Büyük Dil Modellerinin Türkçe Dil Becerileri: Kapsamlı Bir Performans Analizi

Furkan AKSOY
Mühendislik ve Doğa Bilimleri Fakültesi
Maltepe Üniversitesi
İstanbul, Türkiye
furkanaksoy178@gmail.com

***Özet*—Bu çalışma, beş son teknoloji büyük dil modelinin (BDM) Türkçe dil görevlerindeki kapsamlı bir değerlendirmesini sunmaktadır. GPT-5 Thinking, Claude 4 Sonnet, Gemini 2.5 Pro, Grok 3 ve DeepSeek R1 modellerini altı kategori üzerinde değerlendiriyoruz: yazım doğruluğu, dil bilgisi, kültürel anlayış, mantıksal muhakeme, standart sınav performansı ve metin anlama. Güncel Türk Dil Kurumu (TDK) standartlarına dayalı 500 özenle tasarlanmış test sorusu kullanarak, her modelin Türkçenin morfolojik açıdan zengin ve aglütinatif yapısını işleme yeterliliğini değerlendiriyoruz. Sonuçlarımız GPT-5 Thinking (%91,2 genel doğruluk) ve Gemini 2.5 Pro'nun (%91 genel doğruluk) diğer modellerden önemli ölçüde üstün performans sergilediğini göstermektedir. Tüm modeller temel dil bilgisi kurallarında güçlü performans gösterirken (%85-95 doğruluk), kültürel bağlam anlayışında önemli farklılıklar sergilemektedir (%70-100 aralığı). Çalışma, ünlü uyumu ve kapsamlı ek sistemi dahil olmak üzere Türkçenin yapısal karmaşıklığının mevcut BDM'ler için benzersiz zorluklar oluşturduğunu ortaya koymaktadır. Farklı uygulamalarda model seçimi için pratik öneriler sunuyor ve Türkçe dil işlemede iyileştirme gerektiren temel alanları belirliyoruz.**

***Index Terms*—Büyük Dil Modelleri, Türkçe Dil İşleme, Doğal Dil Anlama, Model Değerlendirme, Morfolojik Analiz, Çok Dilli Yapay Zeka**

I. Giriş

Büyük Dil Modelleri (BDM), doğal dil işleme alanında devrim yaratarak çeşitli dilbilimsel görevlerde olağanüstü performans sergilemiştir [1]. Ancak bu modellerin Türkçe gibi morfolojik açıdan zengin dillerdeki değerlendirmesi sınırlı kalmıştır. Türkçe, aglütinatif yapısı, karmaşık ünlü uyumu sistemi ve zengin morfolojik özellikleri nedeniyle benzersiz zorluklar sunmaktadır [2].

Bu çalışmanın motivasyonu, eğitim, medya, hukuk ve teknoloji sektörlerinde Türkçe konuşan uygulamalarda BDM'lerin artan benimsenmesinden kaynaklanmaktadır. Bu artan kullanıma rağmen, modern BDM'lerin Türkçe dil yetenekleri için özel olarak tasarlanmış kapsamlı değerlendirme çerçeveleri eksikliği mevcuttur.

Bu makale aşağıdaki katkıları sunmaktadır:

- BDM'lerin Türkçe dil yetenekleri için kapsamlı bir değerlendirme çerçevesi
- Altı farklı kategoride beş son teknoloji modelin performans analizi
- Türkçe morfolojik karmaşıklığının oluşturduğu özel zorlukların belirlenmesi
- Gerçek dünya uygulamalarında model seçimi için pratik öneriler
- Tekrarlanabilir araştırma için açık kaynak veri kümesi ve metodoloji

II. İlgili Çalışmalar

Türkçenin hesaplamalı işlenmesi üzerine erken çalışmalar morfolojik analiz ve belirsizlik giderme üzerine odaklanmıştır [2], [3]. Sinir dili modellerindeki son gelişmeler morfolojik açıdan zengin diller için umut verici sonuçlar göstermiştir [4].

Çok dilli dil modellerinin önceki değerlendirmeleri öncelikle yüksek kaynaklı dillere odaklanmıştır [5], [6]. Kapsamlı Türkçe dil değerlendirmesi konusunda sınırlı çalışma mevcuttur; mevcut çalışmaların çoğu makine çevirisi [7] veya isimli varlık tanıma [8] gibi belirli görevlere odaklanmıştır.

Çalışmamız, temel ortografik kurallardan kültürel bağlam anlamasına kadar Türkçe dil anlayışının çoklu yönlerini kapsayan bütünsel bir değerlendirme çerçevesi sunması açısından önceki çalışmalardan farklılaşmaktadır.

III. Metodoloji

A. Model Seçimi

Farklı mimari yaklaşımları ve eğitim metodolojilerini temsil eden beş çağdaş BDM'yi değerlendiriyoruz:

GPT-5 Thinking (OpenAI): Görünür düşünme süreçleri ile karmaşık problem çözme görevleri için optimize edilmiş gelişmiş muhakeme yetenekli model.

Claude 4 Sonnet (Anthropic): Hız-doğruluk dengesini ve etik değerleri vurgulayan Anayasal Yapay Zeka ile eğitilmiş model.

Gemini 2.5 Pro (Google): Kapsamlı çok dilli eğitim ile özellikle morfolojik açıdan karmaşık dillerde güçlü olan çok modlu model.

Grok 3 (xAI): Dinamik bilgi erişim yetenekleri ile gerçek zamanlı internet erişimi olan model.

DeepSeek R1: Detaylı düşünme süreçleri ve etimolojik analiz yetenekleri olan açık kaynak muhakeme modeli.

B. Test Tasarımı

Değerlendirme çerçevemiz, Türkçe dil yeterliliğinin farklı yönlerini değerlendirmek için tasarlanmış altı kategoriden oluşmaktadır:

1) Test 1: Yazım Doğruluğu: Güncel TDK kılavuzlarına dayalı 200 kelime (100 doğru, 100 yaygın hatalarla). Modeller ikili sınıflandırma ile yanıt verir ve yanlış yazılmış kelimeler için düzeltme sağlar.

2) Test 2: Temel Dil Bilgisi: Yedi alt kategori boyunca 100 soru:

- De/Da bağlacı kullanımı (15 soru)
- Ki bağlacı tanımlama (15 soru)
- Soru eki uyumu (15 soru)
- Ünlü uyumu kuralları (15 soru)
- Ünsüz benzeşmesi (15 soru)
- Noktalama ve biçimlendirme (15 soru)
- Morfolojik analiz (10 soru)

3) Test 3: Kültürel Anlayış: Türkçe atasözleri ve deyimleri kapsayan, hem anlam açıklaması hem de bağlamsal kullanım örnekleri gerektiren 50 soru.

4) Test 4: Mantıksal Muhakeme: Türkçe bağlamda analitik düşünmeyi test eden sözel mantık (20) ve sayısal mantık (20) arasında bölünmüş 40 soru.

5) Test 5: Standart Sınav Performansı: Paragraf analizi ve kapsamlı dil bilgisi değerlendirmesi dahil YKS TYT Türkçe deneme sınavlarından 40 soru.

6) Test 6: Metin Anlama: Kod değiştirme, arkaik Türkçe ifadeler ve karışık dilbilimsel öğeler içeren 10 karmaşık metin.

C. Değerlendirme Kriterleri

Tüm değerlendirmeler 2025 itibarıyla güncel Türk Dil Kurumu (TDK) standartlarını takip etmektedir. Puanlama, objektif sorular için tam eşleşme ve subjektif yanıtlar için uzman doğrulaması kullanır. Doğruluğun yanı sıra verimliliği değerlendirmek için yanıt süreleri kaydedilir.

IV. Sonuçlar

A. Genel Performans

Tablo I, tüm test kategorilerindeki kapsamlı performansı sunmaktadır. GPT-5 Thinking ve Gemini 2.5 Pro, %91'in üzerinde genel doğrulukla açık liderler olarak ortaya çıkmaktadır.

Tablo I
Genel Performans Sonuçları (% Doğruluk)

Model	T1	T2	T3	T4	T5	T6	Ort
GPT-5 Thinking	94,0	93,0	91,0	88,0	88,0	92,0	**91,2**
Gemini 2.5 Pro	91,5	92,0	98,0	80,0	85,0	88,0	**91,0**
DeepSeek R1	86,0	88,0	80,0	75,0	80,0	84,0	82,2
Claude 4 Sonnet	83,5	86,0	77,0	69,0	78,0	82,0	79,3
Grok 3	80,0	83,0	73,0	63,0	95,0	78,0	78,7

B. Detaylı Kategori Analizi

1) Yazım Doğruluğu (Test 1): Tüm modeller temel yazım kurallarında yeterlilik gösterdi, GPT-5 Thinking en yüksek doğruluğa (%94) ulaştı. DeepSeek R1, düzeltmelerin yanı sıra detaylı kelime kökeni açıklamaları sağlayan benzersiz etimolojik analiz yetenekleri gösterdi.

2) Dil Bilgisi Değerlendirmesi (Test 2): Tüm modellerde performans tutarlı olarak yüksekti (%83-93), bu da eğitim verilerinde Türkçe morfolojik kuralların yeterli temsilini göstermektedir. GPT-5 Thinking ve DeepSeek R1 arasında noktalama kullanımındaki dikkat çekici benzerlikler, potansiyel ortak eğitim kaynaklarını düşündürmektedir.

Tablo II
Dil Bilgisi Alt Kategori Sonuçları (% Doğruluk)

Alt Kategori	GPT-5	Gemini	DeepSeek	Claude	Grok
De/Da Kullanımı	96	94	90	88	84
Ki Bağlacı	94	92	88	86	82
Soru Eki	98	96	92	90	88
Ünlü Uyumu	96	94	90	88	86
Ünsüz Kuralları	94	92	88	86	84
Noktalama	80	88	82	78	74
Morfoloji	92	90	86	84	82

3) Kültürel Anlayış (Test 3): Bu kategoride önemli performans farklılığı ortaya çıktı (%73-98), eğitim verisi kültürel temsilinin etkisini vurguladı. Gemini 2.5 Pro atasözlerinde mükemmel performans sergilerken, Grok 3 en düşük kültürel anlayışı gösterdi.

4) Mantıksal Muhakeme (Test 4): Tüm modeller Türkçe bağlamda sayısal mantık ile mücadele etti, performans sözel görevlerden sayısal görevlere önemli ölçüde düştü. Bu, matematiksel muhakemenin çapraz dilsel transferinde zorlukları göstermektedir.

5) Standart Sınav Performansı (Test 5): Grok 3'ün istisnai performansı (%95), modelin problemleri analitik olarak çözmek yerine çevrimiçi cevap anahtarlarına eriştiği için doğal dil anlamasından ziyade internet arama yeteneklerine atfedildi.

6) Metin Anlama (Test 6): Modeller kod değiştirme senaryolarında tutarlı performans gösterdi (%78-92), güçlü çok dilli anlama yeteneklerini gösterdi. GPT-5 Thinking karmaşık metin yorumlama görevlerinde öne çıktı.

C. Yanıt Süresi Analizi

Tablo III, Claude 4 Sonnet'in hıza öncelik verirken DeepSeek R1'in kapsamlı analizi vurgulamasıyla işleme sürelerinde önemli farklılık göstermektedir.

Tablo III
Modellere Göre Ortalama Yanıt Süreleri

Model	Ort Süre (dk)	Aralık (dk)
Claude 4 Sonnet	2,5	1-4
Gemini 2.5 Pro	6,0	3-9
GPT-5 Thinking	10,5	8-13
Grok 3	8,0	5-12
DeepSeek R1	15,0	10-18

V. Tartışma

A. Türkçe Dil Karmaşıklığının Etkisi

Analizimiz, Türkçenin aglütinatif yapısının mevcut BDM'ler için özel zorluklar oluşturduğunu ortaya

koymaktadır. ”Çalıştırabileceklerimizden” gibi kelimeler yedi morfem içerir ve her biri anlamsal ve gramatikal bilgi katkısında bulunur. Modeller, bu tür karmaşık yapıları ayrıştırma ve anlamada değişen başarı gösterdi.

Özellikle büyük ve küçük uyum örüntülerinin etkileşimi olan ünlü uyumu kuralları, tüm modeller için zorluklar yarattı. Uzun kelime formlarında ünlü uyumu ve ünsüz benzeşmesinin eşzamanlı uygulanması, morfolojik işleme yeteneklerinin sınırlarını test etti.

B. Kültürel Bağlam Sınırlılıkları

Kültürel anlamataki önemli performans farklılığı (%70-100 aralığı) mevcut BDM'lerdeki kritik bir sınırlılığı vurgulamaktadır. İslami geleneğe ve Osmanlı mirasına derinlemesine kök salmış Türkçe kültürel ifadeler, eğitim verilerinde yetersiz temsil edilen bağlamsal bilgi gerektirmektedir.

Bu bulgu, eğitim içeriği üretimi, edebi çeviri ve sosyal medya içerik denetimi gibi kültürel duyarlılık gerektiren uygulamalar için önemli çıkarımlara sahiptir.

C. Araç Entegrasyon Stratejileri

Modeller bilgi erişimi için farklı yaklaşımlar sergiledi. Grok 3'ün bilinmeyen içerik için sistematik internet araması, GPT-5 Thinking'in iç bilgiye dayalı yaklaşımı ile çelişki oluşturmaktadır. Bu farklılık, optimal Türkçe dil işlemenin çoklu modeller ve dış kaynakları birleştiren hibrit yaklaşımlar gerektirebileceğini önermektedir.

D. Eğitim Verisi Kalitesi Sorunları

Özellikle Claude 4 Sonnet'in 2019 öncesi TDK kurallarına bağlılığında güncel olmayan eğitim verisi kanıtı ortaya çıktı. Bu, özellikle Türkçe gibi resmi olarak düzenlenen diller için eğitim verisinin gelişen dil standartları ile güncel tutulmasının önemini vurgulamaktadır.

VI. Pratik Çıkarımlar ve Öneriler

A. Uygulamaya Özel Model Seçimi

Bulgularımıza dayanarak özel öneriler sunuyoruz:

Profesyonel Yazım: Mantıksal tutarlılık kontrolü için GPT-5 Thinking'in ardından ilk taslak için Gemini 2.5 Pro kullanarak çok aşamalı yaklaşım.

Gazetecilik ve Medya: Gerçek zamanlı bilgi erişimi ve doğruluk kontrolü için Grok 3, dilbilimsel doğruluk için diğer modellerle desteklenmeli.

Eğitim Uygulamaları: Temel dil öğretimi için Gemini 2.5 Pro, ileri analitik görevler için GPT-5 Thinking.

Kültürel İçerik: Gemini 2.5 Pro üstün kültürel anlayış göstererek geleneksel Türkçe ifadeler içeren uygulamalar için uygun.

B. Belirli Görevler için Kaçınılması Gereken Modeller

Analizimiz belirli modellerin kaçınılması gereken senaryoları belirlemektedir:

- Grok 3: Tek başına yazım kontrolü için uygun değil (%80 doğruluk)
- Claude 4 Sonnet: Kültürel içerik için yetersiz (%70-80 kültürel doğruluk)
- DeepSeek R1: Gerçek zamanlı uygulamalar için uygunsuz (15 dakika yanıt süresi)

VII. Sınırlılıklar ve Gelecek Çalışmalar

Bu çalışmanın gelecek araştırmalarda ele alınması gereken birkaç sınırlılığı bulunmaktadır:

Örneklem Büyüklüğü: Kapsamlı olmakla birlikte, 500 soruluk veri kümesi daha iyi istatistiksel anlamlılık için 1000+ soruya genişletilebilir.

Model Kapsamı: Alan özelinde Türkçe BDM'ler dahil ek modeller değerlendirilmelidir.

Görev Çeşitliliği: Gelecek değerlendirmeler yaratıcı yazım, teknik çeviri ve alan özelinde görevleri içermelidir.

Uzunlamasına Analiz: Eğitim verisi ve mimarilerin evrimiyle model performans değişikliklerinin zaman içinde takip edilmesi.

İnsan Değerlendirmesi: Sonuç geçerliliğini güçlendirmek için otomatik puanlamanın yanı sıra uzman dil bilimci değerlendirmesi.

VIII. Sonuç

Türkçe dil görevlerinde beş son teknoloji BDM'nin bu kapsamlı değerlendirmesi, mevcut yetenekler ve sınırlılıklar konusunda önemli içgörüler ortaya koymaktadır. GPT-5 Thinking (%91,2) ve Gemini 2.5 Pro (%91,0) genel olarak en yetenekli modeller olarak ortaya çıkmakla birlikte, hiçbir tek model tüm kategorilerde üstünlük göstermemektedir.

Temel bulgular şunları içermektedir:

- Tüm modellerde temel dil bilgisinde güçlü performans (%85-95)
- Kültürel anlamataki önemli farklılık (%70-100 aralığı)
- Türkçe bağlamlar içinde sayısal muhakemede zorluklar
- Bazı modelleri etkileyen eğitim verisi kalitesi sorunlarının kanıtı
- Performansı etkileyen farklı araç entegrasyon stratejileri

Çalışma, Türkçenin morfolojik karmaşıklığı ve kültürel zenginliğinin BDM geliştirmede özel dikkat gerektirdiğini göstermektedir. Mevcut modeller umut verici olmakla birlikte, mükemmele yakın Türkçe dil işleme başarısı, gelişmiş eğitim verisi kalitesi, daha iyi kültürel temsil ve potansiyel olarak çoklu modelleri birleştiren hibrit yaklaşımlar gerektirecektir.

Açık kaynak veri kümemiz ve değerlendirme çerçevemiz, Türkçe konuşan uygulamalar için daha kültürel açıdan farkında ve dilbilimsel açıdan doğru modellerin geliştirilmesini destekleyerek Türkçe dil yapay zekasında sürekli araştırma için bir temel sağlamaktadır.

IX. Veri Erişilebilirliği

Tüm test verileri, detaylı metodolojiler ve değerlendirme betikleri açık kaynak lisansı altında şu adreste erişilebilir: https://github.com/FurkanAksoyy/tr-llm-turkce-performans-analizi

KAYNAKLAR

[1] T. Brown et al., "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877-1901, 2020.

[2] K. Oflazer, "Two-level description of Turkish morphology," *Literary and linguistic computing*, vol. 9, no. 2, pp. 137-148, 1994.

[3] D. Hakkani-Tür, K. Oflazer, and G. Tür, "Statistical morphological disambiguation for agglutinative languages," *Computers and the Humanities*, vol. 36, no. 4, pp. 381-410, 2002.

[4] K. Kann et al., "Fortification of neural morphological segmentation models for polysynthetic minimal-resource languages," *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 47-57, 2018.

[5] A. Conneau et al., "Unsupervised cross-lingual representation learning at scale," *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440-8451, 2020.

[6] J. Hu et al., "XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation," *International Conference on Machine Learning*, pp. 4411-4421, 2020.

[7] J. Tiedemann and S. Thottingal, "OPUS-MT—Building open translation services for the world," *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pp. 479-480, 2020.

[8] G. G. Şahin and M. Steedman, "Data augmentation via dependency tree morphing for low-resource languages," *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 5004-5009, 2018.

[9] J. Devlin et al., "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*, pp. 4171-4186, 2019.

[10] A. Rogers, O. Kovaleva, and A. Rumshisky, "A primer in BERTology: What we know about how BERT works," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 842-866, 2020.