekleme
for her bir kelime w için S içinde **do**
    if w boşluk veya noktalama işareti ise **then**
        `T.ekle(ozel_id_getir(w))`
        devam et
    end if
    if w Büyük Harfle Başlıyorsa **then**
        `T.ekle(ID_BUYUK_HARF)`
        $w \leftarrow$ `kucuk_harfe_cevir`(w)
    end if
    `kok, ekler <- bicimbilimsel_cozumle(w)` ▷ Açgözlü sözlük araması
    if $kok \neq$ `YOK` **then**
        `T.ekle(kok.id)`
        for her bir e için $ekler$ içinde **do**
            `T.ekle(e.id)`
        end for
    else
        `alt_birimler <- bpe(w)` ▷ Yedek mekanizma
        `T.genislet(alt_birimler)`
    end if
end for
Döndür `T`

metne geri dönüşte ek bir boşluk tokenine ihtiyaç duyulmaz.

Aynı konumda birden fazla aday eşleşme bulunduğunda seçim, eşleşen karakter sayısını önceleyen bir uzunluk puanı ve tür önceliği birlikte değerlendirilerek yapılır. Eşitlik durumlarında kökler BPE alt birimlerine, BPE alt birimleri ise eklere göre öncelik kazanır. Ayrıca kalan parçanın en az iki karakterlik geçerli bir ekle devam edebilmesi gibi ileriye dönük sinyaller, belirsiz durumlarda daha tutarlı morfem kapanışları üretmek amacıyla kullanılabilir [6].

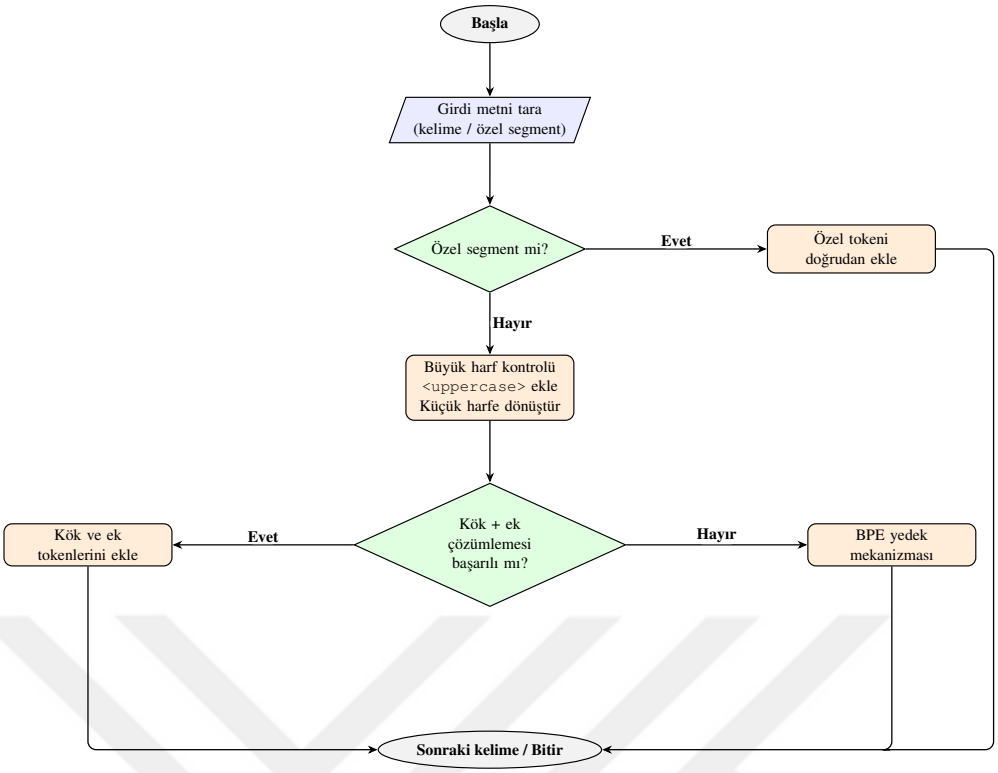
5.3.3 Çözme (Decode) ve Fonolojik Kurallar

MFT yalnızca kodlama değil, aynı zamanda güvenilir bir çözme aşamasını da tanımlar. Ekler kanonik biçimde temsil edildiği için çözme işlemi basit bir birleştirme değildir. Algoritma 2 içinde verilen çözücü iş hattı, token dizisini yüzey metne geri dönüştürürken bağlama duyarlı fonolojik kuralları uygular ve bu sayede doğru yüzey biçimleri yeniden üretilir.

Algorithm 2 MFT Çözme (Decode) İş Hattı

Girdi: Token Kimlikleri dizisi T
Çıktı: Yeniden oluşturulmuş metin dizgesi S
`parcalar <- []; i <- 0`
while $i < |T|$ **do**
    `id <- T[i]`
    if $id =$ `ID_BUYUK_HARF` **then**
        `buyuk <- turkce_buyuk_harf_cevir(bak(T[i+1]))`
        `parcalar.ekle(buyuk)`
        $i \leftarrow i + 2$; **devam et**
    end if
    `adaylar <- yuzey_bicimleri(id)`
    if $|$`adaylar`$| > 1$ **then**
        `baglam <- unlu_baglami_bul(parcalar)`
        `yuzey <- fonoloji_uygula(id, baglam, T, i)`
    else
        `yuzey <- adaylar[0]`
    end if
    `parcalar.ekle(yuzey)`
    $i \leftarrow i + 1$
end while
Döndür `birlestir(parcalar)`

Çözme aşamasında uygulanan fonolojik kurallar sistematiktir. Ünlü uyumu, ek ünlüsünün kelimedeki son ünlünün önlük-artlık özelliklerine göre seçilmesini sağlar. Ünsüz benzeşmesi, sert ünsüzlerden sonra $d \rightarrow t$ dönüşümünü tetikler. Ünlüyle başlayan eklerden önce görülen ünsüz yumuşaması $p \rightarrow b$ ve $k \rightarrow ğ$ gibi dönüşümleri kapsar. İlerleyici zaman eki gibi yapılarda kök sonundaki ünlülerin daralması uygulanır. Ünlü-ünlü birleşimlerinde ise bağlama bağlı olarak *y*, *n* veya *s* kaynaştırma ünsüzleri eklenir. Bu mekanizma sayesinde token dizisi ile yüzey metin arasında yüksek doğruluklu ve tersinir bir kodlama-çözme çevrimi elde edilir [6]. Bu tersinirlik, MFT'yi bir tokenizerın ötesinde, üretim odaklı bir Türkçe *metin çözümleyici* olarak konumlandırır.



Şekil 5.2 MFT tokenizasyon sürecinin algoritmik akış diyagramı.

Şekil 5.2, MFT kodlama hattındaki karar noktalarını tek bakışta gösterir: özel segmentlerin doğrudan aktarımı, kök-ek çözümleme yolu ve başarısız çözümlemede BPE yedek mekanizmasına geçiş.

5.3.4 Kodlama-Çözme Çevrimi Doğruluğu

MFT'nin uygulama bağlamında güvenilir olması yalnızca model eğitimi için değil, değerlendirme ve hata analizi için de kritiktir; çünkü tokenizasyon hataları modelin ürettiği yanıtın okunabilirliğine ve izlenebilirliğine doğrudan yansır. Bu nedenle hibrit yaklaşımda encode edilen token dizisinin tekrar yüzey metne çevrilebilmesi, yani kodlama-çözme çevrimi doğruluğu, büyük ölçekli bir örneklem üzerinde doğrulanmıştır. Cosmos-Corpus veri kümesinden [5] seçilen **66.547** kelimede yapılan değerlendirmede kelime-bazlı `decode(encode(w))` **tam eşleşme doğruluğu %99,48** olarak ölçülmüştür [6]. Aynı veride tek bir seferde tüm metnin encode-decode edilip çıkan kelime dizisinin hizalanmasıyla ölçülen daha gevşek tanılayıcı metrik ise **%99,84** (66.442/66.547) bulunmuştur. Bu iki sayının farklılaşması, Türkçe morfofonolojisinde aynı soyut morfem dizisinin bağlama göre birden fazla yüzey biçime karşılık gelebilmesi (örn. ünlü daralması/ünsüz yumuşaması), ayrıca kaynak metindeki yazım varyantları ve diakritik gürültüsü gibi etkenlerden kaynaklanır. Kalan %0,52'lik tam eşleşme farklarının büyük bölümü

üç tip vakada yoğunlaşmaktadır: **ödünçleme ve özel yazım kalıplarında ünsüz yumuşaması belirsizlikleri** (örn. *ittifakından* ↔ *ittifağından*), **dil içi varyantlara bağlı ünlü düşmesi/daralması türü çiftler** (örn. *çağırıyordu* ↔ *çağrıyordu*) ve **kaynak metindeki yazım gürültüsü** (örn. diakritiklerin atlanması) [6].

MFT'nin başarımı, bu bölümde tanımlanan %TR ve %Pure metrikleriyle izlenir ve deneysel doğrulama Bölüm 6'te raporlanır. Tokenizasyon standartları perspektifi, farklı değerlendirme protokollerinde (TR-MMLU üzerinde sözlük uyumu ve morfem bütünlüğü, farklı cümle kümeleri üzerinde token/kelime oranı) ölçümler değişse bile temel hedefin aynı kaldığını gösterir: Türkçeye daha iyi hizalanmış tokenlar; daha kısa diziler, daha yüksek dil-özgü kapsama ve daha tutarlı temsiller üretmelidir [3].

5.4 DQ-ALR: Kavramsal Çerçeve

Gerçek dünya veri setleri çoğunlukla heterojen kaynaklardan gelir ve örneklerin güvenilirliği, içerdiği uzmanlık düzeyi ve metinsel tutarlılığı açısından değişkendir. Bu değişkenlik yalnızca Türkçe veya NLP'ye özgü değildir; aksine pek çok yapay zeka ve makine öğrenmesi senaryosunda modelin "gürültüye aşırı uyum" göstermesine veya düşük güvenilirlikli örneklerden yanlış genellemelere gitmesine yol açabilir. Bu nedenle tez, veri kalitesinin yalnızca veri temizleme aşamasında ele alınmaması gerektiğini savunur ve bu sinyalin potansiyel olarak öğrenme dinamiklerine yansıtılabileceğini tartışır; sağlık gibi hata toleransı düşük alanlar ise bu perspektifin pratik önemini daha görünür kılar. Bu bölüm, DQ-ALR'yi henüz deneysel olarak doğrulanmış bir yöntem olarak değil, gelecek araştırmalar için motivasyon ve kavramsal temel sunan bir çerçeve olarak tanımlar.

DQ-ALR, kavramsal olarak örnek bazında öğrenme oranı tanımlayarak, güvenilir örneklerin parametre güncellemesinde daha güçlü etki oluşturmasını; düşük güvenilir örneklerde ise daha temkinli güncelleme yapılmasını hedefler. DQ-ALR'nin en sade formu, sağlık veri setindeki sağlayıcı uzmanlığını (doktor unvanı) doğrudan bir ağırlık olarak kullanır [239]. Bu durumda her örnek için öğrenme oranı:

$$\eta_i = \eta_{\text{base}} \cdot w_{\text{title}}(i) \tag{5.3}$$

şeklinde tanımlanır; burada η_{base} global öğrenme oranı, $w_{\text{title}}(i)$ ise örneğin sağlayıcısına ilişkin unvan-ağırlığıdır. Formül 5.3, DQ-ALR'nin en sade unvan-ağırlıklı sürümünü gösterir [239].

Tezde kullanılan çerçeve bu sade formu genişleterek, örnek güvenilirliğini tek

bir kaynaktan değil, birden fazla sinyalden türetir. Güvenilirlik puanı R_i, kaynak güvenilirliği (T_i), metin tutarlılığı (C_i) ve etiket/çıktı güvenilirliği (L_i) bileşenlerinin ağırlıklı toplamı olarak modellenir:

$$R_i = \alpha_1 T_i + \alpha_2 C_i + \alpha_3 L_i \tag{5.4}$$

Formül 5.4, DQ-ALR'yi yalnızca "kim söyledi?" sorusuna değil, aynı zamanda "ne kadar tutarlı?" ve "etiket ne kadar güvenilir?" sorularına da duyarlı hale getirmeyi hedefler. Bu bileşenlerin katkısının gelecek çalışmalarda ayrı ayrı deneysel olarak sınanması öngörülmektedir.

Örnek-bazlı öğrenme oranının güvenilirlik puanı ile ilişkilendirilmesi için doğrusal bir ölçekleme kullanılabilir:

$$\eta_i = \eta_{\text{base}} \cdot \big(1 + \lambda R_i\big) \tag{5.5}$$

Burada λ, güvenilirlik etkisinin şiddetini belirleyen bir katsayıdır. Formül 5.5, R_i yüksek olduğunda öğrenme adımını büyütür; R_i düşük olduğunda ise güncellemenin daha konservatif olmasını sağlar. Pratikte DQ-ALR, AdamW gibi adaptif optimizasyon algoritmalarıyla uyumlu olacak biçimde, örnekten türetilen η_i değerini mikro-batch düzeyinde uygulayacak şekilde entegre edilir. Böylece veri kalitesi, yalnızca "hangi örnekler eğitime girsin?" sorusunu değil, "hangi örnekler öğrenmede ne kadar ağır bassın?" sorusunu da cevaplar.

Örnek-bazlı öğrenme oranı, optimizasyon güncellemesine doğrudan şu biçimde yansıtılabilir. AdamW güncellemesinin örnek i için ölçeklenmiş hali, momentum ve varyans terimleri ile birlikte,

$$\theta_{t+1} = \theta_t - \eta_i \cdot \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} \tag{5.6}$$

şeklinde yazılabilir. Formül 5.6, örneklerin güvenilirliğini "güncelleme büyüklüğü" üzerinden kontrol eder. Alternatif bir bakış açısı ise, η_i yerine kaybı ağırlıklandırmak ve global öğrenme oranını sabit tutmaktır; çünkü lineer kayıp ağırlıklandırması gradyanı ölçeklediği için pratikte benzer bir etki üretir. Tezde amaç, eğitim dinamiğini daha öngörülebilir kılmak adına, global zamanlayıcıyı korurken veri kalitesi sinyalini örnek katkısına yansıtmaktır; bu nedenle DQ-ALR tasarımı, örnek ağırlığı ile öğrenme oranı düzeltmesini birlikte düşünecek şekilde ele alınır.

Bu strateji özellikle tıp alanında iki nedenle kritiktir. Birincisi, hekim unvanı ve uzmanlık alanı gibi sinyaller; örneklerin içerdiği bilginin güvenilirliği hakkında

doğrudan ipucu sunar. İkincisi, tıbbi metinlerde modelin uydurma (hallucination) üretmesi yalnızca başarım düşüklüğü değil, güvenlik riski de doğurabilir; bu nedenle öğrenme sürecinin güvenilir örneklerle daha güçlü biçimde şekillenmesi, etik ve uygulama açısından önleyici bir mekanizma olarak da değerlendirilebilir [239].

Şekil 4.1, 4.3 ve 4.2 (Bölüm 4), veri setindeki unvan, uzunluk ve uzmanlık dağılımlarını sunarak sağlık verisi heterojenliğini somutlaştırır; bu dağılımlar, DQ-ALR çerçevesinde tartışılan güvenilirlik sinyallerinin potansiyel kaynağını oluşturur [241].

DQ-ALR'nin pratikteki uygulanışı, yalnızca "öğrenme oranını çarpmak" olarak düşünülmemelidir; zira modern optimizasyonlarda güncelleme büyüklüğünü belirleyen birden fazla etken vardır. Birincisi, örnek-bazlı η_i değerinin mini-batch içinde nasıl birleştirileceğidir. Tezdeki yaklaşım, örneklerin güvenilirliğini aynı batch içinde ağırlıklandırarak gradyan katkısını düzenlemeyi hedefler; böylece yüksek güvenilirlikli örneklerin "öğrenme baskınlığı" doğrudan gradyan toplamasına yansır. İkincisi, güvenilirlik sinyalinin kararlılığıdır; R_i ölçeği veya dağılımı değiştikçe öğrenme dinamiği aşırı dalgalanabilir. Bu nedenle uygulamada R_i değerlerinin belirli aralıklarda kırpılması (clipping) veya ölçeklenmesi, DQ-ALR'nin eğitim boyunca istikrarlı çalışması açısından kritik bir mühendislik kararıdır. Üçüncüsü, temel zamanlayıcının (ısınma, kosinüs gibi) korunmasıdır; DQ-ALR, global zamanlayıcıyı ortadan kaldırmak yerine, global η_{base} üzerine örnek-bazlı bir düzeltme çarpanı olarak eklenerek öğrenmenin erken safhalarında aşırı büyük güncellemeleri engeller.

5.5 Alan-Özel Model Uyarlama: LoRA ve Slerp

Alan-özel veri setleri çoğu zaman, modelin genel dil yeteneklerini bütünüyle yeniden öğrenmeyi gerektirecek ölçekte değildir; ayrıca milyarlarca parametreli LLM'lerde tam ince ayar yüksek maliyetlidir. Bu nedenle tez, parametre-verimli uyarlamayı temel yaklaşım olarak benimser. LoRA, büyük ağırlık matrislerine düşük-rank güncelleme bileşeni eklenerek, çok küçük bir parametre alt kümesiyle etkili bir uyarlama sağlar [30]. Bir ağırlık matrisi $W_0 \in \mathbb{R}^{d\times k}$ için:

$$W = W_0 + \Delta W = W_0 + BA \tag{5.7}$$

şeklinde tanımlanan güncellemede, $A \in \mathbb{R}^{r\times k}$ ve $B \in \mathbb{R}^{d\times r}$ matrisleri eğitilir; $r \ll \min(d, k)$ seçildiği için eğitilen parametre sayısı belirgin biçimde düşer. Formül 5.7, sağlık gibi alanlara uyarlamada hem bellek hem süre açısından uygulanabilir bir yol

sunar.

LoRA'nın parametre tasarrufu, güncellenen matrisin boyutuna bağlı olarak nicel biçimde görülebilir. Tam güncellemede $d \times k$ parametre eğitilirken, LoRA'da ek parametre sayısı yaklaşık $r(d + k)$ mertebesindedir; r küçük seçildiğinde bu, özellikle attention projeksiyonları gibi büyük matrislerde önemli bir tasarrufa karşılık gelir. Bu tezde LoRA, alan-özel uyarlamanın deneysel döngüsünü hızlandırmak için bir "maliyet azaltıcı" değil, aynı zamanda "kontrollü uzmanlaşma" aracı olarak kullanılır; çünkü güncellemenin düşük-rank bir alt uzayla sınırlandırılması, genel yeteneklerdeki bozulmayı doğal olarak frenleyen bir düzenleme etkisi de yaratır.

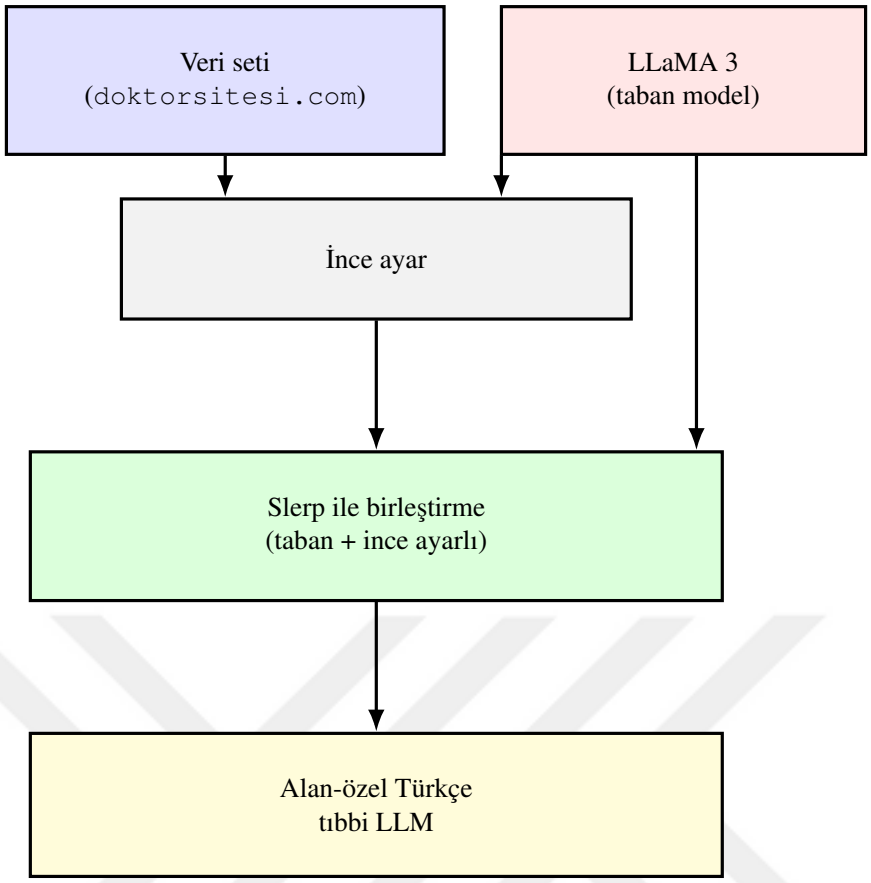
LoRA uygulamasında kritik tasarım kararı, hangi katmanlara adapte edileceğidir. Tezde hedef, alan terminolojisinin ve alan-özel akıl yürütmenin ağırlıklı olarak attention ve MLP projeksiyonlarında temsil edildiği varsayımıyla, güncelleme parametrelerinin modelin tamamına yayılmak yerine belirli projeksiyon matrislerinde yoğunlaşmasını sağlamaktır. Bu, hem eğitimin kararlılığını artırır hem de sınırlı bütçeyle daha hızlı deneme–yanılma yapmayı mümkün kılar. Aynı taban modelin farklı alt alanlara uyarlanması gerektiğinde, LoRA adaptörlerinin alan başına ayrı tutulması, daha sonra birleştirme veya çoklu-adaptör stratejileriyle (ör. Slerp) denge kurulmasını kolaylaştırır.

LoRA ile yapılan alan-özel ince ayarlarda kritik bir risk, modelin genel dil yeteneklerinde gerileme (felaket unutma) göstermesidir. Tezde bu risk, Slerp tabanlı model birleştirme ile azaltılır. Slerp, iki model ağırlık vektörü arasında, uzayın geometrisini daha iyi koruyan küresel doğrusal interpolasyon uygular:

$$\text{Slerp}(A_1, A_2; \alpha) = \frac{\sin((1-\alpha)\theta)}{\sin(\theta)} A_1 + \frac{\sin(\alpha\theta)}{\sin(\theta)} A_2 \tag{5.8}$$

Burada $\alpha \in [0, 1]$ birleştirme katsayısı, θ ise iki vektör arasındaki açıdır. Formül 5.8 bağlamında pratikte A_1 taban modelin, A_2 ise alan-özel uyarlanmış modelin ağırlıklarını temsil eder; seçilen α ile "uzmanlık bilgisini artırırken genel yeteneği koruma" dengesi kurulur [6].

Slerp'in pratik değeri, α seçiminin açık bir "hedef fonksiyon" ile belirlenebilmesinden gelir. Tezde α seçimi, TR-MMLU gibi genel yetenek göstergeleri ile alan-özel doğrulama ölçütlerinin birlikte dikkate alındığı bir denge problemidir; amaç, uzmanlık kazanımı sağlanırken genel performansın aşırı düşmesini engellemektir. Bu perspektif, model birleştirmeyi "tek seferlik bir işlem" değil, doğrulama verisine dayalı bir hiperparametre seçimi problemi olarak ele alır.



Şekil 5.3 Alan-özel uyarlama akışı.

Şekil 5.3, sağlık alanında izlenen iki aşamalı uyarlama hattını özetler: önce alan-özel veri ile uzmanlık performansı yükseltilir, ardından Slerp tabanlı birleştirme ile genel dil yeteneklerinin korunması hedeflenir [241]. Tezin deney bölümünde, bu hedefin TR-MMLU ve alan-özel değerlendirmelerle birlikte raporlanması, "uzmanlaşma–genellik" dengesinin ölçülebilir hale gelmesini sağlar (Bölüm 6).

5.6 Uzun Bağlamlı Türkçe Sentence Embedding Modeli (embeddingmagibu-200m)

Bu tezde, Türkçe odaklı ancak çok dilli sözlük desteğine sahip uzun bağlamlı bir cümle gömme modeli olan **embeddingmagibu-200m** geliştirilmiş ve değerlendirilmiştir [245]. Model, 8K token bağlam penceresi ve 128K çok dilli tokenizer sözlüğüyle 768 boyutlu ℓ_2 normalize edilmiş vektörler üretir ve SentenceTransformers formatında paketlenmiştir. Bu yaklaşım, Türkçe cümle benzerliği ve çok görevli embedding değerlendirmelerinde temsil kalitesini pratik maliyetle artırmayı hedefler.

embeddingmagibu-200m, tam ön-eğitim yerine, üç aşamalı bir uyarlama hattı ile öğretmen modelin semantik uzayını Türkçeye taşır [245]. İlk aşamada çok dilli kapsama hedefiyle 128K sözlük boyutuna sahip bir SentencePiece-BPE tokenizer eğitilir. Bu seçim, Türkçe morfolojik kapsamanın korunması, çok dillilik ve embedding tablosu parametre sayısının kontrolünü birlikte hedefler [246], [244]. Tokenizer eğitimi için Cosmos Turkish Corpus v1.0 kullanılır; bu korpus, Türkçe ön-eğitim senaryolarında geniş kapsama sağlayan büyük ölçekli bir metin kaynağıdır.

İkinci aşama, ağırlık koruyan klonlama ve embedding yeniden eşlemedir. Öğretmen model olarak EmbeddingGemma-300M kullanılır; omurga ağırlıkları korunur, ancak tokenizer değişimi nedeniyle token embedding tablosu yeniden inşa edilir [255], [245]. Her hedef token j için, öğretmen tokenizer altında aynı yüzey formun kodlanmasıyla bir eşleme elde edilir: $\pi(j) \mapsto (i_1, \ldots, i_k)$. Hedef token embedding'i, öğretmen embedding'lerinin ortalaması ile başlatılır:

$$E'_j = \frac{1}{k} \sum_{m=1}^{k} E_{i_m} \tag{5.9}$$

Bu başlangıç, Formül 5.9 ile tanımlandığı üzere rastgele başlatmaya kıyasla daha kararlı bir uyarlama sağlar ve damıtma sürecinin daha hızlı yakınsamasına yardımcı olur. Klonlama ve eşleme işlemleri `transformer-cloner` aracıyla otomatikleştirilmiştir [256].

Üçüncü aşama çevrimdışı embedding damıtmasıdır. Öğretmen modelin her eğitim adımında çalıştırılması maliyetli olduğundan, öğretmen embedding'leri önceden hesaplanır ve veri seti olarak yayımlanır [242]. Damıtma eğitiminde yalnızca öğrenci model çalıştırılır; hedef olarak öğretmenin normalize edilmiş embedding'leri kullanılır [247]. Kayıp fonksiyonu, normalize edilmiş öğrenci ve öğretmen vektörleri arasındaki kosinüs benzerliği üzerinden tanımlanır:

$$\mathcal{L}_{\cos} = \frac{1}{N} \sum_{i=1}^{N} \left(1 - \hat{s}_i^\top \hat{t}_i\right) \tag{5.10}$$

Burada $\hat{s}_i$ ve $\hat{t}_i$ sırasıyla öğrenci ve öğretmenin normalize edilmiş embedding'leridir. Formül 5.10 ile verilen bu objektif, öğrenciyi öğretmenin semantik uzayına hizalamayı amaçlar ve çevrimdışı önhesaplama sayesinde eğitim maliyetini belirgin biçimde düşürür [245].

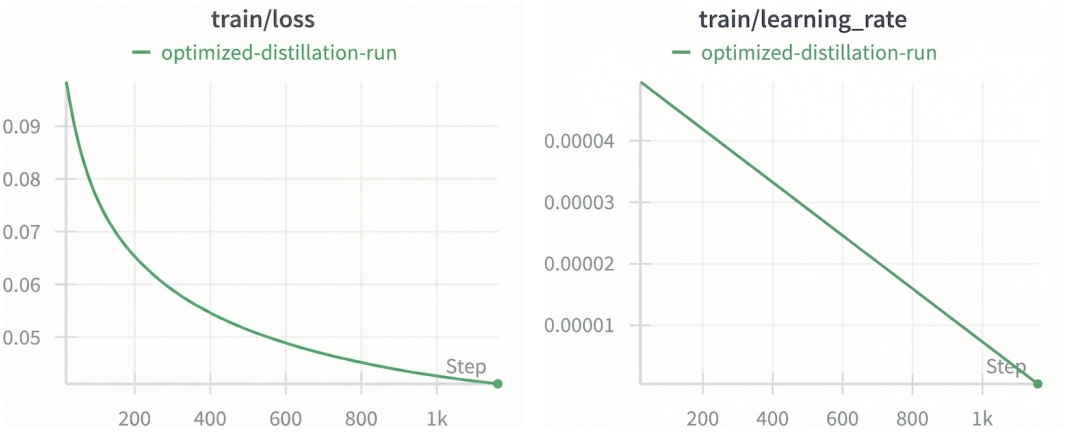
Çevrimdışı damıtma hattının somut veri ayağı için, çok dilli ve dengeli bir Wikipedia korpusu hazırlanmış ve öğretmen embedding'leri bu korpus üzerinde

üretilmiştir. Korpus, Wikimedia Wikipedia akışından 40 dil için örnekler toplanarak oluşturulmuştur [248]. Dil seçimi, Türkçe ve İngilizce gibi yüksek kaynaklı dilleri yüksek kota ile temsil ederken, geri kalan diller için standart kota uygulayacak şekilde kurgulanmıştır. Türkçe ve İngilizce için 100K/30K/10K (train/validation/test), diğer 38 dil için 10K/3K/1K örnek hedeflenmiş, toplamda 812K örneklem büyüklüğü planlanmıştır. Akış modunda çekilen örnek sayısı bazı dillerde hedef kotanın altında kaldığı için dil başına oranlı düzeltme uygulanmış ve nihai veri dağılımı erişilebilir örneklerle sınırlandırılmıştır. Veri çekme süreci 8 iş parçacıklı paralel yürütme ile hızlandırılmış, her dil için aynı metin alanları (lang, title, text, url) toplanmıştır. Yerel diske yazılan veri, ardından HuggingFace Hub'a gönderilerek sürümlenmiştir.

Öğretmen embedding'leri, EmbeddingGemma-300M modeli ile bu korpus üzerinde üretilmiş ve `generate_and_push` akışıyla doğrudan Hub'a aktarılmıştır [249]. Üretimde, metin alanı `text` olarak seçilmiş, çıktı olarak normalize edilmiş nihai vektörler (final) hedeflenmiş ve büyük hacim nedeniyle her 20.000 örnekte bir ara push yapılmıştır. Sonuç veri seti tek bir eğitim split'i altında yaklaşık 580K satır içerir ve otomatik olarak optimize edilmiş Parquet formatına dönüştürülmüştür. Temel alanlar `lang`, `title`, `text`, `url` ve embedding sütunlarıdır. Embedding sütunları 768 boyutlu vektörlerden oluşur (`teacher_embedding_final`, ayrıca pre-dense temsilin saklandığı `teacher_embedding_pre_dense`); bu sayede hem doğrudan damıtma hedefi olarak kullanılabilecek nihai vektörler hem de ara katman temsilleri ayrı ayrı erişilebilir hale getirilmiştir. Bu büyük ölçekli embedding korpusu, öğrenci modelin öğretmen uzayına hizalanmasında ana veri kaynağı olarak kullanılmıştır.

Model kartında raporlanan eğitim düzeni, çevrimdışı damıtma fikrinin pratikte neden avantajlı olduğunu netleştirir: öğretmen embedding'leri bir kez üretildiğinde, öğrenci eğitimi boyunca öğretmen çalıştırılmadığı için eğitim maliyeti belirgin biçimde azalır [245]. Bu tezde kullanılan uyarlama senaryosunda, damıtma verisi yaklaşık 300 bin örnek ölçeğindedir; eğitim tek epoch ile sınırlandırılarak, öğrenmenin ana hedefi "öğretmen uzayına hizalanma" olarak tanımlanmıştır. Örnek bir konfigürasyon, 5×10^{-5} öğrenme oranı, kosinüs zamanlayıcı, kısa bir ısınma oranı ve ağırlık çürümesi gibi standart stabilizasyon bileşenlerini içerir [245]. Bu noktada kritik karar, öğretmen embedding'lerinin hangi temsil düzeyinde hedefleneceğidir; normalize edilmiş nihai vektörleri hedeflemek, benzerlik uzayını doğrudan hizalamaya yardımcı olur.

Şekil 5.4, çevrimdışı damıtma eğitimi boyunca kayıp eğrisi ve öğrenme oranı zamanlayıcısını göstermektedir; bu senaryoda hedef öğretmen uzayına hızlı



Şekil 5.4 Eğitim dinamikleri.

hizalanma amaçlandığı için tek epoch içinde kayıp düşüşü beklenir ve global zamanlayıcının korunması güncellemelerin kararlılığına katkı verir [245].

Bu üç aşamalı yaklaşım, Türkçeye özel bir tokenizer ile bağlam verimliliğini artırırken, öğretmen modelin semantik bilgisini koruyarak daha küçük ve odaklı bir öğrenci model üretmeyi hedefler. STSbTR ve TR-MTEB üzerindeki sonuçlar, yöntemin mutlak liderlikten ziyade pratik ve rekabetçi bir kalite-maliyet dengesi sunduğunu göstermektedir (Bölüm 6) [245], [90].

5.7 Deneysel Altyapı ve Tekrarlanabilirlik

Tez boyunca raporlanan deneyler, tekrarlanabilirliği en üst düzeyde tutacak biçimde tasarlanmıştır. Değerlendirme protokolleri mümkün olduğunca standartlaştırılmış, veri sürümleri ve seed değerleri sabitlenmiş, metrik hesapları aynı betikler üzerinden yürütülmüştür. TR-MMLU değerlendirmelerinde soru formatı ve yanıt ayrıştırma kuralları sabitlenerek model çıktılarının karşılaştırılabilir olması sağlanmıştır [237]. Tokenizasyon değerlendirmelerinde %TR ve %Pure metrikleri aynı sözlük ve aynı hesaplama prosedürü ile üretilmiş, hesaplama maliyeti (token/kelime oranı, işlem süresi) birlikte raporlanmıştır [3]. DQ-ALR deneylerinde, aynı veri kümesi üzerinde sabit öğrenme oranı ile DQ-ALR arasındaki fark; doğruluk, kayıp ve yakınsama davranışı üzerinden raporlanmıştır [239]. Embedding tarafında ise STSbTR ve TR-MTEB değerlendirmeleri, kamuya açık değerlendirme protokollerine uygun biçimde yürütülmüş ve sonuçlar dış denetime açık olacak şekilde paketlenmiştir [245], [250], [91].

5.8 Kontrollü Tokenizasyon Deneyi: Alt Görev Kurulumu

Bu tezde, MFT'nin "içsel" tokenizasyon metriklerinde (TR%, Pure%) gözlenen üstünlüğünün pratikte semantik görev başarımına yansıyıp yansımadığı, **yalnızca tokenizasyon değişkeninin ayrıştırıldığı** kontrollü bir alt görev deney hattı ile test edilmiştir. Bu amaçla dört model, aynı mimari ve aynı eğitim bütçesiyle *sıfırdan* (rastgele başlatım) eğitilmiş; başlangıçta seed=42 ile üretilen tek bir rastgele ağırlık örneklemesi dört modele kopyalanarak başlangıç koşulu tüm karşılaştırıcılar için bire bir aynı tutulmuştur. STS, MTEB-TR ve TurBLiMP üzerinde aynı değerlendirme protokolleriyle karşılaştırma yapılmıştır. Bu kurulumun özeti Tablo 5.2'da verilmektedir.

Tablo 5.2 Kontrollü alt görev deney kurulumu.

Bileşen	**Spesifikasyon**
Öğrenci mimarisi	EmbeddingGemma-300M (SentenceTransformer kodlayıcısı)
Başlatma	Rastgele başlangıç; seed=42 ile tek başlangıç ağırlığı; 4 kopya; sözlük: 32.768
Eğitim hedefi	Kosinüs benzerliği kaybı ile `teacher_embedding_final` hizalaması
Birleşik veri seti	Birleşik kodlanmış veri seti (4 ayrı `input_ids` sütunu)
Bağlam penceresi	2048; herhangi bir tokenizer sınırı aşarsa örnek tüm akışlardan elenir
Hiperparametreler	BF16, batch size=256, 5×10^{-5} LR, 100 adım ısınma + 1 epoch
Donanım	NVIDIA H100 80GB (tek düğüm)
Karşılaştırıcılar	MFT / Tabi / Cosmos / Mursit
Değerlendirme	STSbTR, MTEB-TR (26 görev), TurBLiMP (16 olgu, minimal çift)

Tablo 5.2 içinde geçen bağlantılar kaynakça atıflarıyla verilmiştir: EmbeddingGemma-300M model kartı [257] ve birleşik kodlanmış veri seti [243].

5.8.1 Veri Hazırlama

Eğitim verisi, öğretmen embedding'lerinin önceden hesaplandığı bir Türkçe metin korpusu alt kümesinden başlayarak hazırlanmıştır [242]. Bu korpus üzerinde birleşik bir veri seti üretilmiş [243] ve her örnek için dört farklı tokenizasyon akışına ait `input_ids` sütunları saklanmıştır: `mft_input_ids`, `tabi_input_ids`, `cosmos_input_ids` ve `newmindaiMursit_input_ids`. Adil karşılaştırmayı sağlamak için bağlam penceresi sabitlenmiş (`max_seq_length`=2048) ve *herhangi bir tokenizer* bu

sınırı aşan örnekler tüm akışlardan birlikte elenmiştir; böylece kırpma kaynaklı içerik farkları önlenmiştir.

5.8.2 Modeller ve Karşılaştırıcılar

Dört kontrol modeli, aynı encoder mimarisini paylaşacak şekilde kurulmuştur (EmbeddingGemma-300M omurgası, SentenceTransformer arayüzü). Başlangıçta embedding tablosu aynı sözlük boyutuna yeniden boyutlandırılmış (32.768) ve seed=42 ile tek bir rastgele başlatma yapılmıştır; ardından bu başlangıç ağırlıkları dört modele kopyalanarak değişken uzayı yalnızca tokenizasyona indirgenmiştir. Karşılaştırıcılar: **MFT** (hibrit kök–ek sözlüğü + BPE bileşeni), **Tabi** (32K BPE tabanlı altbirim; TabiBERT tokenizersi), **Cosmos** (32K BPE tabanlı altbirim; CosmosGPT2 tokenizersi) ve **Mursit** (32K BPE tabanlı altbirim; newmindaiMursit tokenizersi). Bu noktada tek bir altbirim tabanlı tokenizera bağlı kalınmamış, farklı araştırma gruplarınca bağımsız biçimde eğitilmiş üç Türkçe BPE tabanlı tokenizer birlikte kullanılmıştır; böylece MFT'nin gözlenen kazanımının belirli bir BPE sözlüğüne özgü bir çıktı paketi etkisi olup olmadığı doğrudan sınanmış ve karşılaştırmanın daha sağlam biçimde yorumlanması hedeflenmiştir.

Eğitim, embedding damıtma hattı ile yürütülmüştür: her öğrenci model, veri setindeki `teacher_embedding_final` hedef vektörlerine kosinüs benzerliği üzerinden hizalanmıştır. Tüm modeller için ortak eğitim ayarları kullanılmıştır: BF16, batch size 256, öğrenme oranı 5×10^{-5}, kısa bir ısınma (100 adım) ardından tam epoch eğitimi ve gradient checkpointing. MFT artık HuggingFace standart tokenizer arayüzüyle uyumlu olsa da, dört tokenizer akışını bire bir ve deterministik biçimde karşılaştırabilmek için `input_ids` çevrimdışı üretilmiş ve eğitim döngüsünde çalışma anında tokenizasyon adımı devreye alınmamıştır.

Önceden eğitilmiş checkpoint'ler üzerinde tokenizer karşılaştırması yapmak, veri karışımı, eğitim bütçesi ve optimizasyon geçmişi gibi etkenleri kontrol etmeyi zorlaştırabilir. Bu nedenle rastgele başlatım kontrol hatları, **başlangıç ağırlıkları, mimari, veri ve hedef (öğretmen) uzayını** sabitleyerek, gözlenen performans farklarının büyük ölçüde tokenizasyonun getirdiği indüktif önyargı ve temsil özelliklerinden kaynaklandığını gösterebilen daha temiz bir deney zemini sağlar.

Tekrarlanabilirlik, sadece aynı kodu çalıştırmakla değil, aynı "kararları" aynı sırada alabilmekle ilgilidir. Bu nedenle tezde; veri sürümü (örn. TR-MMLU'nun karşılaştırıcı alt kümesi ve soru havuzu sürümleri), değerlendirme komutları, seed sabitleme ve çıktı ayrıştırma kuralları birlikte raporlanır [237]. Eğitim tarafında seedlerin sabitlenmesi, veri karıştırma sırasının kontrol edilmesi ve değerlendirme

sırasında aynı istem şablonunun kullanılması gibi kararlar, özellikle küçük farkların sonuçları etkileyebildiği düşük kaynak senaryolarında kritik hale gelir. Embedding tarafında çevrimdışı damıtma yaklaşımı da bu hedefe hizmet eder; öğretmen embedding üretimi bir kez yapılıp arşivlendiği için, öğrenci eğitimi aynı hedef vektörler üzerinde bağımsız olarak tekrar edilebilir [247], [242].

Bu yaklaşımın kritik çıktısı, yöntemlerin yalnızca "bu tezde çalıştı" biçiminde değil, farklı araştırmacılar tarafından da doğrulanabilir biçimde tanımlanmasıdır. Bu nedenle, model ve veri çıktı paketleri, değerlendirme çıktıları ve eğitim araçlarıyla birlikte yayımlanmış; embedding tarafında `transformer-cloner` ve `distil-trainer` gibi araçlar ile çevrimdışı damıtma süreci yeniden üretilebilir hale getirilmiştir [256], [247], [242].

5.9 Bölüm Özeti

Bu bölümde, tokenizasyon standartları, DQ-ALR ve alan-özel uyarlama yaklaşımlarıyla birlikte embeddingmagibu-200m'in geliştirme hattı ayrıntılandırılmıştır. Tokenizasyon tarafında %TR ve %Pure metrikleriyle dilbilimsel kapsama ölçülmüş; DQ-ALR ile veri kalitesi sinyalinin optimizasyona nasıl taşındığı gösterilmiş; LoRA ve Slerp ile alan-özel uzmanlaşma ve genel yetenek dengesi tanımlanmıştır. Embedding hattında ise çok dilli tokenizer eğitimi, ağırlık koruyan klonlama ve çevrimdışı damıtma aşamaları sistematik biçimde açıklanmıştır.

Veri tarafında, çok dilli Wikipedia örneklerinden oluşturulan ve öğretmen embedding'leri ile zenginleştirilen yeni korpusun hazırlanışı detaylandırılmıştır. Dil bazlı kota tasarımı, akış veri toplama, paralel örnekleme ve HuggingFace Hub'a sürümleme adımları, damıtma sürecinin veri ayağını tekrarlanabilir kılan ana bileşenler olarak sunulmuştur. Böylece izleyen bölümde raporlanan deneylerin hangi veri çıktı paketlerine dayandığı, hangi temsil düzeylerinin hedeflendiği ve bu hedeflerin nasıl üretildiği netleşmiştir. Ayrıca deneysel protokoller ve tekrarlanabilirlik ilkeleri tanımlanarak, bulguların dayandığı yöntemsel temel güçlendirilmiştir.

Bölümün sonunda ortaya çıkan genel çerçeve, veri kalitesi, tokenizasyon ve alan uyarlama kararlarının birbirinden bağımsız değil, aynı eğitim hattının birlikte optimize edilen parçaları olduğunu vurgular. Bu bütünlük, deney bölümündeki performans farklarının yöntemsel kökenlerinin daha açık biçimde izlenebilmesini sağlar.

6
DENEYSEL ÇALIŞMALAR

Bu bölümde, tez kapsamında geliştirilen yöntemlerin deneysel doğrulaması sunulmaktadır. Deneylerin ortak amacı; (i) TR-MMLU üzerinde Türkçe değerlendirme standardına göre model başarımını karşılaştırmak, (ii) tokenizasyon standartlarının Türkçe Token Oranı (Turkish Ratio, TR%) ve Saf Token Oranı (Purity Ratio, Pure%) metrikleriyle Türkçe görevlerdeki etkisini nicel biçimde göstermek, (iii) veri kalitesi sinyalini optimizasyona taşıyan DQ-ALR yaklaşımının yakınsama ve genelleme üzerindeki etkisini analiz etmek, (iv) alan-özel Türkçe tıp modeli uyarlamasında LoRA ve Slerp'in rolünü somut ölçütlerle değerlendirmek ve (v) embeddingmagibu-200m modelinin performansını STSbTR ve TR-MTEB üzerinden raporlamaktır [237], [3], [241], [245], [252].

Tüm deneyler, Bölüm 5'de tanımlanan metodolojik çerçeveyle tutarlı biçimde yürütülmüş; veri sürümü ve değerlendirme betikleri sabitlenmiş, mümkün olan yerlerde sabit seed kullanılmış ve sonuçların tekrar üretilebilirliği önceliklendirilmiştir.

Bu bölüm aynı zamanda, yöntem bileşenlerinin birbirinden bağımsız değil, aynı protokol içinde test edilebilir olduğuna dair ortak bir değerlendirme dili kurar. Böylece farklı yöntem aileleri arasında adil karşılaştırma yapılabilmekte ve performans farklarının hangi tasarım kararlarından kaynaklandığı daha açık biçimde izlenebilmektedir.

6.1 Hipotezlerle İlişki

Bu bölüm, Giriş'te tanımlanan **H1-H5** hipotezlerinin deneysel sınanmasını raporlar (bkz. Bölüm 1.3).

6.2 Deneysel Kurulum ve Ortak Protokoller

Bu tezde deneysel doğrulama, farklı yöntem ailelerinin aynı ölçüm zemininde karşılaştırılabilir olmasını hedefler. Bu amaçla ölçüm hattı iki ana prensip etrafında kurgulanmıştır. Birinci prensip, ölçümlerin mümkün olduğunca deterministik ve tekrar üretilebilir yürütülmesidir. Bu kapsamda aynı veri sürümleri kullanılmış, metin normalizasyonu ve değerlendirme betikleri sabitlenmiş ve raporlanan tüm sayıların aynı protokol altında üretildiği kontrol edilmiştir. İkinci prensip ise, yöntemlerin etkisini "kontrollü değişken" yaklaşımıyla izleyebilmektir; örneğin tokenizasyon deneylerinde model ailesi sabit tutulup tokenizer tercihinin etkisi ayrıştırılmaya çalışılmış, alan-özel uyarlamada ise aynı taban model üzerinde farklı adaptasyon stratejilerinin çıktıları karşılaştırılmıştır [3], [241].

Bu bölümde kullanılan ölçütler, görev tipine göre üç ana grupta ele alınmaktadır. Çoktan seçmeli bilgi/muhakeme testleri için doğruluk (accuracy) kullanılırken, embedding tabanlı değerlendirmelerde korelasyon ve çok görevli karşılaştırıcı ortalamaları tercih edilmiştir. Alan-özel tıbbi değerlendirmelerde ise tek bir otomatik metrikle sınırlı kalmamak adına, metin benzerliği, LLM tabanlı otomatik puanlama ve uzman değerlendirmesi birlikte raporlanmıştır. Bu çoğul ölçüm yaklaşımı, yalnızca "benzer görünen" yanıtların değil, klinik bağlamda anlaşılabilir, tutarlı ve doğru yönlendirici yanıtların üretilebildiğini daha güvenilir biçimde sorgulamayı amaçlar [258], [241].

6.3 TR-MMLU Üzerinde Model Karşılaştırmaları

TR-MMLU, Türkçe büyük dil modellerinin çok alanlı bilgi, muhakeme ve kavramsal anlama performansını ölçmek üzere tasarlanmış bir karşılaştırıcıdır [237], [16]. Bu tez kapsamında TR-MMLU, hem modeller arası karşılaştırma hem de tokenizasyon ve alan uyarlaması gibi yöntemlerin etkisini tek bir standart üzerinde izleyebilmek için temel ölçüm zemini olarak kullanılmıştır. Değerlendirmeler, çoktan seçmeli formatın sabitlenmesi, çıktıdan seçenek harfinin düzenli ifadeyle çıkarılması ve kategori bazlı makro-ortalama doğruluğun raporlanması gibi protokol kararlarıyla standardize edilmiştir [237].

TR-MMLU'nun açık liderlik tablosu 05 Şubat 2026 itibarıyla **85 model değerlendirmesi** içermektedir; bu kapsam, TR-MMLU'yu Türkiye'de kamuya açık Türkçe MMLU karşılaştırıcıları içinde en fazla model değerlendirmesi barındıran ölçüm zemini olarak konumlandırmaktadır [4]. Aynı tarihli özet istatistiklerde liderlik tablosu üst sınırı %90,89 (gemini-3-flash-preview) ve model ortalaması %56,44 olarak gözlenmiştir.

Değerlendirme ortamı, modellerin adil karşılaştırılabilirliği için sabitlenmiştir. Hesaplamalı ayrıntılar Bölüm 5'de verildiği üzere aynı donanım ve yazılım altyapısı kullanılmış, istem şablonları Türkçe için optimize edilerek yanıt formatının "yalnızca seçenek harfi" olacak şekilde kısıtlanması hedeflenmiştir. Bu pratik karar, özellikle farklı modellerin açıklama üretme eğilimleri nedeniyle oluşabilecek ölçüm hatalarını azaltmak için kritik görülmüştür [237].

Tablo 6.1, tez boyunca izlenen modellerden seçili bir alt kümenin TR-MMLU doğruluk karşılaştırmasını özetlemektedir. Bu özet, iki temel gözlemi destekler. Birincisi, parametre sayısı tek başına Türkçe başarımı belirlememektedir; dil-özelleştirilmiş ince ayar ve tokenizasyon tercihi, benzer ölçekli modeller arasında anlamlı farklar doğurabilmektedir. İkincisi, Türkçe görevlerde daha iyi hizalanmış tokenizasyonun, TR-MMLU başarımıyla birlikte hareket ettiği görülmektedir; bu ilişki, Bölüm 5.2'da tanımlanan metriklerle ölçülebilir hale gelmektedir.

Tablo 6.1 TR-MMLU üzerinde seçili model doğrulukları (%).

Model	**Parametre**	**Doğruluk**
Gemma-3-12B-TR (ince ayarlı)	12B	71,08
EuroLLM-9B-Instruct	9B	51,29
Mistral-7B-Instruct	7B	32,73
LLaMA-3-8B-Instruct	8B	52,23
Gemma-2-9B (multilingual)	9B	69,26
Phi-4	14B	65,52
GPT-4o (referans)	-	**84,84**

Bu tablo, tezde kullanılan yaklaşımın neden "standartlaştırılmış bir Türkçe karşılaştırıcı" etrafında şekillendiğini de açıklar. Aynı görev seti üzerinde ölçüm yapılmadığında, farklı veri karışımları ve farklı istem protokolleri nedeniyle iyileşmelerin hangi kaynaktan geldiğini ayırt etmek zorlaşır. TR-MMLU'nun bu tezde temel referans olarak kullanılması, hem tokenizasyon gibi altyapı kararlarının hem de alan-özel uyarlama gibi üst seviye stratejilerin tek bir tutarlı hedef üzerinden izlenebilmesini sağlamaktadır [237].

TR-MMLU'nun yerel ve kültürel olarak özgün içeriği, Türkçe modeller arasındaki performans farklarını görünür kılmıştır. Bu tez kapsamında çeviri tabanlı setlerle birebir eş-protokol karşılaştırma yapılmamış olsa da, gözlenen ayrışma H5 hipotezinin "ayırt edicilik" beklentisiyle uyumludur.

Tokenizasyon etkisini daha net izleyebilmek için, aynı temel model ailesinde yalnızca tokenizer değişiminin yapıldığı kontrollü karşılaştırmalarda %TR ve %Pure artışı ile TR-MMLU doğruluğu arasında pozitif bir ilişki gözlenmiştir.

Bu bulgu, tokenizasyon standartları çalışmasında raporlanan "%TR ile başarı arasındaki güçlü korelasyon" sonucu ile uyumludur ve tezin H1 ve H2 hipotezlerinin deneysel dayanaklarından birini oluşturur [3]. Deneysel açıdan kritik olan nokta, bu ilişkinin yalnızca doğruluk düzeyinde değil, aynı zamanda token ayak izi ve işlem maliyeti gibi pratik kısıtlarla birlikte düşünülmesidir; daha "Türkçe uyumlu" tokenizasyon, aynı bağlam penceresinde daha fazla bilgi taşımayı mümkün kılabildiği için muhakeme performansına dolaylı katkı verebilir.

6.4 Tokenizasyon Yaklaşımlarının Deneysel Değerlendirmesi

Bu bölüm, tokenizasyonun Türkçe üzerindeki etkisini iki ölçekli bir deney hattı üzerinden raporlar. İlk hat, büyük model ailelerinin tokenizasyon davranışını TR-MMLU üzerinde standart metriklerle ölçer ve model başarımı ile ilişkilendirir [3]. İkinci hat ise, morfolojik ve fonolojik bilgiyi tokenizasyona entegre eden hibrit yaklaşımın (MFT/turkish_tokenizer) dilbilimsel metrikler bakımından ne ölçüde iyileşme sağladığını gösterir [6]. Bu deney hattı, morfolojik farkındalıklı tokenizasyonun daha tutarlı temsiller üreteceği (**H1**) ve %TR / %Pure metriklerinin tokenizasyon kalitesi için anlamlı sinyaller sunduğu (**H2**) hipotezlerini doğrudan sınamayı hedefler.

Tokenizasyon standartları çalışmasının bu tezdeki rolü, tokenizasyonu "gizli bir mühendislik ayrıntısı" olmaktan çıkarıp ölçülebilir bir araştırma değişkeni haline getirmesidir. %TR, üretilen tokenların ne kadarının Türkçe birimlerle uyumlu olduğunu; %Pure ise tokenların anlamsal olarak bölünemez birimlere ne ölçüde karşılık geldiğini yakalamayı hedefler. Bu iki metrik, TR-MMLU metinleri üzerinde aynı normalize edilmiş ölçüm hattı ile hesaplandığında, farklı tokenizer tercihleri arasında adil karşılaştırma yapılabilir hale gelir [3]. Tez açısından önemli olan nokta, bu metriklerin yalnızca "daha az token" hedeflemediğidir; hedef, Türkçe gibi eklemeli dillerde temsilin izlenebilirliğini artıran ve anlamlı parçalanmayı azaltan bir kodlamayı teşvik etmektir.

6.4.1 Tokenizasyon Ölçüm Hattı

Bu bölümde raporlanan tokenizasyon standartları sonuçlarını üreten ölçüm hattı, tez kapsamında tokenizasyon ölçüm hattı olarak adlandırılmaktadır. Amaç, tokenizerleri sadece sözlük boyutu veya token sayısı üzerinden değil; aynı zamanda dilbilimsel kapsama (%TR) ve morfem bütünlüğü (%Pure) gibi dil-özgü ölçütlerle, aynı metin normalizasyonu altında ve aynı veri sürümü üzerinde karşılaştırılabilir hale getirmektir. Bu ölçüm hattı, her tokenizer için toplam token sayısı, benzersiz

token sayısı ve işlem süresi gibi hesaplamalı metrikleri; %TR / %Pure gibi dilbilimsel metriklerle birlikte raporlar. Metriklerin döngüsel hale gelmesini önlemek için, %TR / %Pure hesaplarında değerlendirilen tokenizerın kendi sözlüğü referans alınmaz; bunun yerine tokenizerdan bağımsız bir morfolojik doğrulama hattı kullanılır (bkz. Bölüm 5) [3]. Bu yaklaşım, tokenizasyonun Türkçe görevlerdeki etkisini (H1-H2) daha şeffaf ve yeniden üretilebilir biçimde tartışmayı mümkün kılar.

Tablo 6.2 TR-MMLU üzerinde dört farklı tokenizerın karşılaştırmalı sonuçlarını vermektedir. Bu tablo, yalnızca TR-MMLU başarımını değil, aynı zamanda token sayısı, işlem süresi ve sözlük ölçeği gibi hesaplamalı özelliklerle %TR ve %Pure gibi dilbilimsel metrikleri birlikte raporlaması açısından önemlidir. Sonuçlar, %TR değerinin TR-MMLU başarımıyla güçlü korelasyon gösterdiğini, %Pure değerinin ise daha zayıf ama anlamlı bir ilişki sunduğunu göstermektedir [3].

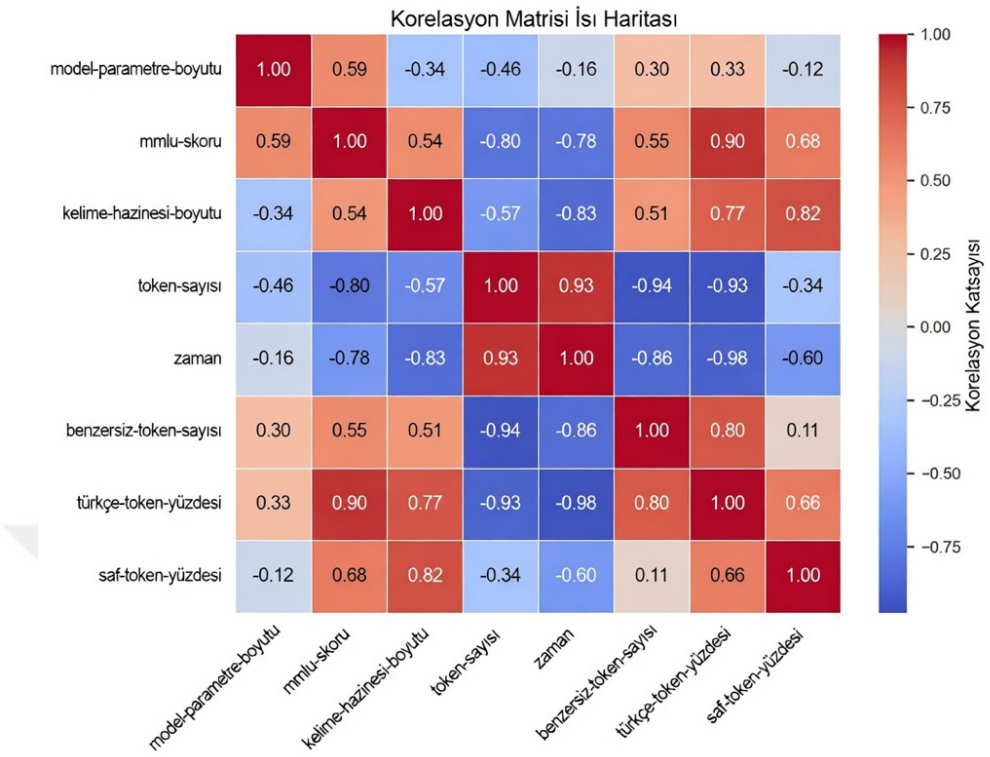
Tablo 6.2 Seçili tokenizer karşılaştırması (TR-MMLU).

Metrik	**gemma-2**	**llama-3.1**	**Qwen2.5**	**aya-expanse**
Model parametreleri (B)	27,2	**70,6**	7,6	32,3
MMLU skoru (%)	**72,10**	70,42	61,68	70,66
Sözlük boyutu	**256.000**	128.256	151.665	255.029
Token sayısı	497.015	488.535	561.866	**434.526**
İşlem süresi (s)	2,95	3,12	3,31	**2,77**
Benzersiz token sayısı	6.383	6.823	5.752	**8.562**
%TR	48,63	45,80	40,33	**50,67**
%Pure	**37,05**	30,91	30,15	32,96

Bu tabloda yer alan karşılaştırıcılar, belirli model aileleri ve sürümleriyle sınırlıdır. Hibrit tokenizasyon çalışmasında ayrıca farklı ölçek/sürümde genel amaçlı tokenizerler için de benzer bir tablo raporlanmaktadır: örneğin Gemma-2-9B [259] için %TR=40,96 ve %Pure=28,49; Llama-3.2-3B [260] için %TR=45,77 ve %Pure=31,45; Aya-Expanse-8B [261] için ise %TR=53,48 değerleri raporlanmıştır [6]. Bu sayılar, Tablo 6.2 ile bire bir aynı değerlendirme koşullarını paylaşmak zorunda değildir; ancak genel amaçlı tokenizasyonların Türkçede dilbilimsel kapsama metriklerinde belirgin sınırlılıklara sahip olduğunu destekleyen ek bir gözlem sunar.

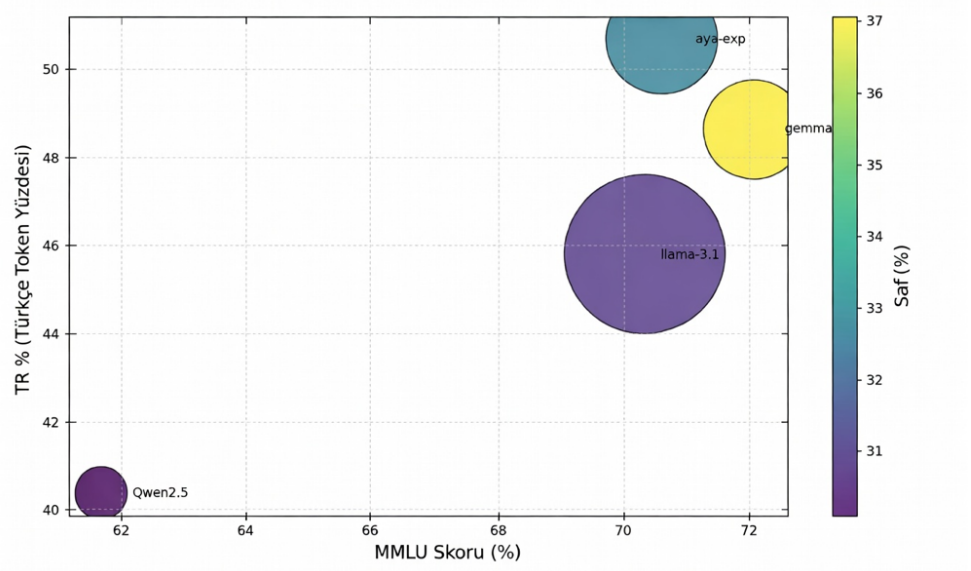
Şekil 6.1 ve Şekil 6.2, tokenizasyon standartları çalışmasındaki korelasyon ve çok boyutlu karşılaştırma görselleştirmelerinin deneysel yorumunu desteklemektedir. Korelasyon matrisi, %TR ile başarı ölçütü arasında güçlü pozitif ilişkiyi; token sayısı ve işlem süresi gibi metriklerin ise dilbilimsel metriklerle negatif ilişkisini görünür kılar. Model karşılaştırma grafiği, %TR ve %Pure değerlerinin birlikte yükseldiği bölgede başarı ölçütlerinin de yükseldiği eğilimi göstermektedir; bu

grafikte işaretçi boyutu parametre ölçeğini, renk ise %Pure değerini temsil eder [3].



Şekil 6.1 Tokenizasyon metrikleri korelasyon matrisi.

Şekil 6.1, metrikler arası doğrusal ilişkinin yönünü ve göreli gücünü özetler; özellikle %TR ile görev başarımı arasındaki pozitif bağın, token sayısı ve işlem süresi gibi maliyet metriklerinden ayrıştığını görselleştirir.



Şekil 6.2 Başarı-%TR ve %Pure ilişkisi.

Hibrit tokenizasyon yaklaşımı (MFT/turkish_tokenizer), genel amaçlı

tokenizasyonlara kıyasla dilbilimsel metriklerde çok daha yüksek değerlere ulaşmayı hedefler. Tablo 6.3 hibrit tokenizerın TR-MMLU metni üzerindeki sonuçlarını özetlemektedir. Bu yaklaşım, dilbilimsel bütünlüğü artırmak adına daha yüksek token sayısı üretebilse bile, tokenların morfem sınırlarına daha iyi hizalanmasıyla temsil tutarlılığını yükseltmektedir [6]. Bu sonuçlar, tezde önerilen yaklaşımın temel savını güçlendirir: Türkçe gibi eklemeli dillerde tokenizasyonun "anlamsal parçalanmayı azaltacak şekilde" tasarlanması, modelin dilsel izlenebilirliğini artırır ve alt görevlerde başarıyı destekleyebilir.

Tablo 6.3 Hibrit `turkish_tokenizer` metrikleri (TR-MMLU).

Metrik	**Değer**
Sözlük boyutu	32.768
Toplam token sayısı	707.727
İşlem süresi (s)	0,6714
Benzersiz token sayısı	11.144
Türkçe token sayısı	10.062
%TR	90,29
Saf token sayısı	9.562
%Pure	85,80

6.4.2 Tokenizasyon Verimliliği (Süre ve Token Ayak İzi)

Hibrit yaklaşım, dilbilimsel bütünlüğü artırmak amacıyla daha ince taneli (morfem hizalı) bir tokenizasyon yaptığı için token sayısını artırabilmektedir. Tablo 6.4 aynı metin kümesi üzerinde farklı tokenizer akışlarının işlem süresi ve token/kelime-token/karakter oranlarını birlikte raporlamaktadır [6].

Tablo 6.4 Tokenizasyon verimliliği karşılaştırması.

Tokenizer	**Süre (ms)**	**Token**	**Tok/Kelime**	**Tok/Kar.**
MFT	1.935	1.899.670	2,91	0,356
Tabi	**1.544**	1.298.725	1,99	0,244
Mursit	1.655	1.187.418	**1,82**	**0,223**
Cosmos	1.620	**1.186.834**	**1,82**	**0,223**

Bu sonuçlar, MFT'nin sıkıştırma odaklı bir tokenizasyon yerine morfem sınırlarını korumayı hedeflemesi nedeniyle token/kelime oranının arttığını (ör. 2,91'e karşı 1,99) göstermektedir. Pratikte bu artış, bağlam penceresinin daha hızlı tüketilmesi pahasına, daha tutarlı morfolojik birim hizalaması ve alt görevlerde ölçülebilir kazanımlar ile birlikte değerlendirilmelidir.

Hibrit yaklaşımın deneysel açıdan bir diğer önemi, tokenizasyonun yalnızca encode aşamasını değil, decode tutarlılığını da bir kalite sinyali olarak ele almasıdır.

Morfolojik çözümleme ve ek kontrolleriyle üretilen token dizisi, yüzey metne geri taşındığında kodlama-çözme çevrimi davranışının daha güvenilir olması beklenir; bu özellik, hem hata analizi hem de uzun metin işleme senaryolarında bağlamın izlenebilirliği açısından faydalıdır [6]. Bu yönüyle MFT, sadece "model girdisi üreten bir tokenizer" değil; yüksek doğrulukla geri birleştirilebilen token dizisi sayesinde pratikte bir Türkçe *metin çözümleyici* olarak da değerlendirilebilir. Şekil 6.2, genel amaçlı tokenizasyonların Türkçe kökleri parçaladığını; hibrit yaklaşımın ise kök–ek ayrımını daha tutarlı biçimde koruduğunu ve %TR / %Pure metriklerinde belirgin üstünlük sağladığını görsel olarak göstermektedir. Kodlama-çözme çevrimi doğrulama, Cosmos-Corpus veri kümesinden [5] seçilen 66.547 kelimelik bir örneklemde kelime-bazlı `decode(encode(w))` tam eşleşme doğruluğu %99,48 olarak ölçülmüştür. Aynı veride tüm metnin tek seferde encode-decode edilip kelime hizalamasıyla ölçülen tanılayıcı metrik ise %99,84 (66.442/66.547) bulunmuştur [6]. Kalan küçük farklar çoğunlukla metinlerdeki ünsüz yumuşaması belirsizlikleri, ünlü düşmesi/daralması türü dil içi varyantlar ve kaynak metin yazım gürültüsü ile açıklanabilmiştir [6].

6.5 Alt Görevlerde Tokenizasyon Etkisi (Kontrollü Deney)

Tokenizasyonun gerçek dünyadaki faydasını gösterebilmek için, yalnızca tokenizasyon değişkeninin ayrıştırıldığı **kontrollü bir alt görev deney hattı** kurulmuştur (bkz. Bölüm 5.8). Bu deneyde dört model, aynı mimari (EmbeddingGemma-300M) [257] ve aynı veri altında *sıfırdan* eğitilmiş; başlangıçta seed=42 ile üretilen tek bir rastgele ağırlık örneklemesi dört modele kopyalanmış ve yalnızca tokenizer akışı (`mft`, `tabi`, `cosmos`, `mursit`) değiştirilmiştir. Amaç, "içsel" tokenizasyon metriklerinde görülen morfolojik bütünlüğün, semantik benzerlik ve bilgi erişim gibi alt görevlerde ölçülebilir bir karşılığı olup olmadığını doğrudan test etmektir. MFT artık standart HuggingFace tokenizer arayüzüyle uyumlu olsa da, aynı eğitim döngüsünün dört tokenizer akışını bire bir karşılaştırabilmesi için `input_ids` çevrimdışı üretilmiş ve eğitim sırasında çalışma anında tokenizasyon adımı devreye alınmamıştır [6].

6.5.1 STSbTR

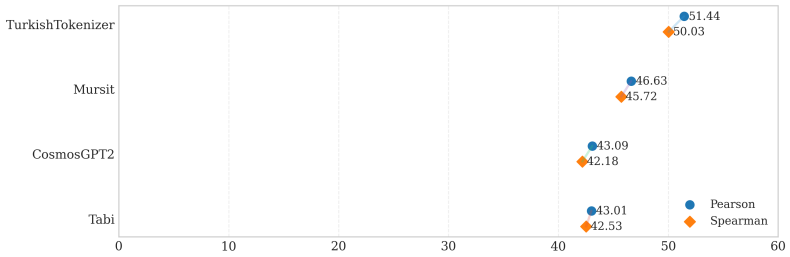
STS değerlendirmesinde, STSbTR veri seti üzerinde cümle embedding'lerinin kosinüs benzerliği ile insan benzerlik puanları arasındaki uyum Pearson ve Spearman korelasyonlarıyla ölçülmüştür [123]. Tablo 6.5 sonuçları, MFT tabanlı kontrol modelinin açık biçimde en yüksek korelasyon değerlerine ulaştığını göstermektedir.

Tablo 6.5 STSbTR kontrollü deney sonuçları.

Model	Pearson	Spearman
`mft-rastgele başlatım`	**51,44**	**50,03**
`newmindaiMursit-rastgele başlatım`	46,63	45,72
`cosmosGPT2-rastgele başlatım`	43,09	42,18
`tabi-rastgele başlatım`	43,01	42,53

Bu sonuç, MFT'nin avantajının özellikle **semantik benzerlik** gibi anlam hizalamasına duyarlı görevlerde belirginleştiğini gösterir ve tezdeki ana anlatıyı güçlendirir: Türkçe gibi eklemeli dillerde morfolojik bütünlüğü koruyan bir tokenizasyon önyargısı, aynı eğitim bütçesi altında daha güçlü semantik temsillere dönüşebilmektedir. Ek olarak, STSbTR eğitim bölümünde de benzer bir sıralama korunmuştur (MFT: %54,76 / %51,90; Mursit: %48,99 / %46,93; Tabi: %45,06 / %43,80; Cosmos: %44,47 / %43,34). Bu tablo, MFT'nin üç farklı Türkçe BPE tabanlı karşılaştırıcıya (Tabi, Cosmos ve Mursit) karşı tutarlı biçimde üstünlük sağladığını ortaya koymaktadır; test bölümünde en güçlü taban karşılaştırıcı olan Mursit'e göre Pearson/Spearman farkı +4,81 / +4,31 puandır.

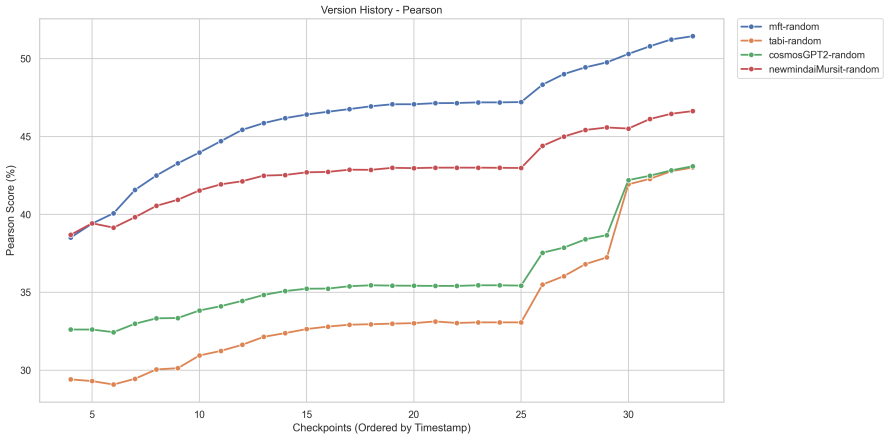
İstatistiksel sağlamlık açısından, Fisher dönüşümü ile hesaplanan güven aralıkları farkın büyüklüğünün belirsizlikle birlikte yorumlanması gerektiğini gösterir: STSbTR test bölümünde MFT Pearson için %95 güven aralığı [47,4 - 55,2] iken Mursit için [42,4 - 50,7] ve Tabi için [38,6 - 47,2] aralığındadır. Aralıkların kısmen örtüşmesi, STSbTR'deki avantajın *tutarlı* olsa da büyüklüğünün örnekleme belirsizliğiyle birlikte değerlendirilmesi gerektiğine işaret eder.



Şekil 6.3 STSbTR kontrollü deney karşılaştırması.

Şekil 6.3, Tablo 6.5'de verilen Pearson/Spearman farklarının görsel karşılığını sunar ve MFT'nin test setinde diğer üç karşılaştırıcıya göre ayrıştığını gösterir.

Şekil 6.4, STSbTR için Pearson skorunun revizyonlar boyunca seyrini görselleştirerek, raporlanan kontrollü sonuçların rastgele bir tek koşuya özgü bir sapma olmadığını nitel olarak destekleyen ek bir tanılayıcı sunar.



Şekil 6.4 STSbTR Pearson revizyon geçmişi.

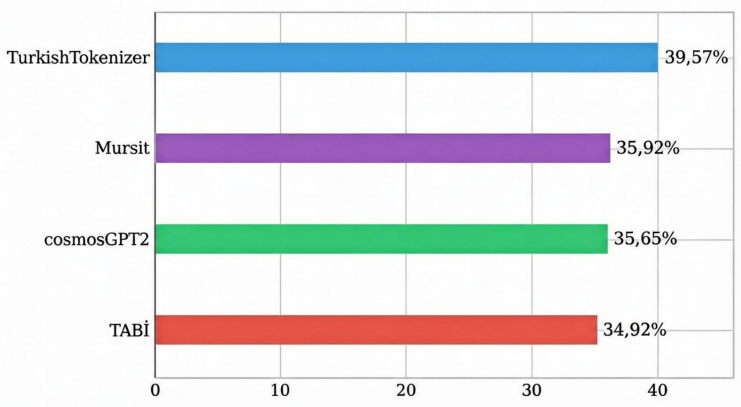
6.5.2 MTEB-TR Özeti

Kontrollü deney hattı, tek bir STS metriği ile sınırlı kalmamak için TR-MTEB protokolüne uygun 26 görevli MTEB-TR değerlendirmesi ile genişletilmiştir [85], [90]. Tablo 6.6 hem genel ortalamayı hem de kategori bazlı özetleri verir. Bulgular, MFT tabanlı modelin genel ortalamada en iyi performansı gösterdiğini; avantajın özellikle **STS** ve **Bilgi Erişim** kategorilerinde yoğunlaştığını ortaya koymaktadır.

Tablo 6.6 MTEB-TR kontrollü deney özet sonuçları.

Kategori/Özet	**MFT**	**Mursit**	**Cosmos**	**Tabi**
Ortalama (26 görev)	**39,57**	35,92	35,65	34,92
STS	**50,04**	45,73	42,14	42,54
Bilgi Erişim	**30,43**	23,27	22,43	21,17
Çift Sınıflandırma	**50,49**	47,40	48,50	47,17
Sınıflandırma	**59,03**	57,63	57,86	57,57
Kümeleme	65,49	65,83	**67,06**	65,18
Diğer	**4,59**	2,71	2,03	2,04
Bitext Eşleme	**1,89**	1,63	1,43	1,49

Bu geniş değerlendirme, tokenizer önyargısının görev tipine göre farklı yansımaları olabileceğini de görünür kılar. MFT özellikle STS ve bilgi erişim görevlerinde belirgin avantaj üretirken; bazı sınıflandırma görevlerinde (örn. *Turkish75NewsClassification*) Tabi gibi altbirim tabanlı karşılaştırıcıların daha yüksek skorlar elde edebildiği görülmektedir (Tablo 6.7). Bu nedenle kontrollü deney hattı, tokenizasyonun "tek yönlü" bir iyileştirme vaadi değil, görev rejimine bağlı bir tasarım değişkeni olduğunu göstermektedir.



Şekil 6.5 MTEB-TR ortalama skor karşılaştırması.

Şekil 6.5, 26 görev ortalamasındaki model sıralamasını görselleştirir ve MFT'nin özet skorda üstte yer aldığını tablo bulgularıyla tutarlı biçimde doğrular.

Tablo 6.7'de her satırda en iyi skor koyu yazılmıştır.

6.5.3 TurBLiMP: Minimal Çiftlerle Dilsel Duyarlılık

TurBLiMP, Türkçe için minimal cümle çiftleri üzerinden 16 farklı dilsel olguyu kapsayan bir değerlendirme kaynağıdır [124]. Bu tezdeki modeller cümle gömme encoder'ları olduğundan, TurBLiMP üzerinde olasılık tabanlı "dilbilgisel kabul edilebilirlik" yerine embedding tabanlı bir vekil ölçüt kullanılmıştır. Her olgu dosyası (good_sentence, bad_sentence) çiftlerinden oluşur; her olgu için önce "iyi" cümle embedding'lerinin merkezi (centroid) hesaplanır ve normalize edilir, ardından her cümle bu centroid'e kosinüs benzerliği ile skorlanır. Bir çift *doğru* sayılır iff $score(good) > score(bad)$; böylece olgu bazında **çift-doğruluk (%)** raporlanır. Bu ölçüm, doğrudan bir dil modeli olasılığı olmadığı için "mutlak dilbilgisel doğruluk" olarak değil; embedding uzayının kontrollü biçimsel bozulmalara karşı duyarlılığını gösteren tanılayıcı bir sinyal olarak yorumlanmalıdır.

Tablo 6.8 sonuçları, MFT'nin olgu ortalamasında en yüksek çift-doğruluğa ulaştığını (%61,2) göstermektedir. Bununla birlikte üstünlük tüm olgulara eşit dağılmamaktadır: *Edilgen Yapılar* ve *Adlaştırma* olgularında Tabi; *Bağıntı Yapısı* ve *İlgi Tümceleri* olgularında Cosmos; *Nicelendiriciler* olgusunda ise Mursit daha yüksek doğruluk üretmektedir. Bu bulgu, morfolojik hizalama önyargısının genel bir avantaj sağlayabildiğini; ancak olgu tipine göre ödünleşimlerin korunduğunu göstermektedir.

Tablo 6.7 MTEB-TR görev bazlı sonuçlar.

Görev	Kategori	MFT	Mursit	Cosmos	Tabi
STS					
STSbTR	STS	**50,04**	45,73	42,14	42,54
Bilgi Erişim					
CQADupstackGaming RetrievalTR	Bilgi Erişim	**13,17**	8,75	8,24	7,27
MSMarcoTRRetrieval	Bilgi Erişim	**15,83**	8,37	7,52	7,11
NFCorpusTR	Bilgi Erişim	**1,32**	0,59	0,51	0,24
QuoraRetrievalTR	Bilgi Erişim	**63,84**	52,20	49,02	47,46
SciFactTR	Bilgi Erişim	**23,97**	21,19	19,86	17,34
SquadTRRetrieval	Bilgi Erişim	**18,74**	10,69	9,94	8,78
TQuadRetrieval	Bilgi Erişim	**46,92**	35,43	33,81	34,96
TurkishAbstractCorpus Clustering	Bilgi Erişim	**47,83**	43,63	44,24	41,20
XQuADRetrieval	Bilgi Erişim	**42,27**	28,55	28,69	26,19
Çift Sınıflandırma					
MnliTr	Çift Sınıf.	**48,38**	46,32	46,23	45,16
SnliTr	Çift Sınıf.	**45,33**	41,47	41,11	40,35
XNLI	Çift Sınıf.	57,76	54,41	**58,17**	55,99
Sınıflandırma					
THYSentimentClassification	Sınıf.	51,53	**53,08**	51,18	47,17
TSTimelineNews CategoryClassification	Sınıf.	**50,19**	45,65	44,35	44,67
Turkish75News Classification	Sınıf.	78,00	72,67	77,33	**80,00**
TurkishIrony Classification	Sınıf.	49,50	52,67	51,17	**53,08**
TurkishMovieSentiment Classification	Sınıf.	**55,09**	54,82	53,45	54,42
TurkishNewsCategory Classification	Sınıf.	**85,28**	82,40	84,00	81,48
TurkishOffensiveLanguage Classification	Sınıf.	49,59	48,41	**50,31**	48,65
TurkishProductSentiment Classification	Sınıf.	**53,09**	51,32	51,10	51,09
Kümeleme					
TurkishColumnWriting Clustering	Küme.	65,49	65,83	**67,06**	65,18
Diğer					
ArguAnaTR	Diğer	**7,28**	3,77	3,35	2,67
FiQA2018TR	Diğer	**5,79**	3,95	2,42	2,81
SCIDOCSTR	Diğer	**0,70**	0,40	0,33	0,65
Bitext Eşleme					
WMT16Bitext Mining	Bitext	**1,89**	1,63	1,43	1,49

Tablo 6.8 TurBLiMP olgu bazlı sonuçlar.

Dilsel olgu	MFT	Tabi	Cosmos	Mursit
Ortalama (16 olgu)	**61,2**	51,6	50,8	52,7
Eksiltme	**62,9**	55,4	47,3	49,3
Serbest Sıralama	**59,1**	55,6	55,3	56,5
Belirleyiciler	**99,7**	58,4	93,6	96,3
Nicelendiriciler	25,5	19,8	21,1	**44,9**
Askıya Alınmış Ekler	**71,2**	47,3	40,4	45,4
İlgi Tümceleri	48,4	49,4	**51,5**	48,7
Bağlama	**86,7**	63,6	70,5	70,7
Artgönderim Uyumu	**52,3**	49,6	49,6	48,6
NPI Lisanslama	**71,9**	48,3	48,7	46,8
Düzensiz Biçimler	**54,9**	45,9	33,3	32,7
Bağıntı Yapısı (Çift Geçişli)	51,0	51,4	**59,3**	36,4
Özne-Yüklem Uyumu	**68,4**	60,8	55,1	56,9
Adlaştırma	45,4	**52,1**	51,7	51,2
Bağıntı Yapısı (Geçişli)	47,4	56,3	**57,0**	54,7
Edilgen Yapılar	41,9	**61,5**	43,1	54,4
Ada Etkileri	**93,1**	50,7	35,7	49,9

6.6 DQ-ALR: Veri Kalitesi Perspektifi

Bu bölüm, Bölüm 5.4'de kavramsal olarak tanımlanan DQ-ALR çerçevesinin motivasyonunu ve mevcut deneysel bağlamını tartışır. DQ-ALR, veri kalitesi sinyallerinin (hekim unvanı, uzmanlık alanı, metin tutarlılığı gibi) öğrenme sürecinde dikkate alınmasını öneren kavramsal bir yaklaşımdır. Bu perspektif, **H3** hipotezinin sınanması için teorik bir temel oluşturur.

Tez kapsamında geliştirilen sağlık veri setinde (Bölüm 4), hekim unvanlarının ve uzmanlık alanlarının dağılımı belirgin bir heterojenlik sergilemektedir [241]. Yayımlanan çalışmada [241], bu meta-bilgiler eğitim sırasında yönerge şablonunda bağlam bilgisi olarak kullanılmış; ancak örnek-bazlı öğrenme oranı ayarlaması uygulanmamıştır. DQ-ALR, bu gözlemden yola çıkarak veri kalitesi sinyallerinin yalnızca yönerge bağlamı olarak değil, doğrudan optimizasyon sürecine entegre edilmesini önermektedir.

Mevcut deneysel kanıtlar, DQ-ALR'nin kontrollü bir ortamda sistematik olarak sınanması yerine, kavramsal motivasyonun sağlık veri setindeki heterojenlik gözlemleriyle desteklenmesinden oluşmaktadır. Bu nedenle H3, gelecek çalışmalarda kontrollü deneylerle doğrulanması gereken bir kavramsal katkı olarak "kısmen desteklendi" şeklinde değerlendirilmiştir. Gelecekte, aynı veri ve başlangıç koşullarında sabit öğrenme oranlı taban model ile DQ-ALR stratejisi uygulanan modelin karşılaştırıldığı kontrollü deneylerin yürütülmesi, bu hipotezin ampirik olarak sınanması için gereklidir.

6.7 Türkçe Sentence Embedding Modeli Deneyleri (embeddingmagibu-200m)

Bu bölümde **embeddingmagibu-200m** sonuçları, STSbTR ve TR-MTEB protokolleri üzerinden raporlanmıştır. Ana performans anlatısı STSbTR (train+test) ve TR-MTEB (26 görev, 26 model) sonuçlarına dayanmaktadır. Amaç, mutlak liderlik iddiası değil; düşük maliyetli bir uyarlama hattı ile rekabetçi Türkçe embedding performansının gösterilmesidir.

6.7.1 STSbTR (Resmi Train+Test Karşılaştırması)

STSbTR için test bölümü ($n = 1379$) ve eğitim bölümü ($n = 5749$) birlikte raporlanmıştır. Tablo 6.9, 21 modelin tamamını test/train Pearson-Spearman skorlarıyla sunmaktadır. embeddingmagibu-200m, test Pearson'da **%77,55** ile **6/21**, eğitim Pearson'da **%82,35** ile **3/21** sıradadır. Spearman sıralaması testte **6/21**, train'de **4/21**'dir. Bu bulgu, modelin eğitim bölümünde üst gruba yaklaştığını; test bölümünde ise güçlü çok dilli ve Türkçe taban karşılaştırıcılarla aynı performans bandında, temkinli bir farkla konumlandığını göstermektedir.

6.7.2 TR-MTEB (26 Model, 26 Görev) Özet Karşılaştırması

Tablo 6.10, TR-MTEB sonuçlarını kategori ortalamalarıyla verir. embeddingmagibu-200m genel ortalamada **63,9** puanla **7/26** sıradadır. Kategori bazında modelin STS $(77,5)$, NLI $(67,9)$ ve BitextMining $(97,0)$ tarafında daha güçlü; Retrieval $(72,2)$ ve Classification $(68,5)$ tarafında ise üst gruba yakın ama sınırlı bir farkla geride kaldığı görülmektedir. Ayrıntılı görev kırılımı Ek A'te sunulur. Sonuçların güncel ve etkileşimli görünümü çevrim içi gezginde izlenebilir [251].

Tablo 6.11, TR-MTEB kategorileri üzerinden 200m ve 152m varyantlarını doğrudan karşılaştırır. 200m modeli özellikle STS, NLI ve BitextMining'de belirgin bir artış üretirken, Clustering tarafında iki modelin birbirine çok yakın olduğu görülmektedir.

6.7.3 TabiBench Notu (İkincil/Supplementary)

TabiBench sonuçları bu revizyonda ana performans iddiası için kullanılmamıştır; ana anlatı STSbTR+TR-MTEB üzerinde kurulmuştur. TabiBench ayrıntılı tabloları, önceki karşılaştırmaların izlenebilirliği için Ek A altında Tablo A.6 ve Tablo A.7 olarak korunmuştur.

Tablo 6.9 STSbTR sonuçları (21 model, test bölümü; tüm değerler %).

Test Sıra	Model	Test P	Test S
1	emrecan/bert-base-turkish-cased-mean-nli-stsb-tr	83,45	83,13
2	intfloat/multilingual-e5-large-instruct	80,50	81,23
3	ytu-ce-cosmos/turkish-e5-large	79,73	79,99
4	intfloat/multilingual-e5-base	78,61	78,35
5	selmanbaysan/turkish_embedding_model_fine_tuned	78,08	78,35
6	**magibu/embeddingmagibu-200m (bu tez)**	**77,55**	**77,45**
7	intfloat/multilingual-e5-small	76,50	76,15
8	newmindai/Mursit-Large-TR-Retrieval	76,42	74,60
9	trmteb/turkish-embedding-model	76,05	74,94
10	newmindai/Mursit-Base-TR-Retrieval	76,02	74,05
11	google/embeddinggemma-300m (öğretmen)	73,84	72,92
12	**magibu/embeddingmagibu-152m**	**72,92**	**71,84**
13	dbmdz/bert-base-turkish-uncased	57,32	56,72
14	Trendyol/TY-ecomm-embed-multilingual-base-v1.2.0	52,13	52,27
15	alibaba-NLP/gte-multilingual-base	51,35	52,01
16	newmindai/TurkEmbed4Retrieval	50,95	51,80
17	boun-tabilab/TabiBERT-Code	50,54	50,97
18	boun-tabilab/TabiBERT	50,54	50,97
19	newmindai/TurkEmbed4STS	50,30	51,66
20	sentence-transformers/all-mpnet-base-v2	50,13	50,11
21	KocLab-Bilkent/BERTurk-Legal	41,82	42,73

Çevrimdışı damıtma sürecinin eğitim dinamikleri, kayıp ve öğrenme oranı eğrileri üzerinden izlenebilir. Şekil 5.4 bu süreci özetlemektedir. Bu görselin temel mesajı, öğretmen uzayına hizalanma hedefinin tek epoch gibi sınırlı bir eğitim bütçesinde istikrarlı biçimde optimize edilebildiğini göstermesidir [245].

Tablo 6.9 STSbTR sonuçları (devamı): train bölümü; tüm değerler %.

Train Sıra	Model	Train P	Train S
1	emrecan/bert-base-turkish-cased-mean-nli-stsb-tr	93,79	92,31
2	intfloat/multilingual-e5-large-instruct	82,75	81,29
5	ytu-ce-cosmos/turkish-e5-large	80,90	79,06
6	intfloat/multilingual-e5-base	79,86	77,38
4	selmanbaysan/turkish_embedding_model_fine_tuned	82,15	80,61
3	**magibu/embeddingmagibu-200m (bu tez)**	**82,35**	**80,27**
7	intfloat/multilingual-e5-small	79,10	76,88
9	newmindai/Mursit-Large-TR-Retrieval	76,33	72,58
8	trmteb/turkish-embedding-model	77,07	74,87
10	newmindai/Mursit-Base-TR-Retrieval	74,47	70,72
11	google/embeddinggemma-300m (öğretmen)	73,91	71,94
12	**magibu/embeddingmagibu-152m**	**73,26**	**70,24**
13	dbmdz/bert-base-turkish-uncased	59,37	57,02
16	Trendyol/TY-ecomm-embed-multilingual-base-v1.2.0	55,64	53,78
15	alibaba-NLP/gte-multilingual-base	55,65	53,56
14	newmindai/TurkEmbed4Retrieval	55,99	54,11
19	boun-tabilab/TabiBERT-Code	51,30	50,13
20	boun-tabilab/TabiBERT	51,30	50,13
17	newmindai/TurkEmbed4STS	55,49	53,59
18	sentence-transformers/all-mpnet-base-v2	51,75	51,46
21	KocLab-Bilkent/BERTurk-Legal	36,80	37,29

Tablo 6.10 TR-MTEB 26 model/26 görev özet sonuçları (güçlü ve zayıf baseline'larla), Bölüm I.

Model	Sıra	Ort.	Retr.	Cls.	Clust.
text-embedding-3-small	1/26	66,5	78,1	69,5	62,1
ytu-ce-cosmos/ turkish-e5-large	2/26	66,0	77,0	72,6	60,7
intfloat/ multilingual-e5- large-instruct	3/26	65,9	76,1	73,7	62,1
google/ embeddinggemma-300m (öğretmen)	4/26	65,2	75,9	71,8	62,4
alibaba-NLP/ gte-multilingual-base	5/26	64,4	75,2	67,8	60,4
magibu/ embeddingmagibu-200m (bu tez)	**7/26**	**63,9**	**72,2**	**68,5**	**61,4**
selmanbaysan/ turkish-emb- model-fine- tuned	8/26	63,4	71,4	70,9	61,5
magibu/ embeddingmagibu-152m	**12/26**	**60,2**	**71,3**	**65,5**	**61,6**
trmteb/ turkish-embedding-model	14/26	59,8	69,5	70,7	63,6
dbmdz/bert-base- turkish-uncased	17/26	49,5	53,6	70,1	60,7
fasttext	18/26	46,9	52,1	63,3	59,7
KocLab-Bilkent/ BERTurk-Legal	22/26	39,3	24,6	64,8	59,1

Tablo 6.10 TR-MTEB 26 model/26 görev özet sonuçları (devamı), Bölüm II.

Model	STS	NLI	Bitext
text-embedding-3-small	70,8	57,2	91,6
ytu-ce-cosmos/ turkish-e5-large	80,0	62,7	99,2
intfloat/ multilingual-e5- large-instruct	81,2	63,0	99,0
google/embeddinggemma-300m (öğretmen)	72,9	60,6	96,8
alibaba-NLP/ gte-multilingual-base	80,7	66,7	98,0
magibu/embeddingmagibu-200m (bu tez)	**77,5**	**67,9**	**97,0**
selmanbaysan/ turkish-emb-model-fine-tuned	78,5	70,8	86,9
magibu/embeddingmagibu-152m	**71,8**	**55,6**	**90,1**
trmteb/ turkish-embedding-model	75,1	56,9	37,8
dbmdz/ bert-base-turkish-uncased	56,8	52,4	7,1
fasttext	50,1	52,1	4,5
KocLab-Bilkent/BERTurk-Legal	42,3	49,5	1,6

Tablo 6.11 200m ve 152m için TR-MTEB özet karşılaştırması (26 görev/kategori).

Ölçüt	**200m**	**152m**	**Fark (200m-152m)**
Genel ortalama (26 görev)	63,9	60,2	+3,7
Genel sıra (26 model)	7/26	12/26	+5 basamak
Retrieval ortalaması	72,2	71,3	+0,9
Classification ortalaması	68,5	65,5	+3,0
Clustering ortalaması	61,4	61,6	-0,2
STS ortalaması	77,5	71,8	+5,7
NLI ortalaması	67,9	55,6	+12,3
BitextMining ortalaması	97,0	90,1	+6,9

6.8 Sağlık Alanı Deneyleri: Türkçe Tıbbi LLM

Bu bölüm, gerçek hasta–doktor etkileşimlerinden derlenen Türkçe tıbbi soru–cevap verisi üzerinde (bkz. Bölüm 4) geliştirilen alan-özel modelin değerlendirmesini sunar [239], [241]. Deney tasarımında iki temel risk dikkate alınmıştır. Birinci risk, genel amaçlı modellerin Türkçedeki kültürel/deyimsel ifadeleri tıbbi bağlamda yanlış yorumlayabilmesidir. İkinci risk ise, alan-özel uyarlama sırasında genel dil yeteneklerinin azalması (felaket unutma) ve bunun güvenilir kullanım senaryolarını zayıflatmasıdır. Bu nedenle değerlendirme, nicel benzerlik ölçümleri, LLM tabanlı otomatik değerlendirme ve uzman hekim değerlendirmesini birlikte kullanır [258], [241]. Bu deney ailesi, PEFT temelli ince ayar ve ağırlık uzayı birleştirme yaklaşımlarının doğru veri altyapısıyla birlikte kullanıldığında uzmanlık kazanımını artırırken genel yetenek kaybını sınırlayabileceği yönündeki **H4** hipotezini sınamayı hedefler.

Değerlendirme hattının ilk bileşeni, yanıtların metin düzeyindeki yakınlığını ölçmek için kosinüs benzerliğidir. Bu ölçüm, uzman hekim yanıtları ile model yanıtlarını aynı temsil uzayına gömerek yapılan, "terim ve ifade düzeyinde yakınlık" sinyali sağlayan bir yaklaşımdır [258]. İkinci bileşen, GPT-3.5 tabanlı otomatik puanlamadır; burada amaç, yanıtların tutarlılık, açıklık ve yardımcı olma düzeyi gibi daha yüksek seviyeli özelliklerini tek bir puan altında özetleyebilmektir [241]. Üçüncü bileşen ise, uzman hekimlerin çok kategorili puanlamasıdır ve bu tezde güvenilirlik açısından en kritik ölçüm katmanını oluşturur; çünkü alan-özel yanlışlar klinik açıdan yüksek risk taşıyabilir ve otomatik metrikler her zaman bu riski yakalayamayabilir.

Tablo 6.12 temel nicel ölçümleri özetler. İlk satır grubu, uzman yanıtları ile model yanıtları arasındaki ortalama kosinüs benzerliğini verir; ince ayarlı modelin taban modele göre daha yüksek benzerlik üretmesi, terminoloji ve ifade düzeyinde "uzman dili"ne yaklaşmayı gösterir. İkinci satır grubu, GPT-3.5 tabanlı puanlamayı raporlar; bu yaklaşım, yanıtların tutarlılığı ve bilgilendiriciliği açısından bir dış değerlendirme sinyali sağlar. Üçüncü satır grubu ise, uzman hekimlerin çok kategorili puanlaması üzerinden insan değerlendirmesini verir [241].

Nitel analiz, genel amaçlı modellerin Türkçedeki mecazlı tıbbi ifadeleri yanlış yorumlayabildiğini göstermektedir. Örneğin "*Boğazıma bir şey batti*" ifadesinin coğrafi bir bağlama (Boğaz) çekilmesi gibi hatalar, alan-özel uyarlamanın yalnızca terminoloji değil, kültürel bağlam çözümleme açısından da gerekli olduğunu ortaya koyar [241]. Bu noktada LoRA ile alan uyarlaması, maliyeti düşürürken; Slerp ile birleştirme, uzmanlık kazanımını korurken genel yeteneğin geri kazanılmasına katkı sağlar (bkz. Bölüm 5).

Tablo 6.12 Sağlık alanı deneyleri özet sonuçları.

Ölçüm grubu	Karşılaştırma	Ölçek	Değer
Kosinüs benzerliği	İnce ayarlı model ile uzman yanıtı karşılaştırması	%	51,51
Kosinüs benzerliği	Taban LLaMA-3 8B ile uzman yanıtı karşılaştırması	%	46,27
Kosinüs benzerliği	Göreli iyileşme (ince ayarlı model - taban model)	%	+11,33
GPT-3.5 otomatik puanı	Uzman yanıt ortalaması	1-10	7,692
GPT-3.5 otomatik puanı	İnce ayarlı model ortalaması	1-10	7,132
Uzman hekim puanı	İnsan doktor ortalaması	1-5	3,07
Uzman hekim puanı	Türkçe Tıbbi LLM ortalaması	1-5	2,59
Uzman hekim puanı	Taban LLaMA-3 8B ortalaması	1-5	2,00

6.9 Hipotezlerin Deneysel Durumu

Bu bölümde raporlanan deneylerin hipotezlerle doğrudan eşleşmesi Tablo 6.13 ile özetlenmiştir. Böylece her hipotezin hangi deney hattıyla sınandığı, hangi ölçümlerle desteklendiği ve mevcut tez kapsamındaki değerlendirme sonucu açık biçimde izlenebilir hale gelir.

Tablo 6.13 H1-H5 hipotezleri için deneysel durum özeti.

Hipotez	Sınama hattı	Temel bulgu	Durum
H1	Bölüm 6.4, Bölüm 6.5	MFT, STSbTR'de en yüksek Pearson/Spearman değerlerine ve TR-MTEB özetinde en yüksek ortalamaya ulaşmıştır (Tablo 6.5, Tablo 6.6).	Desteklendi
H2	Bölüm 6.4	%TR / %Pure metrikleri tokenizer farklarını ayrıştırmış; %TR artışı ile başarı arasında tutarlı pozitif ilişki gözlenmiştir (Tablo 6.2, Tablo 6.3).	Desteklendi
H3	Bölüm 6.6	DQ-ALR kavramsal olarak önerilmiş; sağlık veri setindeki heterojenlik gözlemleriyle motivasyonu desteklenmiştir. Kontrollü deneysel doğrulama gelecek çalışmalara bırakılmıştır.	Kısmen desteklendi
H4	Bölüm 6.8	Alan-özel uyarlama ile uzman yanıtlarına yakınlık ve uzman puanları artmış (Tablo 6.12); çoğul ölçüm yaklaşımı (benzerlik, GPT puanlama, uzman değerlendirmesi) uzmanlık kazanımını tutarlı biçimde doğrulamıştır.	Desteklendi
H5	Bölüm 6.3	TR-MMLU, yerel/kültürel içerikte modeller arası farkları görünür kılmıştır (Tablo 6.1); ancak tez kapsamında çeviri tabanlı setlerle birebir eş-protokol karşılaştırma yapılmamıştır.	Kısmen desteklendi

6.10 Bölüm Özeti

Bu bölümde, tezde önerilen yöntemlerin deneysel doğrulaması TR-MMLU, tokenizasyon standartları, veri kalitesi stratejileri, alan-özel tıbbi model değerlendirmesi ve embedding değerlendirmeleri üzerinden raporlanmıştır. TR-MMLU sonuçları, Türkçe değerlendirmede standardizasyonun ve dil-özelleştirilmiş yaklaşımların önemini göstermiştir [237]. Tokenizasyon deneyleri, %TR ve %Pure metriklerinin model başarımıyla ilişkisinin ölçülebilir olduğunu ve hibrit tokenizasyonun dilbilimsel bütünlük açısından belirgin iyileşme sağladığını ortaya koymuştur [3], [6]. DQ-ALR, veri kalitesi sinyalinin öğrenme sürecinde dikkate alınmasının potansiyel faydalarını kavramsal olarak tartışmış; sağlık veri setindeki heterojenlik gözlemleriyle bu perspektifin motivasyonunu desteklemiştir. Embedding deneyleri, çevrimdışı damıtma ile Türkçe odaklı uzun bağlam embedding modelinin pratik biçimde üretilebildiğini ve STSbTR/TR-MTEB üzerinde rekabetçi bir performans sunduğunu; TabiBench

bulgularının ise ikincil/supplementary kapsamda değerlendirildiğini göstermiştir [245], [252]. Sağlık alanı değerlendirmeleri ise, alan-özel uyarlamanın kültürel bağlam ve uzman diline yakınlık açısından ölçülebilir kazanımlar ürettiğini doğrulamıştır [241].

Bölümün bütüncül mesajı, Türkçe için başarıyı artırmanın tek bir "büyük model" hamlesinden ziyade ölçülebilir ve birbiriyle uyumlu mühendislik kararlarının birleşimiyle mümkün olduğudur. Tokenizasyon standartları, bu kararların etkisini sayısallaştırırken; DQ-ALR veri güvenilirliğinin eğitimde dikkate alınmasını kavramsal olarak önermekte; LoRA ve Slerp gibi adaptasyon stratejileri ise alan-özel kazanımı maliyet-etkin biçimde mümkün kılmaktadır. Embedding tarafında elde edilen sonuçlar da, uzun bağlamlı Türkçe metinlerde semantik temsil kalitesinin pratik maliyetle artırılabildiğini göstermektedir [3], [262], [245].

Bu sonuçlar, tezin hipotezlerini deneysel olarak desteklerken, veri ve model kararlarının ölçülebilir etkilerini ortak bir protokol altında görünür kılmıştır. Böylece izleyen uygulama ve sonuç bölümlerinde tartışılan çıktıların deneysel temeli açık biçimde sabitlenmiştir.

7
GÖSTERİM VE DENEYİM ORTAMLARI

Bu bölüm, tez kapsamında geliştirilen yöntem ve modellerin erişilebilirliğini artırmak için oluşturulan gösterim ve deneyim ortamlarını özetler. Amaç, geliştirilen yaklaşımların yalnızca akademik çıktılar olarak kalmaması; kullanıcılar, araştırmacılar ve uygulayıcılar tarafından kolayca denenebilir, karşılaştırılabilir ve yeniden üretilebilir biçimde sunulmasıdır.

Bu kapsamda, model ve veri çıktı paketlerinin görünür kılınması, değerlendirme çıktılarının etkileşimli sunumu ve açık kaynak araçların paylaşımı birlikte ele alınır. Böylece araştırma sonuçları, sadece raporlanmış metrikler olarak değil, pratik senaryolarda doğrulanabilir deneyimler olarak da erişilebilir hale gelir.

7.1 Etkileşimli Deneyim Ortamları

Turkish Tokenizer için morfolojik bölütleme davranışının gözlemlenebildiği etkileşimli bir deneyimleme ortamı geliştirilmiştir [263]. Bu arayüz, kök–ek ayrışımını görsel olarak incelemeyi, farklı tokenizasyon seçeneklerini karşılaştırmayı ve Türkçe Token Oranı (Turkish Ratio, TR%) ile Saf Token Oranı (Purity Ratio, Pure%) metrikleriyle kalite sinyalini anlık olarak izlemeyi mümkün kılar. Arayüzde tokenlerin kök/ek/BPE türlerine göre renklendirilmesi, kodlanmış kimlik dizisinin gösterilmesi, encode-decode çıktısının karşılaştırılması ve sıkıştırma oranı gibi özet istatistiklerin sunulması; tokenizasyonun yalnızca "ön işlem" değil, ölçülebilir bir temsil bileşeni olduğunu pratik düzeyde görünür kılar.

7.1.1 Tokenizer Karşılaştırma Ortamı

Tezdeki tokenizasyon standartları sonuçlarının kullanılabilirliğini artırmak için, farklı tokenizerların aynı metin üzerinde ürettiği token ayak izini ve dilbilimsel metrikleri (%TR, %Pure) birlikte görünür kılan bir karşılaştırma deneyimleme katmanı da sunulmuştur. Bu katman, tek bir sayısal sonuç yerine; tokenizer

tercihlerinin metin uzunluğu, işlem maliyeti ve morfem bütünlüğü üzerindeki etkisini birlikte okumayı ve Tablo 6.2 ile Şekil 6.1 gibi sonuçları pratik düzeyde yeniden üretmeyi hedefler. Böylece tokenizasyon, yalnızca bir modelleme ayrıntısı olmaktan çıkarılıp ölçülebilir ve karşılaştırılabilir bir mühendislik kararı olarak sunulur.

TR-MTEB Explorer arayüzü, TR-MTEB kapsamındaki Türkçe metin gömme modellerinin tek bir liderlik ekranında karşılaştırmalı analizini sağlar [251]. Değerlendirme kapsamında **07 Şubat 2026** tarihli sonuçlarda **26 model** ve **26 görev** raporlanmakta; özet karşılaştırma altı görev ailesi (Retrieval, Classification, Clustering, STS, doğal dil çıkarımı (Natural Language Inference, NLI), BitextMining) üzerinden okunmaktadır. Arayüzün bu tasarımı, yalnızca tek bir veri setinde en iyi skoru veren modeli değil, farklı görev türlerinde dengeli genelleme yapan modelleri görünür kılar.

embeddingmagibu-152m için hazırlanan etkileşimli demo, modelin uygulama düzeyindeki davranışını doğrudan gözlemlemeyi mümkün kılar [264]. Demo; metin benzerliği, semantik arama (getirim), kümeleme, benzerlik matrisi, getirim değerlendirme, kopya bulma, toplu arama ve az örnekli sınıflandırma gibi görevleri aynı arayüzde sunar. Bu kapsam, tezin embedding sonuçlarını yalnızca tablo düzeyinde değil, etkileşimli kullanım senaryoları üzerinden de doğrulanabilir hale getirir.

embeddingmagibu-200m için farklı testlerin deneyimlenebildiği bir arayüz de hazırlanmıştır. Bu ortam, STSbTR senaryolarında ve TR-MTEB görevlerinde model çıktılarının karşılaştırmalı biçimde gözlemlenmesini sağlar. Modelin uzun bağlam desteğinin pratik sonuçları ve parametre/verimlilik dengesi, kullanıcıya şeffaf biçimde sunulur.

Geliştirilen Türkçe alan-özel modellerin teknik uzmanlığı sınırlı kullanıcılar tarafından da test edilebilmesi için web tabanlı bir sohbet uygulaması üzerinden erişim sağlanmıştır. Bu arayüz, sade bir kullanım akışı sunarken model davranışının gerçek kullanım senaryolarına yakın koşullarda gözlemlenmesine olanak tanır.

Tez kapsamında hazırlanan **17 saatlik sıfırdan LLM geliştirme eğitimi** de Udemy ve YouTube üzerinden yayımlanmıştır. Bu içerik, metodolojik katkıların eğitim boyutunda yaygınlaştırılmasını ve alanın sürdürülebilir biçimde büyümesini desteklemeyi amaçlar.

7.2 Erişilebilirlik ve Yeniden Üretilebilirlik

Geliştirilen veri setleri, modeller ve değerlendirme betikleri açık erişimle paylaşılmış; yeniden üretilebilirliği artırmak için sürümleme ve veri kartı dokümantasyonu benimsenmiştir. Bu yaklaşım, hem akademik çalışmaların karşılaştırılabilirliğini güçlendirir hem de endüstriyel uygulamalar için güvenilir bir temel sağlar. Paylaşılan çıktılar, farklı araştırmacıların aynı veri ve protokoller üzerinde sonuçları bağımsız biçimde doğrulayabilmesine imkan tanır.

7.3 Açık Kaynak Destekleri ve Topluluk Etkisi

Tez kapsamında geliştirilen ve açık kaynak olarak paylaşılan temel araçlar, üç ana bileşende toplanır. **turkish-tokenizer-rs** Rust tabanlı, yüksek performanslı Türkçe tokenizasyon kütüphanesidir. **transformer-cloner**, Transformer modellerini klonlama ve yeniden yapılandırma sürecini otomatikleştiren araçtır. **distil-trainer** ise kapsamlı bilgi damıtma eğitim çerçevesi olarak embeddingmagibu-200m hattında kullanılan damıtma iş akışını standartlaştırır.

Topluluk etkisi açısından, Hugging Face ve Ollama profilleri üzerinden paylaşılan 20'den fazla model ve veri seti yüzlerce kez indirilmiş; akademik ve endüstriyel çalışmalarda doğrudan kullanılmıştır. Ayrıca mlx ve unsloth gibi küresel ölçekte kullanılan projelere yapılan katkılar ana repoya entegre edilmiştir. Bu çıktılar, tezde önerilen yaklaşımın sürdürülebilir ve topluluk destekli bir araştırma ekosistemi oluşturduğunu göstermektedir.

7.4 Bölüm Özeti

Bu bölümde, tezde geliştirilen yöntemlerin erişilebilirliğini artırmak için oluşturulan deneyimleme ortamları ve açık kaynak çıktılar özetlenmiştir. Amaç, Türkçe NLP araştırmalarının yalnızca sonuç odaklı değil; deneyimlenebilir, karşılaştırılabilir ve topluluk tarafından sürdürülebilir biçimde geliştirilebilir bir yapıya kavuşmasını sağlamaktır.

Ayrıca veri setleri ve modeller için sürümleme, dokümantasyon ve paylaşım kanallarının nasıl yapılandırıldığı özetlenmiş; bu sayede tezin çıktılarının farklı kullanıcı profilleri tarafından yeniden kullanılabilirliği vurgulanmıştır. Böylece bölüm, metodolojik katkıları uygulamaya taşıyan "erişim katmanı"nın rolünü görünür kılmıştır.

8
SONUÇ

Bu tezde, düşük kaynaklı diller ve özellikle Türkçe için LLM'lerin daha verimli, anlamlı ve güvenilir biçimde geliştirilebilmesi amacıyla bütüncül bir çerçeve sunulmuştur. Çalışma; değerlendirme standardizasyonu, dilbilimsel bütünlüğü önceleyen tokenizasyon, veri kalitesini optimizasyona taşıyan öğrenme stratejileri ve alan-özel uyarlamayı tek bir hat üzerinde birleştirmiştir. Elde edilen bulgular, Türkçe'de yüksek performansın yalnızca daha büyük modellerle değil; dilin yapısal özelliklerini ölçülebilir hale getiren ve eğitim kararlarına yansıtan tasarım ilkeleriyle mümkün olduğunu göstermektedir.

Bu sonuç bölümü, tez boyunca ortaya konan yöntemleri, deneysel bulguları ve uygulama çıktılarıyla birlikte değerlendirerek, araştırmanın genel etkisini ve sınırlılıklarını özetler. Ayrıca gelecek çalışma alanları için somut genişleme önerileri sunarak, ortaya çıkan çerçevenin nasıl ileri taşınabileceğini tartışır.

8.1 Araştırmanın Genel Değerlendirmesi

Tez kapsamında ilk olarak, Türkçe LLM'lerin değerlendirilmesi için standart bir ölçüm zemini olarak **TR-MMLU** ekosistemi ele alınmış ve standart protokollerle birlikte kullanılmıştır. TR-MMLU karşılaştırıcısı, **62 bölüm/kategori** altında **6.200** çoktan seçmeli sorudan oluşan, Türkçe'de bilgi, muhakeme ve kavramsal anlama yeteneklerini ölçmeyi hedefleyen bir değerlendirme standardıdır. Bu standardizasyon, farklı model aileleri, tokenizasyon tercihleri ve uyarlama stratejilerinin etkisini aynı ölçüm hattında izlemeyi mümkün kılmıştır. Sonuçlar, Türkçeye daha iyi hizalanmış tokenizasyon ve alan-özel uyarlama yaklaşımlarının, çok dilli genel amaçlı modellere kıyasla daha tutarlı Türkçe performans üretebildiğini göstermiştir.

İkinci olarak, Türkçenin biçimbilimsel yapısına duyarlı tokenizasyonun ölçülebilir bir araştırma değişkeni haline gelebilmesi için %TR ve %Pure metrikleri üzerinden standardizasyon yaklaşımı benimsenmiştir. Tokenizasyon standartları, Türkçe'de

temsil bütünlüğünü ve parçalanma davranışını sayısallaştırarak, tokenizasyon tercihlerinin model başarımıyla ilişkisini görünür kılmıştır. Hibrit tokenizasyon yaklaşımı (MFT/turkish_tokenizer) ise kök–ek ayrımını daha tutarlı biçimde koruyarak dilbilimsel metriklerde belirgin iyileşme sağlamış; bu iyileşmenin TR-MMLU başarımıyla birlikte hareket edebildiği deneysel olarak gösterilmiştir. Bu bulgu, eklemeli dillerde tokenizasyonun yalnızca frekans temelli bir sıkıştırma problemi değil, aynı zamanda anlamlı parçalanmayı azaltan bir dil tasarımı problemi olduğunu desteklemektedir.

Tokenizasyon bulguları sayısal olarak özetlendiğinde, hibrit yaklaşımın TR-MMLU metni üzerinde %TR=90,29 ve %Pure=85,80 seviyesine ulaştığı (Tablo 6.3); kontrollü rastgele başlatım deney hattında ise STSbTR'de üç farklı BPE tabanlı Türkçe karşılaştırıcıya (Tabi, Cosmos ve Mursit) kıyasla en yüksek Pearson/Spearman değerlerini ürettiği görülmektedir (Tablo 6.5) [6]. En güçlü taban karşılaştırıcı olan Mursit'e göre test bölümünde +4,81/+4,31 puanlık bir fark söz konusudur. Benzer biçimde, TR-MTEB'de 26 görev ortalamasında 39,57 ile tüm karşılaştırıcılar arasında en iyi skoru üretmiştir (Tablo 6.6) [6]. Bu kazanım, token/kelime oranında artış (örn. 2,91'e karşı 1,99; Tablo 6.4) ve buna bağlı dizi uzaması gibi bir maliyetle birlikte gelmektedir.

Hibrit yaklaşımın mühendislik açısından kritik yönü, MFT'nin yalnızca bir tokenizer değil; aynı zamanda tersinir *tokenize-detokenize* çevrimi sunan bir Türkçe metin çözümleyici olarak tasarlanmasıdır. 66.547 kelimelik bir örneklemde kelime-bazlı `decode(encode(w))` tam eşleşme doğruluğu %99,48 olarak ölçülmüş; tüm metnin tek seferde encode-decode edilip kelime hizalamasıyla ölçülen tanılayıcı metrik ise %99,84 (66.442/66.547) raporlanmıştır [6]. Bildiğimiz kadarıyla, kamuya açık Türkçe araçlar arasında bu düzeyde tersinirlik ile üretim-odaklı tokenizasyon hızını birlikte raporlayan bir yaklaşım sınırlıdır; bu yönüyle MFT, Türkçe NLP için özgün ve pratik bir metin çözümleme hattı sunmaktadır.

Üçüncü katkı olarak, veri kalitesinin eğitim sürecinde dikkate alınması gerektiğini savunan **DQ-ALR** çerçevesi kavramsal olarak önerilmiştir. Bu yaklaşım, sağlık veri setindeki hekim unvanı ve uzmanlık alanı gibi meta-bilgilerin öğrenme sürecine potansiyel olarak entegre edilebileceğini tartışır. Kavramsal motivasyon, veri setindeki heterojenlik gözlemleriyle desteklenmiş olmakla birlikte, yöntemin kontrollü deneysel doğrulaması gelecek çalışmalara bırakılmıştır.

Son olarak, sağlık gibi uzmanlık gerektiren alanlarda alan-özel modelleme ele alınmıştır. Gerçek hasta-doktor etkileşimlerinden türetilen veri üzerinde alan

uyarlaması yapılmış; değerlendirme, metin benzerliği, LLM tabanlı otomatik puanlama ve uzman hekim değerlendirmesini bir araya getiren çoğul bir ölçüm yaklaşımıyla yürütülmüştür. Bu tezde ayrıca, Türkçe odaklı uzun bağlam cümle gömme modeli **embeddingmagibu-200m** geliştirilmiş; model 8K token bağlam penceresi ve 128K çok dilli tokenizer sözlüğüyle STSbTR ve TR-MTEB değerlendirmeleri üzerinde raporlanmıştır [245], [90], [250]. Sonuçlarda model, TR-MTEB genel ortalamasında **63,9** puanla **7/26** sırada; STSbTR test bölümünde **%77,55 / %77,45** (Pearson/Spearman), eğitim bölümünde ise **%82,35 / %80,27** değerlerindedir. Bu sonuçlar, en büyük modelleri geçme iddiasından ziyade, düşük eğitim maliyetiyle (yaklaşık **$5-$20**) rekabetçi bir maliyet-performans dengesi üretildiğini göstermektedir.

8.2 Tezin Katkıları

Bu tez, Türkçe doğal dil işleme araştırmalarına hem metodolojik hem de uygulamalı düzeyde katkı sunmuştur. Birinci katkı, TR-MMLU gibi yerel ve kültürel açıdan uygun bir değerlendirme standardının, karşılaştırılabilir ve izlenebilir bir ölçüm zemini olarak kullanılmasıdır. İkinci katkı, tokenizasyon standartları ile %TR / %Pure metrikleri üzerinden tokenizasyonun ölçülebilir hale getirilmesi ve hibrit tokenizasyon yaklaşımının Türkçe'nin morfolojik doğasına uygun biçimde tasarlanmasıdır. Üçüncü katkı, DQ-ALR yaklaşımı ile veri güvenilirliği sinyalinin eğitimde dikkate alınmasının potansiyel faydalarının kavramsal olarak tartışılmasıdır. Dördüncü katkı, sağlık alanında alan-özel uyarlamanın parametre-verimli tekniklerle sistemli biçimde değerlendirilmesi ve yerel veri kaynaklarının açık biçimde paylaşılmasıdır. Beşinci katkı, embeddingmagibu-200m gibi Türkçe odaklı uzun bağlam embedding modelinin geliştirilmesi ve STSbTR / TR-MTEB sonuçlarıyla maliyet-performans perspektifinde şeffaf biçimde raporlanmasıdır [245], [90], [250].

Veri ekosistemi açısından, TR-MMLU soru havuzu ve Türkçe tıbbi veri seti gibi temel kaynaklar DOI ile yayımlanmış ve lisans/erişim koşulları kayıtlar üzerinden izlenebilir hale getirilmiştir; diğer veri kaynaklarında ise lisans ve kullanım koşulları ilgili platformların veri kartları üzerinden doğrulanabilir biçimde raporlanmıştır. Bu yaklaşım, açık bilim hedefini sürdürürken lisans uyumluluğunu da ölçülebilir bir gereksinim olarak ele alır.

8.3 Hipotezlerin Nihai Değerlendirmesi

Tezin başlangıcında tanımlanan **H1-H5** hipotezleri, deneysel bulgularla birlikte nihai olarak aşağıdaki şekilde değerlendirilmiştir (ayrıntı için bkz. Tablo 6.13):

H1 (morfolojik farkındalıklı tokenizasyon daha güçlü temsiller üretir): Desteklenmiştir. Kontrollü deney hattında MFT, STSbTR ve TR-MTEB özet skorlarında en yüksek değerlere ulaşmıştır (Bölüm 6.5, Tablo 6.5, Tablo 6.6).

H2 (%TR / %Pure tokenizasyon kalitesini öngörür): Desteklenmiştir. Tokenizasyon standartları deneylerinde %TR / %Pure metrikleri model/tokenizer farklarını açıklayıcı biçimde ayrıştırmış ve başarı göstergeleriyle tutarlı ilişki sunmuştur (Bölüm 6.4, Tablo 6.2, Tablo 6.3).

H3 (veri kalitesi sinyalinin öğrenmeye entegrasyonu aynı koşullarda karşılaştırmalı üstünlük sağlayabilir): Kısmen desteklenmiştir. DQ-ALR kavramsal bir çerçeve olarak önerilmiş; sağlık veri setindeki heterojenlik gözlemleriyle motivasyonu desteklenmiştir (Bölüm 6.6). Ancak yöntemin kontrollü deneysel doğrulaması henüz gerçekleştirilmemiş olup, ampirik sınaması gelecek çalışmalara bırakılmıştır.

H4 (PEFT + ağırlık uzayı birleştirme uzmanlık kazanımı sağlarken genel yetiyi korur): Desteklenmiştir. Alan-özel uyarlama ile uzman yanıtlarına yakınlık artmış (+11,33%), GPT tabanlı otomatik puanlama ve uzman hekim değerlendirmeleri tutarlı iyileşme göstermiştir; çoğul ölçüm yaklaşımı uzmanlık kazanımını güvenilir biçimde doğrulamıştır (Bölüm 6.8, Tablo 6.12).

H5 (yerel/kültürel karşılaştırıcılar farkları daha görünür kılar): Kısmen desteklenmiştir. TR-MMLU üzerinde modeller arası farklar görünür biçimde ayrışmıştır; ancak tez kapsamında çeviri tabanlı setlerle birebir eş-protokol karşılaştırma yapılmadığı için, dış karşılaştırma iddiası sınırlı tutulmuştur (Bölüm 6.3, Tablo 6.1).

Bu nihai tablo, tezin ana iddialarının büyük bölümünün deneysel olarak doğrulandığını; kısmen desteklenen hipotezler (H3 ve H5) için ise gelecek çalışma gündeminin açık biçimde tanımlandığını göstermektedir.

8.4 Sonuçların Genel Yorumları

Deneysel gözlemler, Türkçedeki morfolojik karmaşıklığın yalnızca ön işleme aşamasında değil, temsil öğrenimi ve muhakeme davranışı üzerinde de belirleyici

olduğunu göstermiştir. Tokenizasyonun dilbilimsel bütünlüğe daha iyi hizalanması, aynı bağlam penceresinde daha fazla bilginin taşınabilmesi ve daha izlenebilir temsiller üretilebilmesi açısından kritik bir kaldıraçtır. Veri kalitesi sinyalinin eğitimde dikkate alınması, heterojen kaynaklı verilerde genelleme ve kararlılığı potansiyel olarak artırabilecek bir perspektif sunmaktadır. Alan-özel uyarlama ise, genel amaçlı modellerin uzmanlık alanlarındaki performans sınırlarını maliyet-etkin biçimde genişletebilmektedir. Bu üç yönün birlikte ele alınması, Türkçe gibi eklemeli dillerde doğruluk, bağlam bütünlüğü ve anlam tutarlılığını daha güvenilir biçimde artıran bir yaklaşım ortaya koymaktadır.

8.5 Gelecekteki Çalışmalar

Bu tez, gelecekteki araştırmalar için çok sayıda genişleme alanı açmaktadır. Morfolojik farkındalıklı tokenizasyonun çok dilli mimarilere entegrasyonu ve çapraz dil etkisinin ölçülmesi, önerilen yaklaşımın genellenebilirliğini daha net ortaya koyabilir. DQ-ALR yaklaşımının, yalnızca veri güvenilirliği değil; etik kalite, tarafsızlık ve güvenlik sinyallerini de kapsayacak biçimde genişletilmesi, yüksek riskli alanlarda daha kontrollü öğrenme davranışlarını mümkün kılabilir. Ayrıca enerji verimliliği, nicemleme farkındalıklı eğitim ve düşük bellekli donanımlarda ince ayar optimizasyonlarının sistematik biçimde test edilmesi, Türkçe odaklı modellerin sürdürülebilirliğini artırabilir. Son olarak, TR-MMLU ve benzeri görev setlerini kapsayan bütüncül bir otomatik değerlendirme platformunun geliştirilmesi, hem akademik hem de uygulamalı üretim hatlarında kalite kontrolünü standardize edebilir.

8.6 Genel Sonuç

Sonuç olarak bu tez, Türkçe'de büyük dil modellerinin performansını artırmak için değerlendirme standardizasyonu, dilbilimsel olarak hizalanmış tokenizasyon, veri kalitesini eğitime yansıtan stratejiler ve alan-özel uyarlamayı birlikte ele alan bir yaklaşım önermiştir. Bu yaklaşım, yalnızca tek bir modelin başarımını yükseltmekten ziyade, Türkçe için sürdürülebilir ve tekrar üretilebilir bir geliştirme hattı kurmayı hedeflemiştir. Tez boyunca üretilen ölçüm zemini, yöntemsel katkılar ve paylaşım çıktıları; Türkçe NLP topluluğuna model, tokenizer, veri ve değerlendirme pratikleri açısından kalıcı bir temel sunmaktadır.

8.7 Bölüm Özeti

Bu bölüm, tezin ana bulgularını bir araya getirerek katkıları, sınırlılıkları ve geleceğe dönük araştırma yönlerini bütüncül biçimde değerlendirmiştir. Böylece çalışmanın Türkçe NLP alanındaki konumu ve uzun vadeli etkisi açık biçimde ortaya konmuştur.

Özetle, tokenizasyon, veri kalitesi ve alan uyarlaması eksenlerinde geliştirilen yaklaşımın, ölçülebilir performans kazanımlarıyla desteklendiği ve açık bilim pratikleriyle kalıcı bir altyapıya dönüştürüldüğü vurgulanmıştır. Bu bütünlük, tezin yalnızca bir dizi deneysel sonuç değil, sürdürülebilir bir araştırma hattı sunduğunu göstermektedir.

KAYNAKÇA

[1] T. B. Brown ve diğ., "Language Models are Few-Shot Learners," *Advances in Neural Information Processing Systems*, c. 33, Curran Associates, Inc., 2020, ss. 1877–1901.

[2] L. Ramshaw M. Marcus, "Text Chunking using Transformation-Based Learning," *Third Workshop on Very Large Corpora*, 1995. erişim adresi: https://aclanthology.org/W95-0107/.

[3] M. A. Bayram, A. Arda Fincan, A. S. Gümüş, S. Karakaş, B. Diri S. Yıldırım, "Tokenization Standards and Evaluation in Natural Language Processing: A Comparative Analysis of Large Language Models on Turkish," *2025 33rd Signal Processing and Communications Applications Conference (SIU)*, Haz. 2025, ss. 1–4. DOI: 10.1109/SIU66497.2025.11112220. erişim adresi: https://ieeexplore.ieee.org/document/11112220.

[4] M. A. Bayram. "TR-MMLU Leaderboard," erişim tarihi 8 Nis. 2026. erişim adresi: https://huggingface.co/spaces/alibayram/turkish_mmlu_leaderboard.

[5] M. A. Bayram. "Cosmos-Corpus-00-5," erişim tarihi 8 Nis. 2026. erişim adresi: https://huggingface.co/datasets/alibayram/cosmos-corpus-00-5.

[6] M. A. Bayram ve diğ., *Tokens with Meaning: A Hybrid Tokenization Approach for NLP*, Ağus. 2025. DOI: 10.48550/arXiv.2508.14292. erişim adresi: http://arxiv.org/abs/2508.14292.

[7] M. D. Akın A. A. Akın, "Türk Dilleri İçin Açık Kaynaklı Doğal Dil İşleme Kütüphanesi: ZEMBEREK," *elektrik mühendisliği*, c. 431, ss. 38–44, Ağus. 2007.

[8] P. Gage, *A New Algorithm for Data Compression*, 1994. erişim adresi: http://www.pennelynn.com/Documents/CUJ/HTML/94HTML/19940045.HTM.

[9] R. Sennrich, B. Haddow A. Birch, "Neural Machine Translation of Rare Words with Subword Units," *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany: Association for Computational Linguistics, 2016, ss. 1715–1725. DOI: 10.18653/v1/P16-1162.

[10] M. Schuster K. Nakajima, "Japanese and Korean voice search," *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2012, ss. 5149–5152. DOI: 10.1109/ICASSP.2012.6289079. erişim adresi: https://ieeexplore.ieee.org/document/6289079.

[11] T. Kudo, "Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates," *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia: Association for Computational Linguistics, 2018, ss. 66–75. DOI: 10.18653/v1/P18-1007.

[12] C. Toraman, E. H. Yilmaz, F. Şahinuç O. Ozcelik, "Impact of Tokenization on Language Models: An Analysis for Turkish," *ACM Transactions on Asian and Low-Resource Language Information Processing*, c. 22, no. 4, ss. 1–21, Nis. 2023, ISSN: 2375-4699, 2375-4702. DOI: 10.1145/3578707. erişim adresi: http://arxiv.org/abs/2204.08832.

[13] N. Z. Kayalı S. İ. Omurca, "Hybrid Tokenization Strategy for Turkish Abstractive Text Summarization," *2024 8th International Artificial Intelligence and Data Processing Symposium (IDAP)*, Eyl. 2024, ss. 1–6. DOI: 10.1109/IDAP64064.2024.10711036. erişim adresi: https://ieeexplore.ieee.org/document/10711036/?arnumber=10711036.

[14] E. Erkaya T. Güngör, "Analysis of Subword Tokenization Approaches for Turkish Language," *2023 31st Signal Processing and Communications Applications Conference (SIU)*, Tem. 2023, ss. 1–4. DOI: 10.1109/SIU59756.2023.10223973. erişim adresi: https://ieeexplore.ieee.org/document/10223973.

[15] Y. Pan, X. Li, Y. Yang R. Dong, *Morphological Word Segmentation on Agglutinative Languages for Neural Machine Translation*, Ocak 2020. DOI: 10.48550/arXiv.2001.01589. erişim adresi: http://arxiv.org/abs/2001.01589.

[16] A. Yüksel, A. Köksal, L. K. Şenel, A. Korhonen H. Schütze, *TurkishMMLU: Measuring Massive Multitask Language Understanding in Turkish*, en, Ekim 2024. erişim adresi: https://arxiv.org/abs/2407.12402.

[17] E. Dogan ve diğ., *Türkçe Dil Modellerinin Performans Karşılaştırması Performance Comparison of Turkish Language Models*, tr, Nis. 2024. erişim adresi: http://arxiv.org/abs/2404.17010.

[18] B. Baykara T. Güngör, "Abstractive text summarization and new large-scale datasets for agglutinative languages Turkish and Hungarian," *Lang. Resour. Eval.*, c. 56, no. 3, ss. 973–1007, Eyl. 2022, ISSN: 1574-020X. DOI: 10.1007/s10579-021-09568-y. erişim adresi: https://doi.org/10.1007/s10579-021-09568-y.

[19] bezir. "MMLU-pro-TR," erişim tarihi 8 Nis. 2026. erişim adresi: https://huggingface.co/datasets/bezir/MMLU-pro-TR.

[20] K. Huang, J. Altosaar R. Ranganath, *ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission*, en, Kas. 2020. erişim adresi: http://arxiv.org/abs/1904.05342.

[21] Y. Li, Z. Li, K. Zhang, R. Dan, S. Jiang Y. Zhang, "ChatDoctor: A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge," en, *Cureus*, Haz. 2023, ISSN: 2168-8184. DOI: 10.7759/cureus.40895. erişim adresi: https://www.cureus.com/articles/152858-chatdoctor-a-medical-chat-model-fine-tuned-on-a-large-language-model-meta-ai-llama-using-medical-domain-knowledge.

[22] Y. Labrak, A. Bazoge, E. Morin, P.-A. Gourraud, M. Rouvier R. Dufour, *BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains*, en, Tem. 2024. erişim adresi: http://arxiv.org/abs/2402.10373.

[23] F. Soygazi, O. Ciftci, U. Kok S. Cengiz, "THQuAD: Turkish Historic Question Answering Dataset for Reading Comprehension," en, *2021 6th International Conference on Computer Science and Engineering (UBMK)*, Ankara, Turkey: IEEE, Eyl. 2021, ss. 215–220, ISBN: 978-1-6654-2908-5. DOI: 10.1109/UBMK52708.2021.9559013. erişim adresi: https://ieeexplore.ieee.org/document/9559013/.

[24] M. Aydogan V. Kocaman, "TRSAv1: A new benchmark dataset for classifying user reviews on Turkish e-commerce websites," *Journal of Information Science*, Şub. 2022. DOI: 10.1177/01655515221074328.

[25] S. Saxena, N. Vyas D. DeCoste, *Training With Data Dependent Dynamic Learning Rates*, en, May. 2021. erişim adresi: http://arxiv.org/abs/2105.13464.

[26] C. Toraman, "Adapting Open-Source Generative Large Language Models for Low-Resource Languages: A Case Study for Turkish," *Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024)*, J. Sälevä A. Owodunni, ed., Miami, Florida, USA: Association for Computational Linguistics, Kas. 2024, ss. 30–44. DOI: 10.18653/v1/2024.mrl-1.3. erişim adresi: https://aclanthology.org/2024.mrl-1.3/.

[27] G. Eryiğit, "ITU Turkish NLP Web Service," en, *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, Gothenburg, Sweden: Association for Computational Linguistics, 2014, ss. 1–4. DOI: 10.3115/v1/E14-2001. erişim adresi: https://aclanthology.org/E14-2001/.

[28] J. Devlin, M.-W. Chang, K. Lee K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran T. Solorio, ed.,

Minneapolis, Minnesota: Association for Computational Linguistics, Haz. 2019, ss. 4171–4186. DOI: 10.18653/v1/N19-1423. erişim adresi: https://aclanthology.org/N19-1423/.

[29] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei I. Sutskever, "Language Models are Unsupervised Multitask Learners," en,

[30] E. J. Hu ve diğ., *LoRA: Low-Rank Adaptation of Large Language Models*, en, Ekim 2021. erişim adresi: http://arxiv.org/abs/2106.09685.

[31] K. Papineni, S. Roukos, T. Ward W.-J. Zhu, "Bleu: a Method for Automatic Evaluation of Machine Translation," *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, P. Isabelle, E. Charniak D. Lin, ed., Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, Tem. 2002, ss. 311–318. DOI: 10.3115/1073083.1073135. erişim adresi: https://aclanthology.org/P02-1040/.

[32] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," *Text Summarization Branches Out*, Barcelona, Spain: Association for Computational Linguistics, Tem. 2004, ss. 74–81. erişim adresi: https://aclanthology.org/W04-1013/.

[33] umarigan. "Turkish Corpus Small," erişim tarihi 8 Nis. 2026. erişim adresi: https://huggingface.co/datasets/umarigan/turkish_corpus_small.

[34] kadirnar. "Combined Turkish Datasets v4," erişim tarihi 8 Nis. 2026. erişim adresi: https://huggingface.co/datasets/kadirnar/combined-turkish-datasets-v4.

[35] A. Safaya, E. Kurtuluş, A. Goktogan D. Yuret, "Mukayese: Turkish NLP Strikes Back," en, *Findings of the Association for Computational Linguistics: ACL 2022*, Dublin, Ireland: Association for Computational Linguistics, 2022, ss. 846–863. DOI: 10.18653/v1/2022.findings-acl.69. erişim adresi: https://aclanthology.org/2022.findings-acl.69.

[36] A. Köksal, M. Thaler, A. Imani, A. Üstün, A. Korhonen H. Schütze, *MURI: High-Quality Instruction Tuning Datasets for Low-Resource Languages via Reverse Instructions*, en, Eyl. 2024. erişim adresi: http://arxiv.org/abs/2409.12958.

[37] H. T. Kesgin ve diğ., *Introducing cosmosGPT: Monolingual Training for Turkish Language Models*, Nis. 2024. DOI: 10.48550/arXiv.2404.17336. erişim adresi: http://arxiv.org/abs/2404.17336.

[38] A. A. Cengiz ve diğ., *Evaluating the Quality of Benchmark Datasets for Low-Resource Languages: A Case Study on Turkish*, Nis. 2025. DOI: 10.48550/arXiv.2504.09714. erişim adresi: http://arxiv.org/abs/2504.09714.

[39] J. Zhao, Z. Zhang, L. Gao, Q. Zhang, T. Gui X. Huang, *LLaMA Beyond English: An Empirical Study on Language Capability Transfer*, Ocak 2024. DOI: 10.48550/arXiv.2401.01055. erişim adresi: http://arxiv.org/abs/2401.01055.

[40] M. Ismayilzada ve diğ., *Evaluating Morphological Compositional Generalization in Large Language Models*, Şub. 2025. DOI: 10.48550/arXiv.2410.12656. erişim adresi: http://arxiv.org/abs/2410.12656.

[41] E. C. Acikgoz, M. Erdogan D. Yuret, *Bridging the Bosphorus: Advancing Turkish Large Language Models through Strategies for Low-Resource Language Adaptation and Benchmarking*, May. 2024. DOI: 10.48550/arXiv.2405.04685. erişim adresi: http://arxiv.org/abs/2405.04685.

[42] B. E. Çetin, Y. Özen, E. N. Demiryılmaz, K. Engür C. Toraman, *OpenEthics: A Comprehensive Ethical Evaluation of Open-Source Generative Large Language Models*, May. 2025. DOI: 10.48550/arXiv.2505.16036. erişim adresi: http://arxiv.org/abs/2505.16036.

[43] M. K. Yuce M. F. Amasyali, "KAFA-Merge: Model Merging Method With Layer-Based Bayes+Linear Search," *IEEE Access*, c. 13, ss. 108 743–108 755, 2025, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2025.3582015. erişim adresi: https://ieeexplore.ieee.org/document/11048544/.

[44] malhajar. "OpenLLM Turkish Leaderboard v0.2," erişim tarihi 8 Nis. 2026. erişim adresi: https://huggingface.co/spaces/malhajar/OpenLLMTurkishLeaderboard_v0.2.

[45] T. Kudo, "Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates," *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, I. Gurevych Y. Miyao, ed., Melbourne, Australia: Association for Computational Linguistics, Tem. 2018, ss. 66–75. DOI: 10.18653/v1/P18-1007. erişim adresi: https://aclanthology.org/P18-1007/.

[46] K. Bostrom G. Durrett, "Byte Pair Encoding is Suboptimal for Language Model Pretraining," *Findings of the Association for Computational Linguistics: EMNLP 2020*, T. Cohn, Y. He Y. Liu, ed., Online: Association for Computational Linguistics, Kas. 2020, ss. 4617–4624. DOI: 10.18653/v1/2020.findings-emnlp.414. erişim adresi: https://aclanthology.org/2020.findings-emnlp.414/.

[47] I. Provilkov, D. Emelianenko E. Voita, "BPE-Dropout: Simple and Effective Subword Regularization," *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter J. Tetreault, ed., Online: Association for Computational Linguistics, Tem. 2020, ss. 1882–1892. DOI: 10.18653/v1/2020.acl-main.170. erişim adresi: https://aclanthology.org/2020.acl-main.170/.

[48] C. W. Schmidt ve diğ., "Tokenization Is More Than Compression," *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Y. Al-Onaizan, M. Bansal Y.-N. Chen, ed., Miami, Florida, USA: Association for Computational Linguistics, Kas. 2024, ss. 678–702. DOI: 10.18653/v1/2024.emnlp-main.40. erişim adresi: https://aclanthology.org/2024.emnlp-main.40/.

[49] K. Park, J. Lee, S. Jang D. Jung, "An Empirical Study of Tokenization Strategies for Various Korean NLP Tasks," *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, K.-F. Wong, K. Knight H. Wu, ed., Suzhou, China: Association for Computational Linguistics, Ara. 2020, ss. 133–142. DOI: 10.18653/v1/2020.aacl-main.17. erişim adresi: https://aclanthology.org/2020.aacl-main.17/.

[50] S. Ahmadi, "A Tokenization System for the Kurdish Language," *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, M. Zampieri, P. Nakov, N. Ljubešić, J. Tiedemann Y. Scherrer, ed., Barcelona, Spain (Online): International Committee on Computational Linguistics (ICCL), Ara. 2020, ss. 114–127. erişim adresi: https://aclanthology.org/2020.vardial-1.11/.

[51] H. Jabbar, *MorphPiece : A Linguistic Tokenizer for Large Language Models*, Şub. 2024. DOI: 10.48550/arXiv.2307.07262. erişim adresi: http://arxiv.org/abs/2307.07262.

[52] V. Hofmann, J. Pierrehumbert H. Schütze, "Superbizarre Is Not Superb: Derivational Morphology Improves BERT's Interpretation of Complex Words," *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, C. Zong, F. Xia, W. Li R. Navigli, ed., Online: Association for Computational Linguistics, Ağus. 2021, ss. 3594–3608. DOI: 10.18653/v1/2021.acl-long.279. erişim adresi: https://aclanthology.org/2021.acl-long.279/.

[53] N. Diewald, M. Kupietz H. Lüngen, "Tokenizing on scale.Preprocessing large text corpora on the lexical and sentence level.," en, 2022.

[54] L. Xue ve diğ., *ByT5: Towards a token-free future with pre-trained byte-to-byte models*, Mar. 2022. DOI: 10.48550/arXiv.2105.13626. erişim adresi: http://arxiv.org/abs/2105.13626.

[55] Y. Tay ve diğ., *Charformer: Fast Character Transformers via Gradient-based Subword Tokenization*, Şub. 2022. DOI: 10.48550/arXiv.2106.12672. erişim adresi: http://arxiv.org/abs/2106.12672.

[56] M. Videau, B. Y. Idrissi, A. Leite, M. Schoenauer, O. Teytaud D. Lopez-Paz, *From Bytes to Ideas: Language Modeling with Autoregressive U-Nets*, Haz. 2025. DOI: 10.48550/arXiv.2506.14761. erişim adresi: http://arxiv.org/abs/2506.14761.

[57] P. Rust, J. Pfeiffer, I. Vulić, S. Ruder I. Gurevych, "How Good is Your Tokenizer? On the Monolingual Performance of Multilingual Language Models," *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, Online: Association for Computational Linguistics, 2021, ss. 3118–3135. DOI: 10.18653/v1/2021.acl-long.243.

[58] J. Xu, H. Zhou, C. Gan, Z. Zheng L. Li, "Vocabulary Learning via Optimal Transport for Neural Machine Translation," *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, C. Zong, F. Xia, W. Li R. Navigli, ed., Online: Association for Computational Linguistics, Ağus. 2021, ss. 7361–7373. DOI: 10.18653/v1/2021.acl-long.571. erişim adresi: https://aclanthology.org/2021.acl-long.571/.

[59] M. Ali ve diğ., *Tokenizer Choice For LLM Training: Negligible or Crucial?* Mar. 2024. DOI: 10.48550/arXiv.2310.08754. erişim adresi: http://arxiv.org/abs/2310.08754.

[60] A. Petrov, E. L. Malfa, P. H. S. Torr A. Bibi, "Language Model Tokenizers Introduce Unfairness Between Languages," en,

[61] A. Thakur ve diğ., *The Art of Breaking Words: Rethinking Multilingual Tokenizer Design*, Ağus. 2025. DOI: 10.48550/arXiv.2508.06533. erişim adresi: http://arxiv.org/abs/2508.06533.

[62] M. Rashad. "Arabic Tokenizers Leaderboard," erişim tarihi 8 Nis. 2026. erişim adresi: https://huggingface.co/spaces/MohamedRashad/arabic-tokenizers-leaderboard.

[63] NbAiLab. "Tokenizer Benchmark," erişim tarihi 8 Nis. 2026. erişim adresi: https://github.com/NbAiLab/tokenizer-benchmark.

[64] RIOTU Lab. "Aranizer," erişim tarihi 8 Nis. 2026. erişim adresi: https://github.com/riotu-lab/aranizer.

[65] D. Držík F. Forgac, "Slovak morphological tokenizer using the Byte-Pair Encoding algorithm," en, *PeerJ Computer Science*, c. 10, e2465, Kas. 2024, ISSN: 2376-5992. DOI: 10.7717/peerj-cs.2465. erişim adresi: https://peerj.com/articles/cs-2465.

[66] M. A. Vasiu R. Potolea, "Enhancing Tokenization by Embedding Romanian Language Specific Morphology," *2020 IEEE 16th International Conference on Intelligent Computer Communication and Processing (ICCP)*, Eyl. 2020, ss. 243–250. DOI: 10.1109/ICCP51029.2020.9266140. erişim adresi: https://ieeexplore.ieee.org/document/9266140.

[67] Y. Zhang, Z. Lin, C. C.-c. Tong S. W.-y. Ho, "Enhancing tokenization accuracy with dynamic patterns: cumulative logic for segmenting user-generated content in logographic languages," en, *Journal of Computational Social Science*, c. 8, no. 3, s. 80, Tem. 2025, ISSN: 2432-2725. DOI: 10.1007/s42001-025-00406-7. erişim adresi: https://doi.org/10.1007/s42001-025-00406-7.

[68] B. F. P. Dossou C. C. Emezue, *Crowdsourced Phrase-Based Tokenization for Low-Resourced Neural Machine Translation: The Case of Fon Language*, Mar. 2021. DOI: 10.48550/arXiv.2103.08052. erişim adresi: http://arxiv.org/abs/2103.08052.

[69] Y. Li, J. Jiang, J. Yangji N. Ma, "Finding Better Subwords for Tibetan Neural Machine Translation," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, c. 20, no. 2, 24:1–24:11, Mar. 2021, ISSN: 2375-4699. DOI: 10.1145/3448216. erişim adresi: https://doi.org/10.1145/3448216.

[70] C. Ding, M. Utiyama E. Sumita, "NOVA: A Feasible and Flexible Annotation System for Joint Tokenization and Part-of-Speech Tagging," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, c. 18, no. 2, 17:1–17:18, 2018, ISSN: 2375-4699. DOI: 10.1145/3276773. erişim adresi: https://dl.acm.org/doi/10.1145/3276773.

[71] M. Berglund B. v. d. Merwe, "Formalizing BPE Tokenization," *Electronic Proceedings in Theoretical Computer Science*, c. 388, ss. 16–27, Eyl. 2023, ISSN: 2075-2180. DOI: 10.4204/EPTCS.388.4. erişim adresi: http://arxiv.org/abs/2309.08715.

[72] H. v. A. Neubeck Alexander, *So many tokens, so little time: Introducing a faster, more flexible byte-pair tokenizer*, en-US, Ara. 2024. erişim adresi: https://github.blog/ai-and-ml/llms/so-many-tokens-so-little-time-introducing-a-faster-more-flexible-byte-pair-tokenizer/.

[73] D. Z. Hakkani-Tür, M. Saraçlar, G. Tür, K. Oflazer D. Yuret, "Morphological Disambiguation for Turkish," en, içinde *Turkish Natural Language Processing*, K. Oflazer M. Saraçlar, ed., Cham: Springer International Publishing, 2018, ss. 53–67, ISBN: 978-3-319-90165-7. DOI: 10.1007/978-3-319-90165-7_3. erişim adresi: https://doi.org/10.1007/978-3-319-90165-7_3.

[74] S. Kulick, "Exploiting Separation of Closed-Class Categories for Arabic Tokenization and Part-of-Speech Tagging," *ACM Transactions on Asian Language Information Processing*, c. 10, no. 1, 4:1–4:18, Mar. 2011, ISSN: 1530-0226. DOI: 10.1145/1929908.1929912. erişim adresi: https://doi.org/10.1145/1929908.1929912.

[75] M. Rashad, *Arabic-Nougat: Fine-Tuning Vision Transformers for Arabic OCR and Markdown Extraction*, Kas. 2024. DOI: 10.48550/arXiv.2411.17835. erişim adresi: http://arxiv.org/abs/2411.17835.

[76] A. Nayak, H. Timmapathini, K. Ponnalagu V. Gopalan Venkoparao, "Domain adaptation challenges of BERT in tokenization and sub-word representations of Out-of-Vocabulary words," *Proceedings of the First Workshop on Insights from Negative Results in NLP*, A. Rogers, J. Sedoc A. Rumshisky, ed., Online: Association for Computational Linguistics, Kas. 2020, ss. 1–5. DOI: 10.18653/v1/2020.insights-1.1. erişim adresi: https://aclanthology.org/2020.insights-1.1/.

[77] T. Hiraoka, S. Takase, K. Uchiumi, A. Keyaki N. Okazaki, "Joint Optimization of Tokenization and Downstream Model," *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, C. Zong, F. Xia, W. Li R. Navigli, ed., Online: Association for Computational Linguistics, Ağus. 2021, ss. 244–255. DOI: 10.18653/v1/2021.findings-acl.21. erişim adresi: https://aclanthology.org/2021.findings-acl.21/.

[78] X. Zhang, P. Li H. Li, "AMBERT: A Pre-trained Language Model with Multi-Grained Tokenization," *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, C. Zong, F. Xia, W. Li R. Navigli, ed., Online: Association for Computational Linguistics, Ağus. 2021, ss. 421–435. DOI: 10.18653/v1/2021.findings-acl.37. erişim adresi: https://aclanthology.org/2021.findings-acl.37/.

[79] H. Kaing ve diğ., "Towards Tokenization and Part-of-Speech Tagging for Khmer: Data and Discussion," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, c. 20, no. 6, 104:1–104:16, 2021, ISSN: 2375-4699. DOI: 10.1145/3464378. erişim adresi: https://dl.acm.org/doi/10.1145/3464378.

[80] C. Ding ve diğ., "Towards Burmese (Myanmar) Morphological Analysis: Syllable-based Tokenization and Part-of-speech Tagging," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, c. 19, no. 1, 5:1–5:34, May. 2019, ISSN: 2375-4699. DOI: 10.1145/3325885. erişim adresi: https://dl.acm.org/doi/10.1145/3325885.

[81] Z. Lin ve diğ., "Not All Tokens Are What You Need for Pretraining," en,

[82] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy S. Bowman, "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding," *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, T. Linzen, G. Chrupała A. Alishahi, ed., Brussels, Belgium: Association for Computational Linguistics, Kas. 2018, ss. 353–355. DOI: 10.18653/v1/W18-5446. erişim adresi: https://aclanthology.org/W18-5446/.

[83] A. Wang ve diğ., *SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems*, Şub. 2020. DOI: 10.48550/arXiv.1905.00537. erişim adresi: http://arxiv.org/abs/1905.00537.

[84] P. Rajpurkar, J. Zhang, K. Lopyrev P. Liang, *SQuAD: 100,000+ Questions for Machine Comprehension of Text*, Ekim 2016. DOI: 10.48550/arXiv.1606.05250. erişim adresi: http://arxiv.org/abs/1606.05250.

[85] N. Muennighoff, N. Tazi, L. Magne N. Reimers, *MTEB: Massive Text Embedding Benchmark*, https://arxiv.org/abs/2210.07316, 2022.

[86] K. Enevoldsen ve diğ., *MMTEB: Massive Multilingual Text Embedding Benchmark*, https://arxiv.org/abs/2502.13595, 2025.

[87] M. Ciancone, I. Kerboua, M. Schaeffer W. Siblini, *MTEB-French: Resources for French Sentence Embedding Evaluation and Analysis*, https://arxiv.org/abs/2405.20468, 2024.

[88] R. Poświata, S. Dadas M. Perełkiewicz, *PL-MTEB: Polish Massive Text Embedding Benchmark*, https://arxiv.org/abs/2405.10138, 2024.

[89] A. Snegirev, M. Tikhonova, A. Maksimova, A. Fenogenova A. Abramov, *The Russian-focused embedders' exploration: ruMTEB benchmark and Russian embedding model design*, https://arxiv.org/abs/2408.12503, 2024.

[90] S. M. Baysan T. Güngör, *TR-MTEB: A Comprehensive Benchmark and Embedding Model Suite for Turkish Sentence Representations*, https://aclanthology.org/2025.findings-emnlp.471/, 2025.

[91] S. M. Baysan. "MTEB-TR," erişim tarihi 8 Nis. 2026. erişim adresi: https://github.com/selmanbaysan/mteb_tr.

[92] A. Conneau ve diğ., *XNLI: Evaluating Cross-lingual Sentence Representations*, Eyl. 2018. DOI: 10.48550/arXiv.1809.05053. erişim adresi: http://arxiv.org/abs/1809.05053.

[93] E. M. Ponti, G. Glavaš, O. Majewska, Q. Liu, I. Vulić A. Korhonen, "XCOPA: A Multilingual Dataset for Causal Commonsense Reasoning," en, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online: Association for Computational Linguistics, 2020, ss. 2362–2376. DOI: 10.18653/v1/2020.emnlp-main.185. erişim adresi: https://aclanthology.org/2020.emnlp-main.185/.

[94] Y. Liang ve diğ., "XGLUE: A New Benchmark Dataset for Cross-lingual Pre-training, Understanding and Generation," *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He Y. Liu, ed., Online: Association for Computational Linguistics, Kas. 2020, ss. 6008–6018. DOI: 10.18653/v1/2020.emnlp-main.484. erişim adresi: https://aclanthology.org/2020.emnlp-main.484/.

[95] S. Longpre, Y. Lu J. Daiber, "MKQA: A Linguistically Diverse Benchmark for Multilingual Open Domain Question Answering," *Transactions of the Association for Computational Linguistics*, c. 9, ss. 1389–1406, Ara. 2021, ISSN: 2307-387X. DOI: 10.1162/tacl_a_00433. erişim adresi: https://doi.org/10.1162/tacl_a_00433.

[96] D. Hendrycks ve diğ., "Measuring Massive Multitask Language Understanding," *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. erişim adresi: https://openreview.net/forum?id=d7KBjmI3GmQ.

[97] Y. Wang ve diğ., *MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark (Published at NeurIPS 2024 Track Datasets and Benchmarks)*, en, Ekim 2024. erişim adresi: http://arxiv.org/abs/2406.01574.

[98] K. Cobbe ve diğ., *Training Verifiers to Solve Math Word Problems*, en, Kas. 2021. erişim adresi: http://arxiv.org/abs/2110.14168.

[99] D. Hendrycks ve diğ., *Measuring Mathematical Problem Solving With the MATH Dataset*, en, Kas. 2021. erişim adresi: http://arxiv.org/abs/2103.03874.

[100] I. Mirzadeh, K. Alizadeh, H. Shahrokhi, O. Tuzel, S. Bengio M. Farajtabar, *GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models*, en, Ekim 2024. erişim adresi: http://arxiv.org/abs/2410.05229.

[101] P. Clark ve diğ., *Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge*, en, Mar. 2018. erişim adresi: http://arxiv.org/abs/1803.05457.

[102] R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi Y. Choi, "HellaSwag: Can a Machine Really Finish Your Sentence?" *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy: Association for Computational Linguistics, 2019, ss. 4791–4800. DOI: 10.18653/v1/P19-1472.

[103] K. Sakaguchi, R. Le Bras, C. Bhagavatula Y. Choi, "WinoGrande: An Adversarial Winograd Schema Challenge at Scale," *Proceedings of the AAAI Conference on Artificial Intelligence*, c. 34, 2020, ss. 8732–8740. DOI: 10.1609/aaai.v34i05.6399.

[104] M. Suzgun ve diğ., *Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them*, en, Ekim 2022. erişim adresi: http://arxiv.org/abs/2210.09261.

[105] S. Lin, J. Hilton O. Evans, *TruthfulQA: Measuring How Models Mimic Human Falsehoods*, en, May. 2022. erişim adresi: http://arxiv.org/abs/2109.07958.

[106] K. Ethayarajh D. Jurafsky, "Utility is in the Eye of the User: A Critique of NLP Leaderboards," en, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online: Association for Computational Linguistics, 2020, ss. 4846–4853. DOI: 10.18653/v1/2020.emnlp-main.393. erişim adresi: https://aclanthology.org/2020.emnlp-main.393/.

[107] Z. Sun ve diğ., *PHEE: A Dataset for Pharmacovigilance Event Extraction from Text*, en, Ekim 2022. erişim adresi: http://arxiv.org/abs/2210.12560.

[108] Z. Fan ve diğ., *AI Hospital: Benchmarking Large Language Models in a Multi-agent Medical Interaction Simulator*, en, Haz. 2024. erişim adresi: http://arxiv.org/abs/2402.09742.

[109] J. Zhou ve diğ., *Instruction-Following Evaluation for Large Language Models*, en, Kas. 2023. erişim adresi: http://arxiv.org/abs/2311.07911.

[110] W.-L. Chiang ve diğ., *Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference*, en, Mar. 2024. erişim adresi: http://arxiv.org/abs/2403.04132.

[111] F. F. Xu, U. Alon, G. Neubig V. J. Hellendoorn, "A systematic evaluation of large language models of code," *Proceedings of the 6th ACM SIGPLAN International Symposium on Machine Programming*, seri MAPS 2022, New York, NY, USA: Association for Computing Machinery, 2022, ss. 1–10, ISBN: 978-1-4503-9273-0. DOI: 10.1145/3520312.3534862. erişim adresi: https://dl.acm.org/doi/10.1145/3520312.3534862.

[112] L. Bandarkar ve diğ., "The Belebele Benchmark: a Parallel Reading Comprehension Dataset in 122 Language Variants," *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, ss. 749–775. DOI: 10.18653/v1/2024.acl-long.44. erişim adresi: http://arxiv.org/abs/2308.16884.

[113] X. Yue ve diğ., *Pangea: A Fully Open Multilingual Multimodal LLM for 39 Languages*, en, Ekim 2024. erişim adresi: http://arxiv.org/abs/2410.16153.

[114] T. Nguyen ve diğ., "CulturaX: A Cleaned, Enormous, and Multilingual Dataset for Large Language Models in 167 Languages," *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti N. Xue, ed., Torino, Italia: ELRA ve ICCL, May. 2024, ss. 4226–4237. erişim adresi: https://aclanthology.org/2024.lrec-main.377/.

[115] P. J. Ortiz Suárez, B. Sagot L. Romary, "Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures," eng, Leibniz-Institut für Deutsche Sprache, Tem. 2019, ss. 9–16. DOI: 10.14618/ids-pub-9021. erişim adresi: https://ids-pub.bsz-bw.de/frontdoor/index/index/docId/9021.

[116] S. Tedeschi ve diğ., *What's the Meaning of Superhuman Performance in Today's NLU?* May. 2023. DOI: 10.48550/arXiv.2305.08414. erişim adresi: http://arxiv.org/abs/2305.08414.

[117] P. Verga ve diğ., *Replacing Judges with Juries: Evaluating LLM Generations with a Panel of Diverse Models*, May. 2024. DOI: 10.48550/arXiv.2404.18796. erişim adresi: http://arxiv.org/abs/2404.18796.

[118] A. Bavaresco ve diğ., *LLMs instead of Human Judges? A Large Scale Empirical Study across 20 NLP Evaluation Tasks*, Haz. 2025. DOI: 10.48550/arXiv.2406.18403. erişim adresi: http://arxiv.org/abs/2406.18403.

[119] A. Srivastava ve diğ., *Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models*, Haz. 2023. DOI: 10.48550/arXiv.2206.04615. erişim adresi: http://arxiv.org/abs/2206.04615.

[120] M. Mitchell ve diğ., "Model Cards for Model Reporting," *Proceedings of the Conference on Fairness, Accountability, and Transparency*, seri FAT* '19, New York, NY, USA: Association for Computing Machinery, 2019, ss. 220–229, ISBN: 978-1-4503-6125-5. DOI: 10.1145/3287560.3287596. erişim adresi: https://doi.org/10.1145/3287560.3287596.

[121] E. M. Bender, T. Gebru, A. McMillan-Major S. Shmitchell, "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, seri FAccT '21, New York, NY, USA: Association for Computing Machinery, Mar. 2021, ss. 610–623, ISBN: 978-1-4503-8309-7. DOI: 10.1145/3442188.3445922. erişim adresi: https://dl.acm.org/doi/10.1145/3442188.3445922.

[122] Ç. Çöltekin, "A Corpus of Turkish Offensive Language on Social Media," eng, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, N. Calzolari ve diğ., ed., Marseille, France: European Language Resources Association, May. 2020, ss. 6174–6184, ISBN: 979-10-95546-34-4. erişim adresi: https://aclanthology.org/2020.lrec-1.758/.

[123] F. Beken Fikri, K. Oflazer B. Yanikoglu, "Semantic Similarity Based Evaluation for Abstractive News Summarization," *Proceedings of the First Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, A. Bosselut ve diğ., ed., Online: Association for Computational Linguistics, Ağus. 2021, ss. 24–33. DOI: 10.18653/v1/2021.gem-1.3. erişim adresi: https://aclanthology.org/2021.gem-1.3/.

[124] E. Başar, F. Padovani, J. Jumelet A. Bisazza, "TurBLiMP: A Turkish Benchmark of Linguistic Minimal Pairs," *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, C. Christodoulopoulos, T. Chakraborty, C. Rose V. Peng, ed., Suzhou, China: Association for Computational Linguistics, Kas. 2025, ss. 16 495–16 510, ISBN: 979-8-89176-332-6. DOI: 10.18653/v1/2025.emnlp-main.834. erişim tarihi 27 Ocak 2026. erişim adresi: https://aclanthology.org/2025.emnlp-main.834/.

[125] tesibio. "DoctorGPT," erişim tarihi 8 Nis. 2026. erişim adresi: https://github.com/tesibio/DoctorGPT/blob/main/README.md.

[126] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio L. Specia, "SemEval-2017 Task 1: Semantic Textual Similarity - Multilingual and Cross-lingual Focused Evaluation," *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 2017, ss. 1–14. DOI: 10.18653/v1/S17-2001. erişim adresi: http://arxiv.org/abs/1708.00055.

[127] T. Hasan ve diğ., *XL-Sum: Large-Scale Multilingual Abstractive Summarization for 44 Languages*, Haz. 2021. DOI: 10.48550/arXiv.2106.13822. erişim adresi: http://arxiv.org/abs/2106.13822.

[128] Z. Sprague, X. Ye, K. Bostrom, S. Chaudhuri G. Durrett, *MuSR: Testing the Limits of Chain-of-thought with Multistep Soft Reasoning*, en, Mar. 2024. erişim adresi: http://arxiv.org/abs/2310.16049.

[129] D. Rein ve diğ., *GPQA: A Graduate-Level Google-Proof Q&A Benchmark*, en, Kas. 2023. erişim adresi: http://arxiv.org/abs/2311.12022.

[130] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger Y. Artzi, *BERTScore: Evaluating Text Generation with BERT*, en, Şub. 2020. erişim adresi: http://arxiv.org/abs/1904.09675.

[131] R. Snow, B. O'Connor, D. Jurafsky A. Ng, "Cheap and Fast – But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks," *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, M. Lapata H. T. Ng, ed., Honolulu, Hawaii: Association for Computational Linguistics, Ekim 2008, ss. 254–263. erişim adresi: https://aclanthology.org/D08-1027/.

[132] A. Aynetdinov A. Akbik, *SemScore: Automated Evaluation of Instruction-Tuned LLMs based on Semantic Textual Similarity*, en, Şub. 2024. erişim adresi: http://arxiv.org/abs/2401.17072.

[133] M. Aldeen ve diğ., "ChatGPT vs. Human Annotators: A Comprehensive Analysis of ChatGPT for Text Annotation," *2023 International Conference on Machine Learning and Applications (ICMLA)*, Ara. 2023, ss. 602–609. DOI: 10.1109/ICMLA58977.2023.00089. erişim adresi: https://ieeexplore.ieee.org/document/10460013.

[134] Z. A. Nazi W. Peng, *Large language models in healthcare and medical domain: A review*, en, Tem. 2024. erişim adresi: http://arxiv.org/abs/2401.06775.

[135] U. Mumtaz, A. Ahmed S. Mumtaz, "LLMs-Healthcare: Current applications and challenges of large language models in various medical specialties," en, *Artificial Intelligence in Health*, c. 1, no. 2, s. 16, Nis. 2024, ISSN: 3029-2387. DOI: 10.36922/aih.2558. erişim adresi: https://accscience.com/journal/AIH/1/2/10.36922/aih.2558.

[136] S. Bedi ve diğ., *A Systematic Review of Testing and Evaluation of Healthcare Applications of Large Language Models (LLMs)*, en, Nis. 2024. DOI: 10.1101/2024.04.15.24305869. erişim adresi: http://medrxiv.org/lookup/doi/10.1101/2024.04.15.24305869.

[137] X. Zhang, S. Li, B. Hauer, N. Shi G. Kondrak, "Don't Trust ChatGPT when your Question is not in English: A Study of Multilingual Abilities and Types of LLMs," *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino K. Bali, ed., Singapore: Association for Computational Linguistics, Ara. 2023, ss. 7915–7927. DOI: 10.18653/v1/2023.emnlp-main.491. erişim adresi: https://aclanthology.org/2023.emnlp-main.491/.

[138] W. Wang ve diğ., *All Languages Matter: On the Multilingual Safety of Large Language Models*, Haz. 2024. DOI: 10.48550/arXiv.2310.00905. erişim adresi: http://arxiv.org/abs/2310.00905.

[139] Y. Huang ve diğ., *TrustLLM: Trustworthiness in Large Language Models*, Eyl. 2024. DOI: 10.48550/arXiv.2401.05561. erişim adresi: http://arxiv.org/abs/2401.05561.

[140] T. Linzen, "How Can We Accelerate Progress Towards Human-like Linguistic Generalization?" en, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online: Association for Computational Linguistics, 2020, ss. 5210–5217. DOI: 10.18653/v1/2020.acl-main.465. erişim adresi: https://aclanthology.org/2020.acl-main.465/.

[141] P. Henderson, J. Hu, J. Romoff, E. Brunskill, D. Jurafsky J. Pineau, *Towards the Systematic Reporting of the Energy and Carbon Footprints of Machine Learning*, Kas. 2022. DOI: 10.48550/arXiv.2002.05651. erişim adresi: http://arxiv.org/abs/2002.05651.

[142] S. Bedi ve diğ., "Testing and Evaluation of Health Care Applications of Large Language Models: A Systematic Review," *JAMA*, Ekim 2024, ISSN: 0098-7484. DOI: 10.1001/jama.2024.21700. erişim adresi: https://doi.org/10.1001/jama.2024.21700.

[143] R. Baeza-Yates, "Ethical Challenges in AI," *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, seri WSDM '22, New York, NY, USA: Association for Computing Machinery, 2022, ss. 1–2, ISBN: 978-1-4503-9132-0. DOI: 10.1145/3488560.3498370. erişim adresi: https://doi.org/10.1145/3488560.3498370.

[144] A. Vaswani ve diğ., "Attention is All you Need," *Advances in Neural Information Processing Systems*, c. 30, Curran Associates, Inc., 2017. erişim adresi: https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[145] N. Khatri, T. Laakkonen, J. Liu V. Wang-Maścianica, *On the Anatomy of Attention*, en, Tem. 2024. erişim adresi: http://arxiv.org/abs/2407.02423.

[146] F. Locatello ve diğ., "Object-Centric Learning with Slot Attention," *Advances in Neural Information Processing Systems*, c. 33, Curran Associates, Inc., 2020, ss. 11 525–11 538. erişim adresi: https://proceedings.neurips.cc/paper_files/paper/2020/hash/8511df98c02ab60aea1b2356c013bc0f-Abstract.html.

[147] Y. Liu ve diğ., *RoBERTa: A Robustly Optimized BERT Pretraining Approach*, Tem. 2019. DOI: 10.48550/arXiv.1907.11692. erişim adresi: http://arxiv.org/abs/1907.11692.

[148] OpenAI ve diğ., *GPT-4 Technical Report*, Mar. 2024. DOI: 10.48550/arXiv.2303.08774. erişim adresi: http://arxiv.org/abs/2303.08774.

[149] A. Grattafiori ve diğ., *The Llama 3 Herd of Models*, Kas. 2024. DOI: 10.48550/arXiv.2407.21783. erişim adresi: http://arxiv.org/abs/2407.21783.

[150] M. Lewis ve diğ., "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension," *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online: Association for Computational Linguistics, 2020, ss. 7871–7880. DOI: 10.18653/v1/2020.acl-main.703.

[151] G. Uludoğan, Z. Y. Balal, F. Akkurt, M. Türker, O. Güngör S. Üsküdarlı, *TURNA: A Turkish Encoder-Decoder Language Model for Enhanced Understanding and Generation*, Ocak 2024. DOI: 10.48550/arXiv.2401.14373. erişim adresi: http://arxiv.org/abs/2401.14373.

[152] A. Q. Jiang ve diğ., *Mixtral of Experts*, en, Ocak 2024. erişim adresi: http://arxiv.org/abs/2401.04088.

[153] Z. Zhong, M. Xia, D. Chen M. Lewis, *Lory: Fully Differentiable Mixture-of-Experts for Autoregressive Language Model Pre-training*, en, May. 2024. erişim adresi: http://arxiv.org/abs/2405.03133.

[154] O. Lieber ve diğ., *Jamba: A Hybrid Transformer-Mamba Language Model*, en, Mar. 2024. erişim adresi: http://arxiv.org/abs/2403.19887.

[155] P. Glorioso ve diğ., *Zamba: A Compact 7B SSM Hybrid Model*, May. 2024. DOI: 10.48550/arXiv.2405.16712. erişim adresi: http://arxiv.org/abs/2405.16712.

[156] J. Park ve diğ., *Can Mamba Learn How to Learn? A Comparative Study on In-Context Learning Tasks*, en, Nis. 2024. erişim adresi: http://arxiv.org/abs/2402.04248.

[157] B. Workshop ve diğ., *BLOOM: A 176B-Parameter Open-Access Multilingual Language Model*, Haz. 2023. DOI: 10.48550/arXiv.2211.05100. erişim adresi: http://arxiv.org/abs/2211.05100.

[158] O. Shliazhko, A. Fenogenova, M. Tikhonova, V. Mikhailov, A. Kozlova T. Shavrina, *mGPT: Few-Shot Learners Go Multilingual*, Ekim 2023. DOI: 10.48550/arXiv.2204.07580. erişim adresi: http://arxiv.org/abs/2204.07580.

[159] P. Lin, S. Ji, J. Tiedemann, A. F. T. Martins H. Schütze, *MaLA-500: Massive Language Adaptation of Large Language Models*, Nis. 2024. DOI: 10.48550/arXiv.2401.13303. erişim adresi: http://arxiv.org/abs/2401.13303.

[160] M. Abdin ve diğ., *Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone*, Ağus. 2024. DOI: 10.48550/arXiv.2404.14219. erişim adresi: http://arxiv.org/abs/2404.14219.

[161] M. Abdin ve diğ., *Phi-4 Technical Report*, Ara. 2024. DOI: 10.48550/arXiv.2412.08905. erişim adresi: http://arxiv.org/abs/2412.08905.

[162] D. Hillier, L. Guertler, C. Tan, P. Agrawal, C. Ruirui B. Cheng, *Super Tiny Language Models*, en, May. 2024. erişim adresi: http://arxiv.org/abs/2405.14159.

[163] G. Team ve diğ., *Gemma 2: Improving Open Language Models at a Practical Size*, Ağus. 2024. DOI: 10.48550/arXiv.2408.00118. erişim adresi: http://arxiv.org/abs/2408.00118.

[164] A. Q. Jiang ve diğ., *Mistral 7B*, Ekim 2023. DOI: 10.48550/arXiv.2310.06825. erişim adresi: http://arxiv.org/abs/2310.06825.

[165] B. Rozière ve diğ., *Code Llama: Open Foundation Models for Code*, en, Ocak 2024. erişim adresi: http://arxiv.org/abs/2308.12950.

[166] J. Zhao, Z. Zhang, B. Chen, Z. Wang, A. Anandkumar Y. Tian, *GaLore: Memory-Efficient LLM Training by Gradient Low-Rank Projection*, en, Mar. 2024. erişim adresi: http://arxiv.org/abs/2403.03507.

[167] C. Yang ve diğ., *Large Language Models as Optimizers*, en, Nis. 2024. erişim adresi: http://arxiv.org/abs/2309.03409.

[168] bitsandbytes-foundation. "bitsandbytes," erişim tarihi 8 Nis. 2026. erişim adresi: https://github.com/bitsandbytes-foundation/bitsandbytes.

[169] G. Yenduri ve diğ., *Generative Pre-trained Transformer: A Comprehensive Review on Enabling Technologies, Potential Applications, Emerging Challenges, and Future Directions*, May. 2023. DOI: 10.48550/arXiv.2305.10435. erişim adresi: http://arxiv.org/abs/2305.10435.

[170] *Free Dolly: Introducing the World's First Truly Open Instruction-Tuned LLM*, en-US, Nis. 2023. erişim adresi: https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm.

[171] Loodos. "Turkish Language Models," erişim tarihi 8 Nis. 2026. erişim adresi: https://github.com/loodos/turkish-language-models.

[172] *[2101.09635] WangchanBERTa: Pretraining transformer-based Thai Language Models*. erişim adresi: https://arxiv.org/abs/2101.09635.

[173] T. Dettmers, A. Pagnoni, A. Holtzman L. Zettlemoyer, *QLoRA: Efficient Finetuning of Quantized LLMs*, en, May. 2023. erişim adresi: http://arxiv.org/abs/2305.14314.

[174] T. Jiang ve diğ., *MoRA: High-Rank Updating for Parameter-Efficient Fine-Tuning*, en, May. 2024. erişim adresi: http://arxiv.org/abs/2405.12130.

[175] Z. Wu ve diğ., *ReFT: Representation Finetuning for Language Models*, en, May. 2024. erişim adresi: http://arxiv.org/abs/2404.03592.

[176] K. Chen, Y. Pang Z. Yang, *Parameter-Efficient Fine-Tuning With Adapters*, en, May. 2024. erişim adresi: http://arxiv.org/abs/2405.05493.

[177] K. Gupta ve diğ., *Continual Pre-Training of Large Language Models: How to (re)warm your model?* Eyl. 2023. DOI: 10.48550/arXiv.2308.04014. erişim adresi: http://arxiv.org/abs/2308.04014.

[178] Z. Ke, Y. Shao, H. Lin, T. Konishi, G. Kim B. Liu, *Continual Pre-training of Language Models*, en, Nis. 2023. erişim adresi: http://arxiv.org/abs/2302.03241.

[179] Y. Luo, Z. Yang, F. Meng, Y. Li, J. Zhou Y. Zhang, *An Empirical Study of Catastrophic Forgetting in Large Language Models During Continual Fine-tuning*, en, Nis. 2024. erişim adresi: http://arxiv.org/abs/2308.08747.

[180] C.-A. Li H.-Y. Lee, *Examining Forgetting in Continual Pre-training of Aligned Large Language Models*, Ocak 2024. DOI: 10.48550/arXiv.2401.03129. erişim adresi: http://arxiv.org/abs/2401.03129.

[181] R. M. French, "Catastrophic forgetting in connectionist networks," *Trends in Cognitive Sciences*, c. 3, no. 4, ss. 128–135, Nis. 1999, ISSN: 1364-6613. DOI: 10.1016/S1364-6613(99)01294-2. erişim adresi: https://www.sciencedirect.com/science/article/pii/S1364661399012942.

[182] R. Kong ve diğ., *LoRA-Switch: Boosting the Efficiency of Dynamic LLM Adapters via System-Algorithm Co-design*, en, May. 2024. erişim adresi: http://arxiv.org/abs/2405.17741.

[183] T. Addair G. Angus, *LoRA Exchange (LoRAX): Serve 100s of Fine-Tuned LLMs for the Cost of 1 - Predibase*, en, Ekim 2023. erişim adresi: https://predibase.com/blog/lora-exchange-lorax-serve-100s-of-fine-tuned-llms-for-the-cost-of-one.

[184] I. Tufanov, K. Hambardzumyan, J. Ferrando E. Voita, *LM Transparency Tool: Interactive Tool for Analyzing Transformer Language Models*, en, Nis. 2024. erişim adresi: http://arxiv.org/abs/2404.07004.

[185] R. Jha, C. Zhang, V. Shmatikov J. X. Morris, *Harnessing the Universal Geometry of Embeddings*, May. 2025. DOI: 10.48550/arXiv.2505.12540. erişim adresi: http://arxiv.org/abs/2505.12540.

[186] T. Han ve diğ., *MedAlpaca - An Open-Source Collection of Medical Conversational AI Models and Training Data*, en, Ekim 2023. erişim adresi: http://arxiv.org/abs/2304.08247.

[187] J. Li ve diğ., *Agent Hospital: A Simulacrum of Hospital with Evolvable Medical Agents*, en, May. 2024. erişim adresi: http://arxiv.org/abs/2405.02957.

[188] E. Kasneci ve diğ., "ChatGPT for good? On opportunities and challenges of large language models for education," *Learning and Individual Differences*, c. 103, s. 102 274, Nis. 2023, ISSN: 1041-6080. DOI: 10.1016/j.lindif.2023.102274. erişim adresi: https://www.sciencedirect.com/science/article/pii/S1041608023000195.

[189] M. Seyrek, S. Yıldız, H. Emeksiz, A. Şahin M. T. Türkmen, "Öğretmenlerin Eğitimde Yapay Zeka Kullanımına Yönelik Algıları," tr, Nis. 2024, ISSN: 2459-1149. DOI: 10.5281/ZENODO.11113077. erişim adresi: https://zenodo.org/doi/10.5281/zenodo.11113077.

[190] Z. Yang, L. Laki, T. Váradi G. Prószéky, "Mono- and Multilingual GPT-3 Models for Hungarian," içinde, Ağus. 2023, ss. 94–104, ISBN: 978-3-031-40497-9. DOI: 10.1007/978-3-031-40498-6_9.

[191] C. Larcher, M. Piau, P. Finardi, P. Gengo, P. Esposito V. Caridá, *Cabrita: closing the gap for foreign languages*, Ağus. 2023. DOI: 10.48550/arXiv.2308.11878. erişim adresi: http://arxiv.org/abs/2308.11878.

[192] G. S. Kohli ve diğ., *Building a Llama2-finetuned LLM for Odia Language Utilizing Domain Knowledge Instruction Set*, Ara. 2023. DOI: 10.48550/arXiv.2312.12624. erişim adresi: http://arxiv.org/abs/2312.12624.

[193] G. L. Garcia ve diğ., *Introducing Bode: A Fine-Tuned Large Language Model for Portuguese Prompt-Based Task*, Ocak 2024. DOI: 10.48550/arXiv.2401.02909. erişim adresi: http://arxiv.org/abs/2401.02909.

[194] A. Santilli E. Rodolà, *Camoscio: an Italian Instruction-tuned LLaMA*, en, Ara. 2023. DOI: 10.48550/arXiv.2307.16456. erişim adresi: http://arxiv.org/abs/2307.16456.

[195] T. S. Almeida, H. Abonizio, R. Nogueira R. Pires, *Sabiá-2: A New Generation of Portuguese Large Language Models*, Mar. 2024. DOI: 10.48550/arXiv.2403.09887. erişim adresi: http://arxiv.org/abs/2403.09887.

[196] N. K. Corrêa, S. Falk, S. Fatimah, A. Sen N. d. Oliveira, "TeenyTinyLlama: open-source tiny language models trained in Brazilian Portuguese," *Machine Learning with Applications*, c. 16, s. 100 558, Haz. 2024, ISSN: 26668270. DOI: 10.1016/j.mlwa.2024.100558. erişim adresi: http://arxiv.org/abs/2401.16640.

[197] Z. Zhang ve diğ., *Balancing Specialized and General Skills in LLMs: The Impact of Modern Tuning and Data Strategy*, Ekim 2023. DOI: 10.48550/arXiv.2310.04945. erişim adresi: http://arxiv.org/abs/2310.04945.

[198] C. Ling ve diğ., *Domain Specialization as the Key to Make Large Language Models Disruptive: A Comprehensive Survey*, Mar. 2024. DOI: 10.48550/arXiv.2305.18703. erişim adresi: http://arxiv.org/abs/2305.18703.

[199] J. Zheng ve diğ., *Fine-tuning Large Language Models for Domain-specific Machine Translation*, Ara. 2024. DOI: 10.48550/arXiv.2402.15061. erişim adresi: http://arxiv.org/abs/2402.15061.

[200] P. R. Inserte, M. Nakhlé, R. Qader, G. Caillaut J. Liu, *Large Language Model Adaptation for Financial Sentiment Analysis*, Ocak 2024. DOI: 10.48550/arXiv.2401.14777. erişim adresi: http://arxiv.org/abs/2401.14777.

[201] Y. Matias, *Our progress on generative AI in health*, en-us, Mar. 2024. erişim adresi: https://blog.google/technology/health/google-generative-ai-healthcare/.

[202] M. Artetxe, S. Ruder D. Yogatama, "On the Cross-lingual Transferability of Monolingual Representations," *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, ss. 4623–4637. DOI: 10.18653/v1/2020.acl-main.421. erişim adresi: http://arxiv.org/abs/1910.11856.

[203] L. Qin ve diğ., *Multilingual Large Language Model: A Survey of Resources, Taxonomy and Frontiers*, Nis. 2024. DOI: 10.48550/arXiv.2404.04925. erişim adresi: http://arxiv.org/abs/2404.04925.

[204] Y. Cui, Z. Yang X. Yao, *Efficient and Effective Text Encoding for Chinese LLaMA and Alpaca*, Şub. 2024. DOI: 10.48550/arXiv.2304.08177. erişim adresi: http://arxiv.org/abs/2304.08177.

[205] Z. Csaki, P. Pawakapan, U. Thakker Q. Xu, *Efficiently Adapting Pretrained Language Models To New Languages*, Ara. 2023. DOI: 10.48550/arXiv.2311.05741. erişim adresi: http://arxiv.org/abs/2311.05741.

[206] K. Dobler G. de Melo, "FOCUS: Effective Embedding Initialization for Monolingual Specialization of Multilingual Models," *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino K. Bali, ed., Singapore: Association for Computational Linguistics, Ara. 2023, ss. 13 440–13 454. DOI: 10.18653/v1/2023.emnlp-main.829. erişim adresi: https://aclanthology.org/2023.emnlp-main.829/.

[207] A. Ebrahimi K. Kann, "How to Adapt Your Pretrained Multilingual Model to 1600 Languages," *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, C. Zong, F. Xia, W. Li R. Navigli, ed., Online: Association for Computational Linguistics, Ağus. 2021, ss. 4555–4567. DOI: 10.18653/v1/2021.acl-long.351. erişim adresi: https://aclanthology.org/2021.acl-long.351/.

[208] J. Wei ve diğ., *Finetuned Language Models Are Zero-Shot Learners*, Şub. 2022. DOI: 10.48550/arXiv.2109.01652. erişim adresi: http://arxiv.org/abs/2109.01652.

[209] L. Ouyang ve diğ., *Training language models to follow instructions with human feedback*, Mar. 2022. DOI: 10.48550/arXiv.2203.02155. erişim adresi: http://arxiv.org/abs/2203.02155.

[210] C. Zhou ve diğ., *LIMA: Less Is More for Alignment*, en, May. 2023. erişim adresi: http://arxiv.org/abs/2305.11206.

[211] H.-A. Kuulmets, T. Purason, A. Luhtaru M. Fishel, *Teaching Llama a New Language Through Cross-Lingual Knowledge Transfer*, Nis. 2024. DOI: 10.48550/arXiv.2404.04042. erişim adresi: http://arxiv.org/abs/2404.04042.

[212] O. Holmström E. Doostmohammadi, "Making Instruction Finetuning Accessible to Non-English Languages: A Case Study on Swedish Models," *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, T. Alumäe M. Fishel, ed., Tórshavn, Faroe Islands: University of Tartu Library, May. 2023, ss. 634–642. erişim adresi: https://aclanthology.org/2023.nodalida-1.62/.

[213] I. A. Azime ve diğ., *Walia-LLM: Enhancing Amharic-LLaMA by Integrating Task-Specific and Generative Datasets*, Nis. 2024. DOI: 10.48550/arXiv.2402.08015. erişim adresi: http://arxiv.org/abs/2402.08015.

[214] L. Owen, V. Tripathi, A. Kumar B. Ahmed, *Komodo: A Linguistic Expedition into Indonesia's Regional Languages*, Mar. 2024. DOI: 10.48550/arXiv.2403.09362. erişim adresi: http://arxiv.org/abs/2403.09362.

[215] T. Blevins ve diğ., *Breaking the Curse of Multilinguality with Cross-lingual Expert Language Models*, Ekim 2024. DOI: 10.48550/arXiv.2401.10440. erişim adresi: http://arxiv.org/abs/2401.10440.

[216] P. H. Martins ve diğ., *EuroLLM: Multilingual Language Models for Europe*, Eyl. 2024. DOI: 10.48550/arXiv.2409.16235. erişim adresi: http://arxiv.org/abs/2409.16235.

[217] A. Üstün ve diğ., *Aya Model: An Instruction Finetuned Open-Access Multilingual Language Model*, Şub. 2024. DOI: 10.48550/arXiv.2402.07827. erişim adresi: http://arxiv.org/abs/2402.07827.

[218] G. Wang, S. Cheng, X. Zhan, X. Li, S. Song Y. Liu, "OPENCHAT: ADVANCING OPEN-SOURCE LANGUAGE MODELS WITH MIXED-QUALITY DATA," en, 2024.

[219] S. Mukherjee, A. Mitra, G. Jawahar, S. Agarwal, H. Palangi A. Awadallah, *Orca: Progressive Learning from Complex Explanation Traces of GPT-4*, en, Haz. 2023. erişim adresi: http://arxiv.org/abs/2306.02707.

[220] J. A. Husain ve diğ., *RomanSetu: Efficiently unlocking multilingual capabilities of Large Language Models via Romanization*, Haz. 2024. DOI: 10.48550/arXiv.2401.14280. erişim adresi: http://arxiv.org/abs/2401.14280.

[221] E. Almazrouei ve diğ., *The Falcon Series of Open Language Models*, Kas. 2023. DOI: 10.48550/arXiv.2311.16867. erişim adresi: http://arxiv.org/abs/2311.16867.

[222] 0. AI ve diğ., *Yi: Open Foundation Models by 01.AI*, Ocak 2025. DOI: 10.48550/arXiv.2403.04652. erişim adresi: http://arxiv.org/abs/2403.04652.

[223] G. Team ve diğ., *Gemini: A Family of Highly Capable Multimodal Models*, Haz. 2024. DOI: 10.48550/arXiv.2312.11805. erişim adresi: http://arxiv.org/abs/2312.11805.

[224] Y. K. Jang, D. Huynh, A. Shah, W.-K. Chen S.-N. Lim, *Spherical Linear Interpolation and Text-Anchoring for Zero-shot Composed Image Retrieval*, en, May. 2024. erişim adresi: http://arxiv.org/abs/2405.00571.

[225] S. Hong ve diğ., *MetaGPT: Meta Programming for A Multi-Agent Collaborative Framework*, Kas. 2023. DOI: 10.48550/arXiv.2308.00352. erişim adresi: http://arxiv.org/abs/2308.00352.

[226] S. Sabour, N. Frosst G. E. Hinton, "Dynamic Routing Between Capsules," *Advances in Neural Information Processing Systems*, c. 30, Curran Associates, Inc., 2017, ss. 3856–3866.

[227] X. Zhao, Z. Kang, A. Feng, S. Levine D. Song, *Learning to Reason without External Rewards*, en, May. 2025. DOI: 10.48550/arXiv.2505.19590. erişim adresi: http://arxiv.org/abs/2505.19590.

[228] A. Jolicoeur-Martineau, *Less is More: Recursive Reasoning with Tiny Networks*, Ekim 2025. DOI: 10.48550/arXiv.2510.04871. erişim adresi: http://arxiv.org/abs/2510.04871.

[229] X. Wang ve diğ., *Executable Code Actions Elicit Better LLM Agents*, en, May. 2024. erişim adresi: http://arxiv.org/abs/2402.01030.

[230] *Introducing Contextual Retrieval*, en. erişim adresi: https://www.anthropic.com/news/contextual-retrieval.

[231] *Introducing the next generation of Claude*, en. erişim adresi: https://www.anthropic.com/news/claude-3-family.

[232] *Introducing Apple's On-Device and Server Foundation Models*, en-US. erişim adresi: https://machinelearning.apple.com/research/introducing-apple-foundation-models.

[233] H. Leong, Y. Gao, S. Ji, Y. Zhang U. Pamuksuz, "Efficient Fine-Tuning of Large Language Models for Automated Medical Documentation," *2024 4th International Conference on Digital Society and Intelligent Systems (DSInS)*, Kas. 2024, ss. 204–209. DOI: 10.1109/DSInS64146.2024.10992195. erişim adresi: https://ieeexplore.ieee.org/document/10992195/.

[234] H. Y. Leong, Y. Gao S. Ji, "A GEN AI Framework for Medical Note Generation," *2024 6th International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, Kas. 2024, ss. 423–429. DOI: 10.1109/ICAICA63239.2024.10823004. erişim adresi: https://ieeexplore.ieee.org/document/10823004/.

[235] *White House Unveils America's AI Action Plan*, en-US, Tem. 2025. erişim adresi: https://www.whitehouse.gov/articles/2025/07/white-house-unveils-americas-ai-action-plan/.

[236] "America's AI Action Plan," en,

[237] M. A. Bayram, A. A. Fincan, A. S. Gümüş, B. Diri, S. Yıldırım Ö. Aytaş, *Setting Standards in Turkish NLP: TR-MMLU for Large Language Model Evaluation*, Ocak 2025. DOI: 10.48550/arXiv.2501.00593. erişim adresi: http://arxiv.org/abs/2501.00593.

[238] M. A. Bayram, *Turkish MMLU: Yapay Zeka ve Akademik Uygulamalar İçin En Kapsamlı ve Özgün Türkçe Veri Seti*, tur, Ağus. 2024. DOI: 10.5281/zenodo.13375018. erişim adresi: https://zenodo.org/records/13375018.

[239] M. A. Bayram, *Türkçe Tıbbi Soru-Cevap Veri Seti: 167 Bin Sağlık Sorusu ve Cevabı*, tur, Tem. 2024. DOI: 10.5281/zenodo.12770916. erişim adresi: https://zenodo.org/records/12770916.

[240] M. A. Bayram. "Healthcare-Focused Turkish Medical Dataset," erişim tarihi 8 Nis. 2026. erişim adresi: https://huggingface.co/datasets/alibayram/doktorsitesi.

[241] M. A. Bayram, B. Diri S. Yıldırım, "Healthcare-Focused Turkish Medical LLM: Training on Real Patient-Doctor Question-Answer Data for Enhanced Medical Insight," *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, c. 37, no. 4, Article 111, 2024, Manuscript/PDF in repository under papers-latex.

[242] M. A. Bayram. "Cosmos Corpus with Precomputed Embeddings," erişim tarihi 8 Nis. 2026. erişim adresi: https://huggingface.co/datasets/alibayram/cosmos-corpus-0-05-with-embeddings.

[243] M. A. Bayram. "Cosmos Corpus Encoded," erişim tarihi 8 Nis. 2026. erişim adresi: https://huggingface.co/datasets/alibayram/cosmos-corpus-encoded.

[244] Cosmos AI Research Group. "Cosmos Turkish Corpus v1.0," erişim tarihi 8 Nis. 2026. erişim adresi: https://huggingface.co/datasets/ytu-ce-cosmos/Cosmos-Turkish-Corpus-v1.0.

[245] M. A. Bayram, *embeddingmagibu-200m: A Turkish-Focused Sentence Embedding Model with an 8,192-Token Context Window*, Model kartı (çevrim içi erişim), 2026. erişim adresi: https://huggingface.co/magibu/embeddingmagibu-200m.

[246] T. Kudo J. Richardson, *SentencePiece: A Simple and Language Independent Subword Tokenizer and Detokenizer for Neural Text Processing*, 2018.

[247] M. A. Bayram. “distil-trainer,” erişim tarihi 8 Nis. 2026. erişim adresi: https://pypi.org/project/distil-trainer/.

[248] M. A. Bayram. “Wikipedia-40 Languages,” erişim tarihi 8 Nis. 2026. erişim adresi: https://huggingface.co/datasets/alibayram/wikipedia-40-langs.

[249] M. A. Bayram. “Wikipedia-40 Languages with Embeddings,” erişim tarihi 8 Nis. 2026. erişim adresi: https://huggingface.co/datasets/alibayram/wikipedia-40-langs-with-embeddings.

[250] G. Beken Fikri. “STSbTR,” erişim tarihi 8 Nis. 2026. erişim adresi: https://huggingface.co/datasets/figenfikri/stsb_tr.

[251] Magibu AI Research Group. “TR-MTEB Turkish Results Explorer,” erişim tarihi 8 Nis. 2026. erişim adresi: https://huggingface.co/spaces/magibu/mteb-turkish.

[252] M. Türker, A. E. Kızıloğlu, O. Güngör S. Üsküdarlı, *TabiBERT: A Large-Scale ModernBERT Foundation Model and A Unified Benchmark for Turkish*, arXiv:2512.23065 [cs], Ocak 2026. DOI: 10.48550/arXiv.2512.23065. erişim tarihi 22 Ocak 2026. erişim adresi: http://arxiv.org/abs/2512.23065.

[253] Y. Wu ve diğ., *Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation*, Eyl. 2016. erişim adresi: http://arxiv.org/abs/1609.08144.

[254] T. Kudo J. Richardson, *SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing*, en, Ağus. 2018. erişim adresi: http://arxiv.org/abs/1808.06226.

[255] H. S. Vera ve diğ., *EmbeddingGemma: Powerful and Lightweight Text Representations*, https://arxiv.org/abs/2509.20354, 2025.

[256] M. A. Bayram. “transformer-cloner,” erişim tarihi 8 Nis. 2026. erişim adresi: https://pypi.org/project/transformer-cloner/.

[257] Google. “EmbeddingGemma-300M,” erişim tarihi 8 Nis. 2026. erişim adresi: https://huggingface.co/google/embeddinggemma-300m.

[258] N. Reimers I. Gurevych, *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*, en, Ağus. 2019. erişim adresi: http://arxiv.org/abs/1908.10084.

[259] Google. “Gemma-2-9B,” erişim tarihi 8 Nis. 2026. erişim adresi: https://huggingface.co/google/gemma-2-9b.

[260] Meta. “Llama-3.2-3B,” erişim tarihi 8 Nis. 2026. erişim adresi: https://huggingface.co/meta-llama/Llama-3.2-3B.

[261] Cohere For AI. “Aya Expanse 8B,” erişim tarihi 8 Nis. 2026. erişim adresi: https://huggingface.co/CohereForAI/aya-expanse-8b.

[262] A. Bayram, B. Diri S. Yıldırım, "Data Quality-Based Adaptive Learning Rate: A Case Study on Medical Text Classification," *Proceedings of the IV. International Congress on Artificial Intelligence in Health*, İzmir Bakırçay University, Turkey, Kas. 2024, ss. 124–125. erişim adresi: https://kongresyz.bakircay.edu.tr/Yuklenenler/kongreler/kongresyz/ProceedingBook_07May.pdf.

[263] M. A. Bayram. "Turkish TikToken/Tokenizer Demo," erişim tarihi 8 Nis. 2026. erişim adresi: https://huggingface.co/spaces/alibayram/turkish_tiktokenizer.

[264] Magibu AI Research Group. "embeddingmagibu-152m Demo," erişim tarihi 8 Nis. 2026. erişim adresi: https://huggingface.co/spaces/magibu/embeddingmagibu-152m.

[265] M. A. Bayram. "Turkish MMLU Dataset," erişim tarihi 8 Nis. 2026. erişim adresi: https://huggingface.co/datasets/alibayram/turkish_mmlu.

[266] HuntNes. "Turkish Fact-Checking Dataset," erişim tarihi 8 Nis. 2026. erişim adresi: https://huggingface.co/datasets/HuntNes/teyit-analizleri.

[267] M. A. Bayram. "Turkish Sentiment Corpus," erişim tarihi 8 Nis. 2026. erişim adresi: https://huggingface.co/datasets/alibayram/yorumbudur.

[268] ceyyyh. "Türkçe Masallar," erişim tarihi 8 Nis. 2026. erişim adresi: https://huggingface.co/datasets/ceyyyh/turkce_masallar_extended.

[269] ceyyyh. "Tarihte Bugün," erişim tarihi 8 Nis. 2026. erişim adresi: https://huggingface.co/datasets/ceyyyh/tarihtebugun.

[270] M. A. Bayram. "malibayram GitHub Repository Index," erişim tarihi 8 Nis. 2026. erişim adresi: https://github.com/malibayram.

[271] M. A. Bayram. "M. Ali Bayram Hugging Face Profile," erişim tarihi 8 Nis. 2026. erişim adresi: https://huggingface.co/alibayram.

[272] Magibu AI Research Group. "Magibu AI Research Group," erişim tarihi 8 Nis. 2026. erişim adresi: https://huggingface.co/magibu.

[273] M. A. Bayram. "Turkish Tokenizer," erişim tarihi 8 Nis. 2026. erişim adresi: https://github.com/malibayram/turkish-tokenizer.

[274] M. A. Bayram, *embeddingmagibu-152m: A Turkish-Focused Sentence Embedding Model with a 2,048-Token Context Window*, Model kartı (çevrim içi erişim), 2025. erişim adresi: https://huggingface.co/magibu/embeddingmagibu-152m.

A
EKLER

Bu ek, tezde üretilen veri setleri, modeller ve araçların erişim bilgilerini tek bir yerde toplamak için hazırlanmıştır. Amaç, yeniden üretilebilirliği kolaylaştırmak ve ilgili kaynaklara hızlı erişimi sağlamaktır.

Ek bölüm, tezde geçen veri ve model çıktı paketlerinin izlenebilirliğini artırmak için platform bağlantıları, sürüm bilgileri ve lisans notlarını birlikte sunar. Böylece araştırmacılar aynı kaynaklara doğrudan erişebilir ve deneyleri tutarlı biçimde yeniden çalıştırabilir.

A.1 Tez Kapsamında Geliştirilen Veri Setleri ve Paylaşım Bilgileri

Bu ekte, tez kapsamında oluşturulan tüm Türkçe NLP veri kümeleri; içerik kapsamı, paylaşım platformu, DOI numarası ve sürüm bilgisiyle birlikte detaylı biçimde sunulmaktadır. Veri kümeleri, Türkçe LLM araştırmalarında hem **eğitim** hem de **değerlendirme** amacıyla kullanılabilecek biçimde tasarlanmış ve açık erişimli olarak yayımlanmıştır. Her veri setinin lisans koşulları, ilgili platformun (Zenodo, Hugging Face) veri kartı üzerinden doğrulanabilir. Bu yaklaşım, Türkçe NLP alanında açık bilim, şeffaflık, tekrarlanabilirlik ve sürdürülebilirlik ilkelerini desteklemektedir.

Tez kapsamında geliştirilen başlıca veri kümeleri Tablo A.1'te özetlenmiştir.

Tablo A.1 için bağlantılar kaynakça atıflarıyla verilmiştir: TR-MMLU veri setinin Hugging Face bağlantısı [265], TR-MMLU veri setinin Zenodo bağlantısı [238], Healthcare-Focused Turkish Medical Dataset için Hugging Face bağlantısı [240], Healthcare-Focused Turkish Medical Dataset için Zenodo bağlantısı [239], Turkish Fact-Checking Dataset [266], Turkish Sentiment Corpus [267], `TÃijrkÃğeMasallar` [268], `TarihteBugÃijn` [269].

Veri setlerinin açıklamaları şu şekildedir. **TR-MMLU**, 62 farklı akademik disiplinden 6.200 çoktan seçmeli soru içeren bir ölçüm standardıdır; sorular Ölçme, Seçme ve Yerleştirme Merkezi (ÖSYM) ve İstanbul Üniversitesi Açık ve Uzaktan Eğitim Fakültesi (AUZEF) gibi resmi sınav kaynaklarından derlenmiş, çekirdek deney fazında 39 modelle raporlanmış ve açık liderlik tablosu 05 Şubat 2026

Tablo A.1 Tez veri setleri: platform, DOI ve sürüm bilgileri.

Veri Seti Adı	Platform	DOI / Bağlantı	Sürüm
TR-MMLU	Hugging Face Zenodo	`10.5281/zenodo.13375018`	v1.2
Healthcare-Focused Turkish Medical Dataset	Hugging Face Zenodo	`10.5281/zenodo.12770916`	v1.0
Turkish Fact-Checking Dataset	Hugging Face	N/A	v1.0
Turkish Sentiment Corpus	Hugging Face	N/A	v1.1
Türkçe Masallar	Hugging Face	N/A	v1.0
Tarihte Bugün	Hugging Face	N/A	v1.0

itibarıyla 85 model değerlendirmesine ulaşmıştır. Bu kapsam, TR-MMLU'yu Türkiye'de kamuya açık Türkçe MMLU karşılaştırıcıları içinde en fazla model değerlendirmesi barındıran ölçüm zemini haline getirmiştir. **Healthcare-Focused Turkish Medical Dataset**, 167.732 hasta–doktor etkileşiminden oluşan büyük ölçekli bir tıbbi soru-cevap derlemidir; içerik Doktorsitesi.com platformundan anonimleştirilmiş biçimde toplanmış ve DoctorLlama/DoctorGemma modellerinin eğitiminde kullanılmıştır. **Turkish Fact-Checking Dataset**, teyit analizlerinden türetilen doğrulama metinlerini içererek dezenformasyon tespiti çalışmalarına kaynak sağlar. **Turkish Sentiment Corpus**, ürün ve hizmet yorumlarından oluşan, pozitif–negatif–nötr etiketli geniş bir duygu analizi derlemidir. **Türkçe Masallar**, Türk halk masalları ve kültürel anlatılardan oluşan 50.000 cümlelik bir korpustur; dilin kültürel kalıplarını modellemeye yöneliktir. **Tarihte Bugün** ise 12.000 kronolojik kaydı içeren tarihsel olay anlatılarını kapsar ve zaman odaklı metin analizi çalışmalarında kullanılmaktadır.

A.2 Açık Bilim İlkeleri ve Tekrarlanabilirlik

Bu tez, yalnızca bilimsel katkı sağlamakla kalmayıp, **Türkçe NLP için yeniden üretilebilir ve açık erişimli bir ekosistem** oluşturmayı hedeflemiştir. Tüm veri kümeleri, modeller ve kodlar Findable, Accessible, Interoperable, Reusable (FAIR) veri paylaşım ilkeleri doğrultusunda tasarlanmış ve arşivlenmiştir.

Uygulanan açık bilim politikaları, veri ve model çıktı paketlerinin uzun vadeli izlenebilirliğini güvence altına alacak şekilde kurgulanmıştır. Tezin temel veri setleri Zenodo üzerinde DOI ile kalıcı olarak arşivlenmiş ve erişim standardı desteklenmiştir (ör. TR-MMLU: `10.5281/zenodo.13375018` [238], tıbbi soru-cevap: `10.5281/zenodo.12770916` [239]). Veri temizleme ve ön işleme süreçleri açık kaynak Python kodlarıyla (`pandas`, `nltk`, `regex`, `langdetect`) belgelenmiş; modellerin değerlendirme sonuçları TR-MMLU liderlik tablosu üzerinden kamuya açık biçimde yayımlanmıştır. Zenodo arşivleri, **DataCite** standardına uygun meta verilerle düzenlenerek sürüm ve açıklama bilgilerinin izlenebilirliği sağlanmıştır.

Tekrarlanabilirlik garantileri, araştırma sürecinin tüm kritik bileşenlerinin açık biçimde belgelenmesine dayanır. Deneysel ortam konfigürasyonları (GPU tipi, bellek, sürümler) ayrıntılı olarak raporlanmış; eğitim parametreleri (`learning rate`, `batch size`, `LoRA rank`) açık biçimde paylaşılmıştır. Seçili veri setleri ve artefakt sürümleri Zenodo'da arşivlenmiş, kod ve model paylaşımları GitHub/Hugging Face üzerinden sürümlendirilerek birbirleriyle ilişkilendirilmiştir.

Bu kapsamda oluşturulan altyapı, Türkçe doğal dil işleme alanında açık veri paylaşımı kültürünün yaygınlaştırılmasına katkı sağlamaktadır.

A.3 Ek Kod, Model ve Kaynak Erişim Bilgileri

Tez kapsamında geliştirilen tüm modeller, veri kümeleri ve araçlar kamuya açık depolarda paylaşılmıştır. **GitHub (Kod Deposu)** [270] adresinde model eğitim kodları, tokenizasyon algoritması (`turkish_tokenizer.py`), veri ön işleme betikleri ve deneysel not defterleri yer almaktadır. **Hugging Face (Modeller ve Veri Kümeleri)** [271] üzerinden TR-MMLU karşılaştırıcısı, sağlık veri kümeleri, tokenizasyon araçları ve değerlendirme tablolarına erişim sağlanmaktadır. **Magibu AI Research Group** sayfası [272] ise araştırma grubunun yayımladığı modelleri ve topluluk katkılarını bir araya getirir. **Zenodo** arşivleri (TR-MMLU, DOI: 10.5281/zenodo.13375018; Türkçe tıbbi soru-cevap, DOI: 10.5281/zenodo.12770916) [238], [239] kayıtlarıyla kalıcı arşivleme sağlar.

embeddingmagibu-200m için yeniden üretilebilirlik notları, modelin tokenizer uyarlaması ve çevrimdışı damıtma hattının açıkça belgelenmesine dayanır. Bu tezde geliştirilen Türkçe cümle gömme modeli **embeddingmagibu-200m**, tokenizer uyarlaması + çevrimdışı damıtma yaklaşımıyla üretilmiştir [245]. Model kartı ve ek çıktılar üzerinden tekrar üretim için asgari adımlar aşağıdaki gibidir (özet):

A.3.1 Kurulum

```
    pipinstall-Usentence-transformers
    datasetssentencepiece
    pipinstall-Utransformer-cloner
    distil-trainer
```

A.3.2 Tokenizer Eğitimi (SentencePiece BPE)

```
import sentencepiece as spm
spm.SentencePieceTrainer.train(
    input='cosmos_corpus.txt',
    model_prefix='turkish_bpe',
    vocab_size=131072,
    model_type='bpe'
)
```

A.3.3 Öğretmen Modelin Yeni Tokenizer ile Klonlanması (Embedding Eşleştirme)

```
from transformer_cloner import TransformerCloner
cloner = TransformerCloner(
    source_model='google/embeddinggemma-300m',
    target_tokenizer='./turkish_bpe.model'
)
cloner.clone(output_path='./cloned_model')
```

A.3.4 Öğretmen Embedding'lerinin Önceden Hesaplanması (Çevrimdışı Veri Seti)

```
    fromdistil_trainer.dataimport
    TeacherEmbeddingsGenerator
    generator=TeacherEmbeddingsGenerator(teacher_model=
    'google/embeddinggemma-300m')
    generator.generate(dataset='cosmos_subset',
    output_path='./embeddings_dataset')
```

A.3.5 Damıtma Eğitimi (Kosinüs Kaybı)

```
from distil_trainer import EmbeddingDistillationTrainer
trainer = EmbeddingDistillationTrainer(
    student_model='./cloned_model',
    embeddings_dataset='./embeddings_dataset',
    target_type='final',
    loss='cosine',
    batch_size=256,
    learning_rate=5e-5,
    num_epochs=1,
    precision='bf16'
)
trainer.train()
```

A.3.6 Yayımlanan Ek Çıktılar

Model ağırlıkları/tokenizer dosyaları, STSbTR checkpoint değerlendirmeleri ve TR-MTEB çıktı CSV'leri, aynı zamanda demo ve sonuç gezgini Space'leri model kartında referanslanmıştır [245], [264], [251].

A.4 Embedding Sonuçları: Tam Döküm

Bu bölümde TR-MTEB değerlendirmesinin tam sıralama ve görev matrisi dökümü sunulmaktadır. STSbTR'nin 21 modeli içeren tam tablosu, ana metindeki embedding deneyleri bölümüne taşınmıştır (Bölüm 6.7, Tablo 6.9). TR-MTEB'nin güncel çevrim içi sonuç gezgini için bkz. [251].

A.5 TR-MTEB Tam Sıralama (26 model)

Tablo A.2 TR-MTEB genel sıralama (26 model, 26 görev ortalaması).

Sıra	Model	Ortalama
1	text-embedding-3-small	66,5
2	ytu-ce-cosmos/turkish-e5-large	66,0
3	intfloat/multilingual-e5-large-instruct	65,9
4	google/embeddinggemma-300m	65,2
5	alibaba-NLP/gte-multilingual-base	64,4
6	intfloat/multilingual-e5-base	64,2
7	**magibu/embeddingmagibu-200m**	**63,9**
8	selmanbaysan/ turkishembeddingmodelfinetuned	63,4
9	intfloat/multilingual-e5-small	62,4
10	newmindai/Mursit-Large-TR-Retrieval	62,0
11	sentence-transformers/ paraphrase-multilingual-mpnet-base-v2	61,7
12	**magibu/embeddingmagibu-152m**	**60,2**
13	newmindai/Mursit-Base-TR-Retrieval	59,9
14	trmteb/turkish-embedding-model	59,8
15	turkishgeneraltextembeddingmodel	59,7
16	emrecan/ bert-base-turkish-cased-mean-nli-stsb-tr	54,8
17	dbmdz/bert-base-turkish-uncased	49,5
18	fasttext	46,9
19	glove	41,3
20	boun-tabilab/TabiBERT-Code	41,1
21	boun-tabilab/TabiBERT	40,8
22	KocLab-Bilkent/BERTurk-Legal	39,3
23	sentence-transformers/all-mpnet-base-v2	37,6
24	Trendyol/ TY-ecomm-embed-multilingual-base-v1.2.0	36,6
25	newmindai/TurkEmbed4Retrieval	36,6
26	newmindai/TurkEmbed4STS	36,6

A.6 TR-MTEB Tam Görev Matrisi (26x26)

Aşağıdaki iki tablo, 26 görevin tamamı için model skorlarını içerir. Tablo A.4 Retrieval ve Classification, Tablo A.5 ise kalan görev kategorilerini kapsar. Okumayı kolaylaştırmak için sütun kısaltmaları Tablo A.3 içinde toplu verilmiştir.

Tablo A.3 MTEB sütun kısaltmaları (kategori, kısa kod ve açık görev adı).

Kategori	**Kısaltma**	**Açık Ad**
Retrieval	CQA	`CQADupstackGamingRetrievalTR`
Retrieval	MSM	`MSMarcoTRRetrieval`
Retrieval	QUO	`QuoraRetrievalTR`
Retrieval	SQU	`SquadTRRetrieval`
Retrieval	TQD	`TQuadRetrieval`
Retrieval	XQD	`XQuADRetrieval`
Classification	THY	`THYSentimentClassification`
Classification	TSN	`TSTimelineNews CategoryClassification`
Classification	T75	`Turkish75NewsClassification`
Classification	IRN	`TurkishIronyClassification`
Classification	MOV	`TurkishMovieSentiment Classification`
Classification	NEW	`TurkishNewsCategoryClassification`
Classification	OFF	`TurkishOffensiveLanguage Classification`
Classification	PRD	`TurkishProductSentiment Classification`
Clustering	TAC	`TurkishAbstractCorpusClustering`
Clustering	TCW	`TurkishColumnWritingClustering`
STS	STS	`STSbTR`
NLI	MNL	`MnliTr`
NLI	SNL	`SnliTr`
NLI	XNL	`XNLI`
BitextMining	WMT	`WMT16BitextMining`
Reranking	ARG	`ArguAnaTR`
Reranking	FIQ	`FiQA2018TR`
Reranking	NFC	`NFCorpusTR`
Reranking	SCI	`SCIDOCSTR`
Reranking	SCF	`SciFactTR`

Tablo A.4 TR-MTEB tam görev matrisi - Bölüm I-A: Retrieval (26 model).

Model	CQA	MSM	QUO	SQU	TQD	XQD
text-embedding-3-small	80,7	53,4	98,8	65,6	79,5	90,3
ytu-ce-cosmos/ turkish-e5-large	50,8	58,7	95,9	76,4	84,2	95,8
intfloat/ multiling.-e5-large-inst.	51,2	56,8	95,9	73,0	84,9	94,5
google/ embeddinggemma-300m	54,7	54,8	94,3	73,3	85,0	93,1
alibaba-NLP/ gte-multiling.-base	56,3	56,8	95,1	68,6	81,9	92,3
intfloat/ multilingual-e5-base	56,0	57,1	95,2	76,9	86,8	95,1
embeddingmagibu-200m	**51,2**	**57,4**	**94,6**	**62,4**	**79,5**	**88,1**
selmanbaysan/ turkish-emb-fine-tuned	51,7	57,2	94,1	58,1	80,4	87,0
intfloat/ multilingual-e5-small	53,3	56,5	94,2	73,9	88,7	94,3
newmindai/ Mursit-Large-TR-Retr.	53,6	59,0	95,0	62,1	81,5	90,1
sent-transf./ paraphrase-multil.	46,2	51,5	94,7	58,0	65,7	86,0
embeddingmagibu-152m	**47,8**	**51,7**	**92,5**	**66,2**	**79,4**	**90,2**
newmindai/ Mursit-Base-TR-Retr.	50,5	58,6	94,3	58,3	80,9	87,6
trmteb/ turkish-embedding-model	54,1	54,3	94,6	54,1	74,7	85,4
turkish general text emb. model	54,0	54,4	94,6	54,1	74,8	85,4
emrecan/ bert-base-tr-nli-stsb	38,5	43,5	93,0	47,9	73,0	79,1
dbmdz/ bert-base-turkish-uncased	23,4	34,6	86,2	44,5	60,6	72,4
fasttext	23,7	31,6	75,7	44,4	65,6	71,8
glove	16,4	20,5	56,7	27,0	42,5	50,6
boun-tabilab/ TabiBERT-Code	10,4	16,4	81,0	16,7	28,6	33,1
boun-tabilab/TabiBERT	10,6	16,8	81,5	15,8	28,5	34,1
KocLab-Bilkent/ BERTurk-Legal	8,3	9,7	56,3	16,8	20,2	36,3
sent-transf./ all-mpnet-base-v2	22,6	25,5	74,7	12,8	27,5	36,5
Trendyol/ TY-ecomm-embed-v1.2.0	18,1	19,8	68,6	24,7	59,4	45,0
newmindai/ TurkEmbed4Retrieval	18,1	19,8	68,6	24,6	59,0	45,0
newmindai/ TurkEmbed4STS	18,1	19,8	68,6	24,6	59,0	45,0

Tablo A.4 TR-MTEB tam görev matrisi - Bölüm I-B: Classification (devamı).

Model	THY	TSN	T75	IRN	MOV	NEW	OFF	PRD
text-embedding-3-small	59,1	64,1	98,7	52,6	71,9	92,2	60,8	56,8
ytu-ce-cosmos/ turkish-e5-large	64,8	66,4	93,3	58,9	82,6	93,3	61,0	60,8
intfloat/ multiling.-e5-large-inst.	66,2	64,9	92,7	58,8	85,2	93,8	60,7	67,2
google/ embeddinggemma- 300m	59,2	63,1	96,0	55,2	78,2	90,9	64,7	67,3
alibaba-NLP/ gte-multiling.-base	59,4	64,3	84,0	52,9	70,8	90,2	58,7	62,1
intfloat/ multilingual-e5-base	62,6	62,6	84,0	54,6	72,6	90,8	57,8	60,6
embeddingmagibu- 200m	**59,6**	**58,7**	**91,3**	**52,9**	**71,9**	**88,8**	**63,9**	**61,0**
selmanbaysan/ turkish-emb-fine-tuned	62,3	60,7	95,3	53,3	78,2	91,9	64,6	61,1
intfloat/ multilingual-e5-small	55,9	60,3	82,0	51,8	67,8	90,4	59,1	58,0
newmindai/ Mursit-Large-TR-Retr.	59,3	61,6	97,3	51,8	74,0	89,6	59,7	58,0
sent-transf./ paraphrase-multil.	59,6	59,4	94,0	54,4	75,9	86,5	60,5	62,3
embeddingmagibu- 152m	**50,7**	**59,5**	**92,7**	**49,3**	**63,4**	**91,4**	**58,8**	**58,1**
newmindai/ Mursit-Base-TR-Retr.	59,9	60,3	90,0	51,1	68,2	90,2	57,9	56,7
trmteb/ turkish-embedding model	63,8	66,6	89,3	56,8	78,9	86,0	65,2	58,9
turkish general text emb. model	63,8	66,6	89,3	56,8	78,9	85,9	65,2	58,9
emrecan/ bert-base-tr-nli-stsb	63,7	55,5	77,3	53,1	77,9	80,4	60,5	63,2
dbmdz/ bert-base-turkish uncased	66,0	62,9	88,7	55,7	71,3	90,1	64,4	62,0
fasttext	56,1	46,2	86,7	52,6	64,9	86,3	58,2	55,8
glove	54,6	51,1	78,7	52,5	63,6	87,3	57,4	55,1
boun-tabilab/ TabiBERT-Code	56,6	56,6	89,3	58,0	59,8	84,7	62,5	55,9
boun-tabilab/TabiBERT	56,1	56,7	82,7	58,9	59,4	84,0	63,1	56,3
KocLab-Bilkent/ BERTurk-Legal	54,9	47,2	97,3	53,1	62,7	85,8	60,0	57,7
sent-transf./ all-mpnet-base-v2	47,8	29,8	78,7	51,3	54,7	68,4	52,7	52,0
Trendyol/ TY-ecomm-embed v1.2.0	52,7	18,0	47,3	52,7	54,8	48,0	52,0	52,0
newmindai/ TurkEmbed4Retrieval	52,7	18,0	47,3	52,7	54,8	48,2	52,0	52,0
newmindai/ TurkEmbed4STS	52,7	18,0	47,3	52,7	54,8	48,2	52,0	52,0

Tablo A.5 TR-MTEB tam görev matrisi - Bölüm II-A: Clustering, STS ve NLI (26 model).

Model	TAC	TCW	STS	MNL	SNL	XNL
text-embedding-3-small	60,6	63,5	70,8	56,5	50,5	64,6
ytu-ce-cosmos/ turkish-e5-large	61,5	60,0	80,0	62,5	54,8	70,9
intfloat/ multiling.-e5-large-inst.	62,2	62,0	81,2	63,9	52,4	72,7
google/ embeddinggemma-300m	61,9	62,8	72,9	59,4	54,7	67,6
alibaba-NLP/ gte-multiling.-base	59,0	61,8	80,7	64,2	60,3	75,7
intfloat/ multilingual-e5-base	58,5	62,0	78,4	61,2	54,0	68,8
embeddingmagibu-200m	**59,0**	**63,7**	**77,5**	**66,8**	**60,9**	**76,0**
selmanbaysan/ turkish-emb-fine-tuned	59,7	63,3	78,5	69,3	63,2	80,0
intfloat/ multilingual-e5-small	56,5	61,1	76,2	60,4	55,1	68,1
newmindai/ Mursit-Large-TR-Retr.	58,9	60,3	74,6	54,1	53,0	60,0
sent-transf./ paraphrase-multil.	58,3	57,6	82,2	71,0	60,7	82,8
embeddingmagibu-152m	**59,8**	**63,5**	**71,8**	**53,6**	**53,2**	**60,1**
newmindai/ Mursit-Base-TR-Retr.	58,5	60,6	74,0	54,0	52,7	61,4
trmteb/ turkish-embedding-model	63,8	63,3	75,1	55,8	51,3	63,4
turkish general text emb. model	63,6	63,3	74,9	55,8	51,3	63,4
emrecan/ bert-base-tr-nli-stsb	55,2	58,4	83,1	64,5	59,6	74,5
dbmdz/ bert-base-turkish-uncased	54,4	66,9	56,8	52,6	43,0	61,6
fasttext	50,9	68,6	50,1	51,2	46,0	59,2
glove	51,8	65,0	44,1	45,9	45,7	54,9
boun-tabilab/ TabiBERT-Code	49,5	72,0	51,0	49,5	42,0	58,3
boun-tabilab/TabiBERT	50,9	69,8	50,7	49,4	41,8	57,9
KocLab-Bilkent/ BERTurk-Legal	52,5	65,8	42,3	48,9	40,8	58,7
sent-transf./ all-mpnet-base-v2	39,8	53,5	50,2	48,2	43,7	56,2
Trendyol/ TY-ecomm-embed-v1.2.0	21,5	60,7	51,7	47,7	40,7	55,5
newmindai/ TurkEmbed4Retrieval	19,1	62,9	51,7	47,7	40,7	55,5
newmindai/ TurkEmbed4STS	19,1	62,9	51,7	47,7	40,7	55,5

Tablo A.5 TR-MTEB tam görev matrisi - Bölüm II-B: BitextMining, Reranking ve ortalama (devamı).

Model	WMT	ARG	FIQ	NFC	SCI	SCF	AVG
text-embedding-3-small	91,6	40,6	73,1	22,9	16,3	93,6	66,5
ytu-ce-cosmos/ turkish-e5-large	99,2	50,1	39,3	7,7	2,0	84,2	66,0
intfloat/ multiling.-e5-large-inst.	99,0	44,9	36,9	8,3	2,8	80,0	65,9
google/ embeddinggemma-300m	96,8	50,3	40,5	9,1	3,3	86,3	65,2
alibaba-NLP/ gte-multiling.-base	98,0	47,7	40,6	7,8	3,9	81,6	64,4
intfloat/ multilingual-e5-base	98,5	44,8	37,9	8,3	3,1	79,8	64,2
embeddingmagibu-200m	**97,0**	**45,4**	**45,2**	**10,7**	**2,7**	**74,8**	**63,9**
selmanbaysan/ turkish-emb-fine-tuned	86,9	36,1	41,7	8,7	2,2	62,9	63,4
intfloat/ multilingual-e5-small	97,5	42,7	32,6	7,9	2,1	75,5	62,4
newmindai/ Mursit-Large-TR-Retr.	86,7	45,7	40,6	9,5	2,7	72,8	62,0
sent-transf./ paraphrase-multil.	96,1	39,3	27,9	5,5	3,2	63,8	61,7
embeddingmagibu-152m	**90,1**	**40,0**	**33,0**	**8,2**	**2,9**	**76,9**	**60,2**
newmindai/ Mursit-Base-TR-Retr.	74,6	38,8	36,6	8,4	3,1	70,7	59,9
trmteb/ turkish-embedding-model	37,8	39,0	39,1	9,4	2,9	70,3	59,8
turkish general text emb. model	37,7	38,9	39,0	9,4	2,9	70,3	59,7
emrecan/ bert-base-tr-nli-stsb	31,5	20,7	21,3	4,8	2,3	42,8	54,8
dbmdz/ bert-base-turkish-uncased	7,1	20,6	11,2	2,6	1,0	25,9	49,5
fasttext	4,5	11,6	11,6	2,9	0,6	42,0	46,9
glove	0,2	8,4	9,6	1,1	0,9	31,8	41,3
boun-tabilab/ TabiBERT-Code	8,6	11,5	3,8	0,7	0,7	11,6	41,1
boun-tabilab/TabiBERT	7,7	10,7	4,3	0,5	0,8	11,9	40,8
KocLab-Bilkent/ BERTurk-Legal	1,6	11,4	5,8	0,8	0,4	25,6	39,3
sent-transf./ all-mpnet-base-v2	5,1	2,3	7,1	2,6	1,4	31,7	37,6
Trendyol/ TY-ecomm-embed-v1.2.0	5,1	5,3	6,6	1,8	0,8	41,5	36,6
newmindai/ TurkEmbed4Retrieval	5,1	4,6	6,8	1,6	0,8	41,9	36,6
newmindai/ TurkEmbed4STS	5,1	4,6	6,8	1,6	0,8	41,9	36,6

A.7 TabiBench Sonuçları (İkincil / Supplementary)

TabiBench tabloları ana performans iddiası için kullanılmamış; yalnızca karşılaştırmalı bağlamı korumak için burada sunulmuştur.

Tablo A.6 TabiBERT ve embeddingmagibu-200m karşılaştırması (ikincil).

Görev	TabiBERT	embeddingmagibu-200m
BilTweetNews-Sentiment-Analysis	35,87	**42,07**
Gender-Hate-Speech-TR	43,05	**39,91**
News-Cat	84,60	**88,76**
Pubmed-RCT-10K-TR	46,45	**44,46**
Sci-Cite-TR	47,49	**47,54**
Thesis-Abstract-Classification-11K	-	**36,67**
Turkish-Product-Reviews	62,63	**73,86**
Med-NLI-TR	36,56	**36,16**
MultiNLI-TR	36,33	**36,71**
SNLI-TR	37,03	**38,52**
Apps-Retrieval-TR	-	**6,07**
Code-Search-Net-21K-TR	-	**73,67**
Cos-QA-TR	2,16	**51,80**
Fiqa-TR	4,30	**38,06**
MsMarco-TR	8,17	**79,15**
NFCorpus-TR	0,07	**1,16**
Quora-TR	58,55	**73,52**
Scifact-TR	11,12	**68,04**
Stackoverflow-QA-TR	-	**75,52**
STSb-TR	0,00	**0,00**

Tablo A.7 TabiBench tür ortalamaları (ikincil).

Tür	TabiBERT	embeddingmagibu-200m
Sınıflandırma	53,35	**53,32**
Sınıflandırma (NLI)	36,64	**37,13**
Bilgi Erişim	14,06	**51,89**
STS	0,00	**0,00**

A.8 Bölüm Özeti

Bu ekte, tez sürecinde geliştirilen tüm Türkçe veri setlerinin, modellerin ve açık kaynak kodların erişim bilgileri sistematik biçimde sunulmuştur. Bu veri kümeleri, Türkçe doğal dil işleme alanında **sürdürülebilir, şeffaf ve tekrarlanabilir bir altyapı** oluşturmuştur. Tez kapsamında geliştirilen TR-MMLU sistemi, morfolojik farkındalıklı tokenizasyon aracı (MFT; tersinir tokenize-detokenize hatlı Türkçe metin çözümleyici) ve tıbbi veri setleri; Türkiye'de düşük kaynaklı diller için açık bilim modelinin ilk örnekleri arasında yer almaktadır. Bu yaklaşım, yalnızca Türkçe'nin değil, diğer biçimbilimsel olarak zengin dillerin de gelecekteki yapay zeka araştırmalarında etik, açıklanabilir ve erişilebilir biçimde temsil edilmesine katkı sağlamaktadır.

Ek olarak bu bölüm, veri ve model paylaşımının yalnızca teknik bir gereklilik değil, araştırma sonuçlarının doğrulanabilirliği için temel bir altyapı olduğunu vurgular. Böylece tezin açık bilim perspektifi, somut erişim noktaları üzerinden yeniden üretilebilirlik hedefiyle ilişkilendirilmiştir.

TEZDEN ÜRETİLMİŞ YAYINLAR

Makale

M. A. Bayram, A. A. Fincan, A. S. Gümüş, S. Karakaş, B. Diri ve S. Yıldırım, "Tokenization Standards for Linguistic Integrity: Turkish as an Evaluation Standard," *ACM Transactions on Asian and Low-Resource Language Information Processing*, 2025. (Hakem incelemesi yapılıyor)

M. A. Bayram, B. Diri ve S. Yıldırım, "Healthcare-Focused Turkish Medical LLM: Training on Real Patient-Doctor Question-Answer Data for Enhanced Medical Insight," *ACM Transactions on Asian and Low-Resource Language Information Processing*, 2025, doi: 10.1145/3772000. (Yayımlandı)

M. A. Bayram, A. A. Fincan, A. S. Gümüş, B. Diri, S. Yıldırım ve Ö. Aytaş, "Setting Standards in Turkish NLP: TR-MMLU for Large Language Model Evaluation," *International Journal of Pattern Recognition and Artificial Intelligence*, 2025. (Hakem incelemesi yapılıyor)

M. A. Bayram, B. Diri ve S. Yıldırım, "Adapting Pretrained Embedding Models to Turkish via Token Remapping and Distillation," *Transactions of the Association for Computational Linguistics*, 2026. (Editör incelemesi yapılıyor)

M. A. Bayram, A. A. Fincan, A. S. Gümüş, S. Karakaş, B. Diri, S. Yıldırım ve D. Çelik, "Tokens with Meaning: A Hybrid Tokenization Approach for Turkish," *Language Resources and Evaluation*, 2026. (Editör incelemesi yapılıyor)

Konferans Bildirisi

M. A. Bayram, A. A. Fincan, A. S. Gümüş, S. Karakaş, B. Diri ve S. Yıldırım, "Tokenization Standards and Evaluation in Natural Language Processing: A Comparative Analysis of Large Language Models on Turkish," *2025 33rd Signal Processing and Communications Applications Conference (SIU)*, IEEE, 2025, doi: 10.1109/SIU66497.2025.11112220 [3].

M. A. Bayram, A. A. Fincan, A. S. Gümüş, B. Diri, S. Yıldırım ve Ö. Aytaş, "TR-MMLU Evaluation Standard for Large Language Models: Performance Evaluation, Challenges, and Opportunities for Improvement," *2025 33rd Signal Processing and Communications Applications Conference (SIU)*, IEEE, 2025, doi: 10.1109/SIU66497.2025.11112154.

A. Bayram, B. Diri ve S. Yıldırım, "Data Quality-Based Adaptive Learning Rate: A Case Study on Medical Text Classification," *Proceedings of the IV. International Congress on Artificial Intelligence in Health*, İzmir Bakırçay University, Turkey, pp. 124–125, Nov. 15–17, 2024 [262].

M. A. Bayram, "Towards Linguistically Grounded Sentence Embeddings for Morphologically Rich Languages: A Turkish Case Study," submitted to *UniDive 4th General Meeting (UniDive 2026)*, Bucharest, Romania, 2026. (Özet / extended abstract gönderildi)

M. A. Bayram, "Morphology-Aware Tokenization Improves Turkish Sentence Embeddings under Controlled Training Conditions," submitted to *The Seventh International Conference Computational Linguistics in Bulgaria (CLIB 2026)*, Sofia, Bulgaria, 2026. (Kısa makale gönderimi hazırlandı)

Proje

Tez kapsamında yayımlanan açık kaynak projeler ve çevrim içi kaynaklar; **MFT / Turkish Tokenizer** [273], **Turkish TikToken/Tokenizer demo** [263], **TR-MMLU leaderboard** [4], **embeddingmagibu-152m** [274], **embeddingmagibu-152m demo** [264], **MTEB Turkish Results Explorer** [251], **transformer-cloner** [256], **distil-trainer** [247] ve **Magibu AI Research Group** [272] başlıkları altında toplanmaktadır.