T.C.

TRAKYA ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

DİL MODELLERİNİN TÜRKÇE DİLİNDE PERFORMANSLARININ İYİLEŞTİRİLMESİ İÇİN FARKLI İSTEM FORMATLARININ KULLANIMI

Eray NURTEKİN

YÜKSEK LİSANS TEZİ

VERİ BİLİMİ VE ANALİTİĞİ ANABİLİM DALI

Tez Danışmanı: Dr. Ögr. Üyesi Emir ÖZTÜRK

EDİRNE-2025

Eray Nurtekin'in hazırladığı **Dil Modellerinin Türkçe Dilinde Performanslarının İyileştirilmesi için Farklı İstem Formatlarının Kullanımı** başlıklı bu tez, tarafımızca okunmuş, kapsam ve niteliği açısından **Veri Bilimi ve Analitiği** Anabilim Dalında bir **Yüksek lisans tezi** olarak kabul edilmiştir.

Jüri Üyeleri	İmza
Dr. Öğr. Üyesi Emir ÖZTÜRK	……
Doç. Dr. Altan MESUT	……
Dr. Öğr. Üyesi Edip Serdar GÜNER	……

Tez Savunma Tarihi: 26/06/2025

Bu tezin Yüksek Lisans tezi olarak gerekli şartları sağladığını onaylarım.

İmza

Dr. Öğr. Üyesi Emir ÖZTÜRK

Tez Danışmanı ……

Trakya Üniversitesi Fen Bilimleri Enstitüsü onayı

……

Doç. Dr. Filiz UMAROĞULLARI

Fen Bilimleri Enstitüsü Müdürü

T. Ü. FEN BİLİMLERİ ENSTİTÜSÜ

VERİ BİLİMİ VE ANALİTİĞİ ANABİLİM DALI YÜKSEK LİSANS PROGRAMI DOĞRULUK BEYANI

Trakya Üniversitesi Fen Bilimleri Enstitüsü, tez yazım kurallarına uygun olarak hazırladığım bu tez çalışmasında, tüm verilerin bilimsel ve akademik kurallar çerçevesinde elde edildiğini, kullanılan verilerde tahrifat yapılmadığını, tezin akademik ve etik kurallara uygun olarak yazıldığını, kullanılan tüm literatür bilgilerinin bilimsel normlara uygun bir şekilde kaynak gösterilerek ilgili tezde yer aldığını ve bu tezin tamamı ya da herhangi bir bölümünün daha önceden Trakya Üniversitesi ya da farklı bir üniversitede tez çalışması olarak sunulmadığını beyan ederim.

26 /06/ 2025

Eray NURTEKİN

İmza

Yüksek Lisans Tezi

Dil Modellerinin Türkçe Dilinde Performanslarının İyileştirilmesi İçin Farklı İstem Formatlarının Kullanımı

T.Ü. Fen Bilimleri Enstitüsü

Veri Bilimi ve Analitiği Anabilim Dalı

ÖZET

Dil modelleri, etkin ince-ayarlama metotları ile herkes tarafından eğitilebilmekte ve özelleştirilerek farklı alanlarda kullanılabilmektedir. Dil modellerinin başarımı için genellikle kayıp veya doğruluk metrikleri kontrol kabul edilse de istem formatlarının seçimi de bu başarımı etkilemektedir. Bu çalışmada Llama2, Llama3 ve Mistral7B gibi büyük dil modellerinin yanı sıra küçük dil modeli olan Phi3 ile farklı iki veri seti üzerinde çeşitli yaygın olarak kullanılan istem formatlarının ve özelleştirilmiş Türkçe istem formatları ile birlikte denenmesi ile eğitilmiş ve başarım sonuçları karşılaştırılmıştır.

Elde edilen bulgular ve sonuçlar neticesinde, yalnızca model mimarisinin değil, kullanılan istem formatının yapısının da başarım üzerinde kritik bir rol oynadığını göstermiş ve istem formatının değişikliğinin dil modellerinde performansı etkilediği gözlenmiştir. Bunlarla birlikte özelleştirilmiş istem formatı kabul edilebilir kayıp değerlerinde en iyi metrik sonuçlarını elde etmiştir.

Yıl : 2025

Sayfa Sayısı : 60

Anahtar Kelimeler : Llama2, Llama3, Phi3, Mistral7B, İstem Formatı, Dil Modelleri

Master's Thesis

The Use of Different Prompt Formats to Improve the Performance of Language Models in the Turkish Language

Trakya University Institute of Natural and Applied Sciences

Department of Data Science and Analytics

ABSTRACT

Language models can be trained and customized by anyone using effective fine-tuning methods, making them versatile tools across various domains. While metrics like loss and accuracy are commonly used to assess the performance of language models, the choice of prompt formats also plays a crucial role. In this study, large language models such as Llama2, Llama3 and Mistral7B, as well as the smaller Phi3 model, were trained and evaluated on two different datasets using both commonly used prompt formats and customized Turkish prompt formats.

Based on the findings and results obtained, it has been demonstrated that not only the model architecture but also the structure of the prompt format plays a critical role in performance, and that changes in prompt formats can affect the performance of language models. Furthermore, the customized prompt format achieved the best metric results within acceptable loss values.

Year : 2025

Number of Pages : 60

Keywords : Llama2, Llama3, Phi3, Mistral7B, Prompt Format, Language Models

TEŞEKKÜR

Yüksek lisans eğitimimin sonuna kadar geçen bu zorlu akademik süreçte, her konuda desteğini ve tavsiyelerini benden esirgemeyen, bilgi birikim ve tecrübesiyle bana yol gösteren, istisnasız her durumda beni anlayışla karşılayan, karşılaştığım her sorunda yapıcı tavsiyleriyle çözüm üreten ve hiçbir zaman desteğini benden esirgemeyen danışman hocam Dr. Emir ÖZTÜRK'e en içten teşekkürlerimi sunmayı bir borç bilirim. Saygıdeğer hocamın akademik ve insani yaklaşımı, bu çalışmanın her aşamasında bana ilham kaynağı olmuştur.

Hayatım boyunca attığım her adımda olduğu gibi bu zorlu süreçte de maddi ve manevi tüm imkanlarını seferber eden, koşulsuz sevgileriyle bana güç veren, her zorlukta sabır ve özveriyle yanımda duran, beni daha iyi bir yaşam seviyesine ulaştırmak adına yıllardır fedakarlıkla çaba gösteren sevgili anneme ve babama da teşekkür etmeyi bir borç bilir, bu çalışmayı onlara ithaf etmekten büyük bir onur duyarım.

İÇİNDEKİLER

ÖZET iv

ABSTRACT v

TEŞEKKÜR vi

İÇİNDEKİLER vii

SİMGELER VE KISALTMALAR DİZİNİ ix

ŞEKİLLER DİZİNİ x

ÇİZELGELER DİZİNİ xii

BÖLÜM 1 1

GİRİŞ 1

BÖLÜM 2 5

DOĞAL DİL İŞLEME, DİL MODELLERİ VE DEĞERLENDİRME YAKLAŞIMLARI 5

2.1.Doğal Dil İşleme (NLP) 5

2.1.1. NLP Sınırlamaları 5

2.1.2. Ön İşleme İçin Kullanılan Teknikler 6

2.1.3. Doğal Dil İşleme Uygulamaları 8

2.2. Derin Öğrenme 9

2.2.1. Derin Öğrenme Uygulamaları ve Önemi 9

2.2.2. Derin Öğrenme Sürecinde Karşılaşılan Temel Zorluklar 10

2.2.3. Derin Öğrenme Modelleri 10

2.3. Büyük Dil Modelleri (LLM) 12

2.3.1. Llama 12
2.3.2. Mistral 13
2.4. Küçük Dil Modelleri (SLM) 13
2.4.1. Phi-3 14
2.5. Model Performans Değerlendirme Metrikleri 15
2.5.1. BLEU 15
2.5.2. ROUGE 15
2.5.3. METEOR 16
2.5.4. BERTScore 16
2.5.5. BLEURT 17
2.5.6. Perplexity 17
2.5.7. Distinct-n 17
BÖLÜM 3 19
DENEY SONUÇLARI VE PERFORMANS DEĞERLENDİRMESİ 19
3.1. Kullanılan İstem Formatları 20
3.2. Kullanılan Veri Setleri 22
3.3. Dil Modellerinin Farklı İstem Formatları İle Karşılaştırılması 23
3.3.1. Llama2 Modelinin Farklı istem formatları için performans analizi 23
3.3.2.Llama3 Modelinin Farklı İstem Formatları İçin Performans Analizi 28
3.3.3.Mistral Modelinin Farklı İstem Formatları İçin Performans Analizi 32
3.3.4. Phi3 Modelinin Farklı İstem Formatları İçin Performans Analizi 36
BÖLÜM 4 41
SONUÇ ve TARTIŞMA 41
KAYNAKLAR 44
ÖZGEÇMİŞ 48

SİMGELER VE KISALTMALAR DİZİNİ

BERT:	BERTScore
BLT:	BLEURT
D-1:	Distinct-1
D-2:	Distinct-2
MET:	METEOR
NLP:	Doğal Dil İşleme
PPL:	Perplexity
R-1:	Rouge-1
R-2:	Rouge-2
R-L:	Rouge-L

ŞEKİLLER DİZİNİ

Şekil 3.1. Llama2 Modelinin NovusResearch Veri Seti Üzerinde Türkçe İstem Formatlarıyla Karşılaştırması (I: TR1, II: TR2, III: TR3, IV: TR4)..... 23

Şekil 3.2. Llama2 Modelinin Yudum Veri Seti Üzerinde Türkçe İstem Formatlarıyla Karşılaştırması (I: TR1, II: TR2, III: TR3, IV: TR4) 24

Şekil 3.3. Llama2 Modelinin NovusResearch Veri Seti Kullanılarak Farklı İstem Formatlarıyla Karşılaştırması (I: LLaMA2, II: LLaMA3, III: Phi3, IV: TR4) 25

Şekil 3.4. Llama2 Modelinin Yudum Veri Seti Üzerinde Farklı İstem Formatlarıyla Karşılaştırması (I: Llama2, II: Llama3, III: Phi3, IV: TR4) 25

Şekil 3.5. Llama3 Modelinin NovusResearch Veri Seti Üzerinde Türkçe İstem Formatlarıyla Karşılaştırması (I: TR1, II: TR2, III: TR3, IV: TR4)..... 28

Şekil 3.6. Llama3 Modelinin Yudum Veri Seti Üzerinde Türkçe İstem Formatlarıyla Karşılaştırması (I: TR1, II: TR2, III: TR3, IV: TR4) 29

Şekil 3.7. Llama3 Modelinin NovusResearch Veri Seti Kullanılarak Farklı İstem Formatlarıyla Karşılaştırması (I: Llama2, II: Llama3, III: Phi3, IV: TR4)..... 30

Şekil 3.8. Llama3 Modelinin Yudum Veri Seti Kullanılarak Farklı İstem Formatlarıyla Karşılaştırması (I: Llama2, II: Llama3, III: Phi3, IV: TR4) 30

Şekil 3.9. Mistral Modelinin NovusResearch Veri Seti Üzerinde Türkçe İstem Formatlarıyla Karşılaştırması (I: TR1, II: TR2, III: TR3, IV: TR4)..... 32

Şekil 3.10. Mistral Modelinin Yudum Veri Seti Üzerinde Türkçe İstem Formatlarıyla Karşılaştırması (I: TR1, II: TR2, III: TR3, IV: TR4) 33

Şekil 3.11. Mistral Modelinin NovusResearch Veri Seti Kullanılarak Farklı İstem Formatlarıyla Karşılaştırması (I: Llama2, II: Llama3, III: Phi3, IV: TR4)..... 34

Şekil 3.12. Mistral Modelinin Yudum Veri Seti Kullanılarak Farklı İstem Formatlarıyla Karşılaştırması (I: LLaMA2, II: LLaMA3, III: Phi3, IV: TR4) 34

Şekil 3.13. Phi3 Modelinin NovusResearch Veri Seti Üzerinde Türkçe İstem Formatlarıyla Karşılaştırması (I: TR1, II: TR2, III: TR3, IV: TR4)..... 36

Şekil 3.14. Phi3 Modelinin Yudum Veri Seti Üzerinde Türkçe İstem Formatlarıyla Karşılaştırması (I: TR1, II: TR2, III: TR3, IV: TR4) 37

Şekil 3.15. Phi3 Modelinin NovusResearch Veri Seti Kullanılarak Farklı İstem Formatlarıyla Karşılaştırması (I: Llama2, II: Llama3, III: Phi3, IV: TR4)..... 38

Şekil 3.16. Phi3 Modelinin Yudum Veri Seti Kullanılarak Farklı İstem Formatlarıyla Karşılaştırması (I: Llama2, II: Llama3, III: Phi3, IV: TR4) 38

ÇİZELGELER DİZİNİ

Çizelge 3.1. Çalışmada Kullanılan Parametreler 20

Çizelge 3.2. Modellerin Kullanmış Olduğu İstem Formatları 21

Çizelge 3.3. NovusResearch Veri Seti Örneği 22

Çizelge 3.4. Yudum Veri Seti Örneği 22

Çizelge 3.5. Llama2 Modeli: NovusResearch Veri Setinde İstem Formatlarına Göre Performans Metrikleri 26

Çizelge 3.6. Llama2 Modeli: Yudum Veri Setinde İstem Formatlarına Göre Performans Metrikleri 27

Çizelge 3.7. Gerçek Çıktı İle Üretilen Çıktı Metnin Karşılaştırılması 27

Çizelge 3.8. Llama3 Modeli: NovusResearch Veri Setinde İstem Formatlarına Göre Performans Metrikleri 31

Çizelge 3.9. Llama3 Modeli: Yudum Veri Setinde İstem Formatlarına Göre Performans Metrikleri 31

Çizelge 3.10. Mistral Modeli: NovusResearch Veri Setinde İstem Formatlarına Göre Performans Metrikleri 35

Çizelge 3.11. Mistral Modeli: Yudum Veri Setinde İstem Formatlarına Göre Performans Metrikleri 35

Çizelge 3.12. Phi3 Modeli: NovusResearch Veri Setinde İstem Formatlarına Göre Performans Metrikleri 39

Çizelge 3.13. Phi3 Modeli: Yudum Veri Setinde İstem Formatlarına Göre Performans Metrikleri 39

BÖLÜM 1

GİRİŞ

Yapay zeka teknolojileri, son yıllarda bilgi işlem süreçlerini köklü biçimde dönüştürerek birçok sektörde yenilikçi çözümler sunmaya başlamıştır. Artan veri hacmi ve hesaplama gücü ile birlikte, bu teknolojiler yalnızca büyük veri kümelerinde üzerinde karmaşık analizler yapmakla kalmayıp, aynı zamanda insan zekasına özgü algılama, öğrenme, karar verme ve problem çözme yeteneklerini taklit edebilen akıllı sistemlerin geliştirilmesini mümkün kılmaktadır. Yapay zeka kendi içerisinde, makine öğrenmesi, uzman sistemler, evrimsel algoritmalar ve derin öğrenme gibi birçok alt dala ayrılmaktadır. Özellikle son yıllarda, derin öğrenme tekniklerinin gelişimiyle birlikte insan performansına yakın sonuçlar elde edilebilmiştir.

Yapay zeka modelleri, doğal dil girdilerinden kod üretimi (Austin vd., 2021), duygu analizi (Dang vd., 2020), durum tespit (Erdi vd., 2021), metin üretimi (Yuan vd., 2022), metin özetleme (Zhang vd., 2024), veri sıkıştırma (Goyal vd., 2021; Öztürk, 2024), soru cevap botları (Karri & Kumar, 2020), güvenlik (Öztürk vd., 2024; Tüfekci & Önal, 2024), ve tıp (Esmer vd., 2020) gibi çeşitli görevlerde kullanılmaktadır.

Derin öğrenme birden fazla işlem katmanı kullanarak verilerin karmaşık temsillerini öğrenen bir yapay zeka yöntemidir (LeCun, Bengio & Hinton, 2015). Derin öğrenme, özellikle büyük veri ve yüksek işlem gücü gerektiren görevlerde sağladığı başarılarla dikkat çekmektedir. Günümüzde derin öğrenmenin etkisi her alanda olduğu gibi doğal dil işleme (NLP) alanında da gözlemlenmektedir.

NLP, insan dilini anlamlandırma, yorumlama ve üretme yetenekleriyle çok çeşitli sektörlerde geniş bir uygulama yelpazesi sunmaktadır. Metin sınıflandırma, duygu analizi, makine çevirisi ve özetleme gibi görevlerde gösterdiği başarıyla NLP, insan ve

makine arasındaki etkileşimi daha doğal ve akıcı hale getirmektedir. Her geçen gün artan bu popülarite, farklı dilleri destekleyebilecek daha yüksek doğruluk ve verimlilik sağlayan NLP modelleri geliştirme ihtiyacını da beraberinde getirmektedir. Bu kapsamda, Türkçe için tasarlanmış modellerin az sayıda olması ve mevcut modellerin gereksinimleri tam olarak karşılayamaması sebebiyle, Türkçe için özel olarak tasarlanmış daha etkin ve performanslı modellerin oluşturulması gerekmektedir. Türkçe modellerinin geliştirilmesi veya iyileştirilmesi ve dilin kendine özgü yapısal özelliklerini de içeren çözümler sunulabilmesi neticesinde Türkçe için kullanılabilecek etkin modellerin hem anlama hem de üretim görevlerinde daha başarılı sonuçlar vermesi sağlanabilecektir.

NLP alanında, RNN, LSTM (Hochreiter & Schmidhuber, 1997) ve GRU (Chung, Gulcehre, Cho, & Bengio, 2014) gibi modeller uzun yıllar boyunca yaygın bir şekilde kullanılmış ve çeşitli amaçlar için başarılı sonuçlar elde edilmiştir. Bu modellerin ardından, dil modellerinin geliştirilmesinde 2017 yılında Google tarafından ortaya koyulan transformer mimarisi dil modelleri açısından dönüm noktası haline gelmiştir. transformer mimarisi, özellikle dikkat (attention) mekanizmasına dayalı yapısıyla ön plana çıkarak, modelin girdi içerisindeki tüm sözcükler arasındaki ilişkiyi etkili bir biçimde öğrenmesini sağlamakta ve bu ilişkileri daha isabetli tahminler üretmek amacıyla kullanmaktadır. Böylece transformer yapısıyla tekrarlayan yapılara (RNN, LSTM) ihtiyaç duymadan etkili sonuçlar elde edilmesi sağlanmıştır (Vaswani, 2017). Zaman içerisinde önerilen ve geliştirilen her yeni mimariyle birlikte, modellerin sahip olduğu parametre sayısında kayda değer bir artış gözlemlenmekte ve eğitim sonuçlarını etkilemektedir.

Dil modellerinin eğitimi sürecinde, gerekli giriş verileri ile bunlara karşılık gelen beklenen çıktıları içeren veri kümeleri kullanılmaktadır. Bir dil modelinin soru yanıtlama amacıyla eğitilmesi durumunda, sorular ve bu sorulara ait doğru yanıtları içeren bir veri kümesi kullanılmakta ve modelin eğitimi sırasında modeller kendilerine ait bir istem formatı kullanmaktadır. Bu formatlar genellikle İngilizce dilinde anahtar kelimelere sahip belirli etiketlerden oluşmakta ve bu etiketlerin içerisinde eğitim verisi bulunmaktadır. Genellikle her dil modeli kendi varsayılan etiketi ile eğitilmiş olsa da istem formatının değişmesinin etkisi daha önce incelenmemiştir.

Bu çalışma, dil modellerinin Türkçe dilindeki verilerle eğitimi sırasında kullanılan farklı istem formatlarının model performansı üzerindeki etkilerini incelemeyi hedeflemektedir. Çalışmanın temel odak noktası, büyük ve küçük dil modeli olarak adlandırılan farklı kapasitedeki modellerde, farklı istem formatlarının performans üzerindeki etkilerini analiz etmek olacaktır. Türkçe dilinde yüksek doğruluk ve etkinlik sağlayan bir model geliştirme ihtiyacı göz önünde bulundurulduğunda, bu tez, istem formatlarının model performansı üzerindeki kritik rolünü ortaya koymayı amaçlamaktadır.

Araştırmanın en temel amaçlarından birisi, farklı Türkçe eğitim veri setleri ile eğitilen dil modellerinde kullanılan istem formatlarının BLEU (Papineni vd., 2002), ROUGE (Barbella & Tortora, 2022), METEOR (Lavie & Denkowski, 2009), BERTScore (Zhang vd., 2020), BLEURT (Sellam, Das, & Parikh, 2020), Perplexity, Distinct-n gibi performans metrikleri ile ölçümünden sonra elde edilen sonuçları değerlendirmektir. Bu bağlamda, her istem formatının modelin öğrenme kapasitesini, hata oranlarını ve üretim başarısını nasıl etkilediği detaylı bir biçimde analiz edilmiştir. Bu analizler, Türkçe dil işleme alanında, daha doğru ve verimli sonuçlar elde etmenin yollarını belirlemeye yönelik önemli bir veri kaynağı sunacaktır.

Araştırmanın bir diğer önemli amacı, Türkçe diline özgü yeni bir istem formatı geliştirmek ve bu formatın standart istem formatlarına kıyasla performans avantajı sağlayıp sağlamadığını incelemek olacaktır. Türkçe'nin kendine has dil bilgisel yapısına ve özelliklerine uygun olarak tasarlanan bu özel istem formatlarının, modelin Türkçe veri ile daha etkili bir etkileşim kurarak daha etkili sonuçlar verebileceği görülmektedir. Bu özgün formatın sağladığı katkılar doğrultusunda, Türkçe dil işleme çalışmalarında daha verimli ve özgün dil modelleri oluşturmak için gereken altyapının sağlanması amaçlanmaktadır.

Çalışmanın son aşamasında Türkçe dil işleme görevlerinde en iyi performansı sağlayan istem formatının belirlenerek bu sonuçların başka diller ile birlikte uygulanabilirliğinin değerlendirilmesi amaçlanmaktadır. Farklı istem formatının kullanılması ile elde edilecek bulgular, yalnızca Türkçeye yönelik değil, aynı zamanda diğer dillerde yapılacak çalışmalar için de yol gösterici nitelikte olacaktır. Ayrıca düşük kaynaklı dillerde format duyarlılığına odaklanan bu tür çalışmalar, gelecekteki çok dilli

yapay zeka uygulamalarının yerelleştirilmesi ve kültürel uyumlandırılması adına da önemli çıkarımlar sunacaktır.

BÖLÜM 2

DOĞAL DİL İŞLEME, DİL MODELLERİ VE DEĞERLENDİRME YAKLAŞIMLARI

2.1.Doğal Dil İşleme (NLP)

Doğal Dil İşleme (NLP), bilgisayar bilimleri ve yapay zekanın (AI) bir dalı olarak, makinelerin insan dilini anlaması, analiz etmesi ve üretmesi üzerine yoğunlaşır. NLP, dilin kurallara dayalı modellemesini sağlayan hesaplamalı dilbilimi ile istatistiksel yöntemler, makine öğrenimi ve derin öğrenmeyi bir araya getirerek, dijital sistemlerin metin ve konuşmayı tanıyıp anlamlı yanıtlar üretebilmesine olanak tanır.

NLP araştırmaları, özellikle büyük dil modellerinin (LLM'ler) karmaşık sorgulara metin tabanlı yanıtlar üretebilme yeteneğiyle, üretici yapay zekanın gelişiminde temel bir rol oynamıştır. Ayrıca, görüntü oluşturma modellerinin metin tabanlı istekleri anlamasına olanak tanır. IBM'e (2024) göre günümüzde NLP, arama motorlarından müşteri hizmetlerinde kullanılan sohbet botlarına, Amazon Alexa, Apple Siri ve Microsoft Cortana gibi sesli asistanlara kadar çeşitli uygulamalarda yaygın olarak kullanılmakta olup, günlük dijital etkileşimlerin ayrılmaz bir parçası haline gelmiştir (IBM 2024).

2.1.1. NLP Sınırlamaları

Büyük dil modelleri her ne kadar günümüzde çok çeşitli görevlerde etkileyici başarılar elde etmiş olsalar da halen çözülmesi gereken önemli sınırlılıklar ve birçok hata barındırmasından dolayı kusursuz yapılar değillerdir. Bu modeller, sahip oldukları yüksek parametre sayısı ve büyük ölçekli veriyle eğitilmelerine rağmen, doğal dilin

doğasındaki belirsizlik, insan dilindeki çok anlamlılık (mecazlar, deyimler, kültürel referanslar) ve bağlamsal çeşitlilik gibi unsurlar nedeniyle bazı zorluklar doğmaktadır. Bu zorluklara göz atacak olursak;

- **Önyargılı Eğitim Verisi:** Büyük dil modelleri, önyargılı veriyle eğitildiğinde yanlış sonuçlar üretebilir. Belirli gruplara karşı ayrımcılık yaparak sonuçların adil olmayan bir şekilde dağılmasına sebep olabilir. Bu sorun çeşitli kullanıcı grupları için riski daha da arttırabilir, özellikle kamu hizmetleri, sağlık sektörü ve insan kaynakları gibi alanlarda ön yargılı eğitim verisi ile çalışmak tahminlerin güvenilirliğini olumsuz yönde etkileyebilir.
- **Yanlış Yorumlama:** Yanlış yorumlama, verinin analiz edilme sürecinde karşılaşılabilecek önemli bir sorundur. Bu durum, hatalı kararların alınmasına, modelin güvenilirliğinin azalmasına ve hedef kitlenin doğru belirlenememesi nedeniyle yanlış yönlendirmelere yol açabilir; dolayısıyla bilgi kaybının önünü açabilir.
- **Yeni Kelimeler:** Diller yaşayan bir varlık gibi, sosyal, kültürel ve teknolojik gelişmelerden sürekli olarak etkilenmektedirler. Modeller bu dinamik yapıya uyum sağlamaya çalışırken bağlamı anlama güçlüğü yaşayabilir bunun neticesinde yanlış çıkarımlarda bulunabilirler.
- **Ses Tonu ve Anlam:** Konuşmada kullanılan vurgu, abartı veya alaycılık, kelimelere farklı anlamlar katabilir. Bu durum, modellerin anlam analizi yapmasını daha zorlaştırabilir ve güvenilirliğini azaltabilir (Oracle, 2024).

2.1.2. Ön İşleme İçin Kullanılan Teknikler

Doğal dil işleme, dilin yapısını ve kurallarını anlamak ve bu bilgileri kullanarak çeşitli dil görevlerini otomatikleştirmek için bazı ön işleme tekniklerinden yararlanmaktadır. Bu alanda en yaygın kullanılan teknikler arasında belirteçlere ayırma, sözcük torbası modelleri, atlanacak sözcük çıkarma, kök alma ve kök çözümleme, sözcük türü etiketleme ve sözdizimi ayrıştırma teknikleri kullanılmaktadır.

- **Belirteçlere Ayırma (Tokenization):** Metni anlamlandırmak için metni kelimeler halinde ayırarak ham metni işlemesi daha kolay bir forma dönüştürür. Birden çok yöntemle belirteçlere ayırma yapılabilir. Eğer metin iyi bir tokenizasyon süreci ile yönetilebilirse NLP modelinin doğruluğunu ve

performansını arttırabilir, ayrıca metin verisinin analizini daha kolay bir hale getirebilmek için temel bir adımdır.

- **Sözcük Torbası Modelleri (Bag of Words - BoW):** Bu teknik, metindeki her bir kelimenin kaç kez geçtiğini sayarak metni bir vektör olarak temsil eder. Ayrıca, metin kümesindeki eşsiz kelimeler bir liste hâlinde tanımlanabilir. Uygulaması kolaydır ve büyük veri setlerinde hızlı çalışabilir. Kelime sıklıklarına dayandığı için elde edilen sonuçlar genellikle kolayca yorumlanabilir.
- **Önemsiz Sözcük Çıkarma (Stop Words Removal):** NLP süreçlerinde metindeki anlamsal olarak önemsiz kelimeleri çıkararak, modeli gereksiz bilgilerle eğitmekten kaçınarak performans üzerinde etkili olmaya yardımcı olmaktadır. Bu yöntemle NLP süreçlerinin verimliliği büyük ölçüde artmaktadır.
- **Kök Eldesi ve Kök Çözümleme (Stemming ve Lemmatization):** Kök eldesi ile kelimenin sonundaki ekler kaldırılarak sözcük kök haline getirilir. Örnek olarak, "Koşuyorum" kelimesi kök haline dönüştürülerek "Koş" kök haline dönüştürülür, Kök çözümleme ile kelimenin sözlükteki temel veya kök haline indirgenmesi sağlanır. Örnek olarak "Koşuyorum" kelimesi "Koşmak" haline dönüştürülür. Bu işlemler ile birlikte metni standartlaştırmak ve dil bilgisi analizi süreçlerini daha işlevsel ve verimli hale getirmeyi amaçlamaktadır.
- **Sözcük Türü Etiketleme (Part-of-Speech Tagging, POS Tagging) ve Sözdizimi Ayrıştırma (Syntax Parsing):** Sözcük türü etiketleme ile her kelimenin dilbilgisi türünü (isim, sıfat, fiil vb.) belirleme işlemidir. Bu işlem sayesinde kelimelerin bağlamını göz önünde bulundurur. Genellikle duygu analizi cümle anlamı çıkarımı ve kelime öbeklerinin tespitinde kullanılır. Sözdizimi ayrıştırma ile cümlenin dilbilgisel yapısını analiz edip, cümlenin her kelimesini başka bir kelimeye bağlayarak analiz eder bu yöntem daha çok makine çevirisi sesli asistanlar gibi alanlarda kullanılır.

2.1.3. Doğal Dil İşleme Uygulamaları

Doğal Dil İşleme (NLP) uygulamaları, birçok alanda verimlilik ve kullanıcı deneyimini geliştirmeye yardımcı olur. NLP'nin gelişen algoritmik altyapısı sayesinde, hem metin tabanlı verilerin işlemesi kolaylaşmakta hem de insan-makine etkileşimi daha doğal hale gelmektedir. Bu bağlamda, NLP'nin başlıca uygulama alanları şu şekilde sıralanabilir:

- **Rutin Görevleri Otomatikleştirme:** Sohbet botları ve dijital asistanlar, rutin müşteri hizmeti görevlerini üstlenerek çalışanların daha karmaşık görevlere odaklanmasına imkan tanır.
- **Arama İyileştirme:** Doğal dilin bağlamını analiz edebilen NLP sistemleri, kelimelerin bağlam içindeki anlamını belirleyerek ve eşanlamlıları eşleştirerek daha doğru arama sonuçları sağlar.
- **Arama Motoru Optimizasyonu:** NLP kullanarak aramaları analiz etmek, içeriği optimize ederek işletmenin arama motoru sonuçlarında üst sıralara çıkmasına yardımcı olur.
- **Büyük Belgeleri Analiz Etme:** Belge kümeleme ve konu modelleme, büyük belge koleksiyonlarının içerik çeşitliliğini anlamayı basitleştirir. Özellikle yasal incelemelerde kullanılır.
- **Sosyal Medya Analizi:** NLP ile müşteri yorumları ve sosyal medya içerikleri analiz edilerek duygu analizi yapılabilir; olumlu veya olumsuz geri bildirimler izlenir, müşteri memnuniyeti artırılabilir.
- **Pazar İç Görüleri:** Müşteri dilini ve davranışlarını analiz ederek, tüketici ihtiyaçlarını daha iyi anlamak ve ürün-pazarlama stratejilerini geliştirmek mümkün hale gelebilmektedir.
- **İçerik Yönetimi:** Yüksek hacimli müşteri yorumlarını analiz eden NLP, yorumların amacını ve tonunu değerlendirerek kalite ve nezaketin korunmasına yardımcı olur (Oracle 2024).

2.2. Derin Öğrenme

Makine öğrenmesi teknolojisi, günümüzde birçok yönüyle ortaya çıkmaktadır. Bunların başında; web aramalarından sosyal ağlarda içerik filtrelemeye, e-ticaret sitelerindeki öneri sistemlerinden kameralar ve akıllı telefonlar gibi tüketici ürünlerine kadar geniş bir alanda kullanılmaktadır. Bu uygulamaların gerçekleştirilmesinde, son yıllarda derin öğrenme tekniklerinin ön plana çıktığı görülmektedir (LeCun, Bengio & Hinton, 2015).

Derin öğrenme, temel olarak verileri katmanlar halinde işleyerek daha basit kavramlardan daha karmaşık temsiller oluşturur. Bu süreçte, bir bilgisayar sistemine karmaşık verileri yorumlama yeteneği kazandırmak amacıyla her katmanda bir dizi matematiksel işlem gerçekleştirilir. Bu işlemler sonucunda, sistem, verilerin özünü giderek daha ayrıntılı bir biçimde anlamlandırabilir. Derin öğrenmenin temel yapı taşı, çok katmanlı yapılar ve ileri beslemeli ağlar gibi modellerdir.

Derin öğrenme hayatımızda birçok alanda artık kendini göstermiş durumdadır. Bengio, Goodfellow ve Courville (2017), derin öğrenmenin günümüzde Google, Microsoft, Facebook, IBM, Apple, Netflix ve NVIDIA gibi birçok üst düzey teknoloji şirketi tarafından kullanıldığını belirtmiştir (s. 23).

2.2.1. Derin Öğrenme Uygulamaları ve Önemi

Mevcut elektronik dünya; iş verileri, finansal veriler, sağlık verileri, çoklu ortam verileri, nesnelerin interneti (IOT) verileri, siber güvenlik verileri, sosyal medya verileri gibi birçok çeşitli türde verilerle doludur (Ślusarczyk, 2018). Bu veriler ışığında derin öğrenmenin potansiyelini anlamak ve bunun üzerine yoğunlaşma günümüzde oldukça önemli hale gelmektedir. Derin öğrenme yaklaşımları, özellikle yüksek boyutlu verilerdeki karmaşık yapıları ortaya çıkarmak için mükemmel bir çözüm olarak, güvenlik teknolojileri de dahil olmak üzere geniş bir uygulama yelpazesinde performans açısından önemli ölçüde gelişmiştir (Sarker, 2021, s. 3). Bu nedenle, derin öğrenme teknikleri, tarihsel verilerden öğrenme yetenekleri sayesinde, günümüzün ihtiyaçlarına uygun akıllı veri odaklı sistemler oluşturmada kilit bir rol oynayabilir (Sarker, 2021, s. 3).

2.2.2. Derin Öğrenme Sürecinde Karşılaşılan Temel Zorluklar

Derin öğrenme algoritmaları, özellikle büyük hacimli ve yüksek kaliteli veri üzerinde eğitildiğinde daha iyi sonuçlar vermektedir. Burada dikkat edilmesi gereken en önemli husus olarak girdi veri kümesindeki aykırı öğelerin veya hataların derin öğrenme sürecini önemli bir şekilde etkileyeceği ve performansını olumsuz yönde etkileyeceği göz önünde bulundurulmalıdır. Bu nedenle, model eğitimi öncesinde kapsamlı bir veri ön işleme sürecinin gerçekleştirilmesi, verinin temizlenmesi ve doğruluğunun sağlanması büyük önem taşımaktadır. Tüm bunlarla birlikte derin öğrenme algoritmaları yoğun işlem gücü gerektirmektedir. Bu algoritmaların verimli ve sağlıklı çalışabilmesi için yeterli bilgi işlem kapasitesine sahip bir alt yapı ihtiyacı duymaktadır. Aksi takdirde, modelin eğitimi ve çıktılarının işlenmesi hem zaman hem de kaynak açısından verimsiz hale gelebilmektedir.

2.2.3. Derin Öğrenme Modelleri

Derin öğrenme modelleri, çeşitli veri türlerine ve öğrenme görevlerine uygun olarak farklı mimarilerde geliştirilmiştir. Bu mimariler;

Evrişimli Sinir Ağları (CNN): Görüntü işleme gibi yüksek boyutlu verilerde yaygın olarak kullanılır. CNN'ler, veri içindeki özellikleri katmanlar halinde çıkararak nesne tanıma, görüntü sınıflandırma ve segmentasyon gibi görevlerde kullanılırlar. En yaygın olarak kullanılan CNN mimarileri ZFNet, GoogLeNet, VGGNet, AlexNet, ResNet (Mosavi, Ardabili, & Várkonyi-Kóczy, 2019, s. 4) olarak adlandırılabilir.

Yinelemeli Sinir Ağları (RNN): Yinelenen sinir ağları, verilerin önceki bilgilerine bağlı olarak çalışan yapılardır, RNN'ler, birbirleriyle bağlantılı hücrelerin oluşturduğu bir ağdır. RNN, konuşma, el yazısı ve metin gibi ardışık verileri tanıyıp işleyebilmek için tasarlanmıştır (Mosavi, Ardabili, & Várkonyi-Kóczy, 2019, s. 5).

Uzun Kısa Süreli Bellek Ağları (LSTM): RNN'lerin kısa vadeli bir hafızası olmasından dolayı bazı problemlerde iyi çalışmazlar. Bu sebeple, RNN'lerin geliştirilmiş bir versiyonu olan LSTM'ler, uzun süreli bağımlılıkları öğrenebilir ve karmaşık dizilerde daha yüksek performans gösterir. Dil modeli oluşturma ve metin tahmini gibi görevlerde sıklıkla kullanılır.

Gürültü Giderici Otomatik Kodlayıcılar (DAE): Veriden gürültüyü temizlemek ve önemli özellikleri çıkarmak amacıyla kullanılır. Özellikle veri önişleme ve sıkıştırma görevlerinde tercih edilir. Ayrıca enerji tahmini, siber güvenlik, bankacılık, dolandırıcılık tespiti, görüntü sınıflandırma ve konuşmacı doğrulama alanlarında da DAE halihazırda kullanılmakta ve umut verici sonuçlar vermektedir. (Mosavi, Ardabili, & Várkonyi-Kóczy, 2019, s. 9).

Derin İnanç Ağları (DBN): Veriyi katmanlar halinde işleyerek özellik çıkarımı yapar ve karmaşık verilerde daha derin özelliklerin anlaşılmasını sağlar. Yaygın kullanım alanı olarak ise insan duygularının tespiti, zaman serisi tahmini, yenilenebilir enerji tahmini, ekonomik öngörüler ve kanser teşhisi gibi alanlarda uygulanmamaktadır. (Mosavi, Ardabili, & Várkonyi-Kóczy, 2019, s. 11).

Derin öğrenme mimarileri, farklı veri türlerine ve görev alanlarına yönelik olarak geliştirilmiş etkili çözümler sunmakta; özellikle görüntü işleme, ardışık veri analizi ve öznitelik çıkarımı gibi uygulamalarda yüksek performans sergilemektedir. Bununla birlikte, bu mimarilerin büyük bir bölümü, tekil görevler için optimize edilmiş olup, genel dil anlayışı ve üretimi gibi çok yönlü problemler karşısında yetersiz kalabilmektedir. Bu nedenle, doğal dil işleme alanında daha kapsamlı, bağlam duyarlı ve çok işlevli yapay zekâ sistemlerine ihtiyaç duyulmaktadır. Bu gereksinimi karşılamak amacıyla geliştirilen Büyük Dil Modelleri (LLM'ler), geniş veri kümeleri üzerinde eğitilerek dilin yapısal ve anlamsal özelliklerini derinlemesine öğrenebilmekte ve farklı dil görevlerinde yüksek başarı oranlarıyla kullanılabilmektedir. Dolayısıyla, LLM'ler, klasik derin öğrenme yaklaşımlarının ötesine geçerek, çok amaçlı ve esnek bir dil modelleme paradigması sunmaktadır.

2.3. Büyük Dil Modelleri (LLM)

Büyük Dil Modelleri (LLM - Large Language Models), doğal dil işleme (NLP) alanında insan dilini anlamak, üretmek ve yorumlamak için derin öğrenme yöntemlerini kullanan büyük ölçekli yapay sinir ağı modelleridir (AWS 2024). Büyük dil modelleri, büyük ölçekli, gelişmiş bir sinir ağıdır ve çok sayıda veri örneği kullanarak yüksek performanslı bilgisayar işlem gücüne dayalı olarak eğitilmektedir (Myers vd., 2024, s. 2). Son yıllarda, LLM'ler önemli bir yükseliş göstermekte ve popülaritesi gün geçtikçe artmaktadır. Çeşitli alanlarda eğitilen LLM'ler, doğal dil programlama (kod üretimi) ve kod analizi gibi görevlerde geniş bir kullanım alanı bulmuştur. Nitekim birçok alanda uygulanmasıyla birlikte kullanımı sırasında bir dizi zorlukla karşılaşılmaktadır. Bu zorluklar, temel olarak çalışılan alanın karmaşıklığı ve verilerin işlenmesinin zorluğundan kaynaklanmaktadır. Örneğin, veri kaynaklarının çeşitliliğinin yetersizliği, alana özgü bilgi birikiminin karmaşıklığı ve uygulama hedefleri ile bu hedeflere ulaşmada karşılaşılan kısıtlamaların özgün nitelikleri, bu zorlukların başlıca nedenleridir (Li vd., 2024, s. 1).

Günümüzde büyük dil modelleri çeşitli alanlarda kullanılmaktadır. Bunlar doğal dil anlama ve çeviri, metin üretme ve yazı asistanlığı, sohbet robotları ve sanal asistanlar, sağlık alanında kullanım, kod yazma ve yazılım geliştirme, siber güvenlik ve dolandırıcılık tespiti olarak örneklendirilebilirler.

Büyük Dil Modelleri, çok çeşitli görevlerde yüksek doğruluk sağlayan, bağlamı anlayabilen ve doğal dili etkili biçimde işleyebilen güçlü yapılar olarak ön plana çıkmaktadır. Bu alandaki gelişmelerle birlikte, farklı kurumlar ve araştırma toplulukları tarafından çeşitli mimariler geliştirilmiştir. Bu doğrultuda, son dönemde öne çıkan ve açık kaynaklı yapılarıyla dikkat çeken modellerden ikisi olan Llama ve Mistral, LLM ekosisteminde önemli bir yer edinmiştir.

2.3.1. Llama

Llama, 7 milyar (7B) ile 65 milyar (65B) parametre arasında değişmekte ve mevcut büyük dil modelleriyle rekabet edebilecek performans sunmaktadır (Touvron vd., 2023, s. 1). Bu modeller, çok büyük miktarda token kullanılarak eğitilmiş ve yalnızca halka açık veri setleriyle, özel verilere ihtiyaç duymadan ileri düzeyde performans

sergileyebileceğini kanıtlamaktadır. Ayrıca LLaMA-13B, boyut olarak 10 kat daha küçük olmasına rağmen, GPT-3 (175B) modelini çoğu benchmarkta geride bırakırken, LLaMA-65B, Chinchilla-70B ve PaLM-540B gibi büyük modellerle benzer seviyede performans göstermektedir (Touvron vd., 2023, s. 1). Bu çalışmada Llama model ailesinden Llama2 (Touvron vd., 2023) seçilmiş ve sonuçların alınmasında kullanılmıştır.

2.3.2. Mistral

Mistral, yüksek performanslı ve verimli bir büyük dil modeli olarak geliştirilmiştir. Özellikle metin üretimi, matematiksel işlemler ve kod oluşturma alanlarında sıkça kullanılmaktadır. AWS ve Azure gibi bulut platformlarında kolayca dağıtılabilmesi sayesinde popülaritesi artmaktadır.

Mistral, Apache 2.0 lisansı altında yayımlanmıştır. Bu sürüm, yerel olarak veya AWS, GCP ve Azure gibi bulut platformlarında kolayca dağıtılmasını sağlayan bir referans uygulama ile birlikte sunulmaktadır (Github - Mistralai 2025). Mistral, birçok farklı görev için kolayca ince ayar (fine-tuning) yapılacak şekilde tasarlanmıştır (Jiang vd., 2023, s. 2). Bu çalışmada Mistral'in 7B büyüklüğündeki modeli kullanılmıştır.

2.4. Küçük Dil Modelleri (SLM)

Küçük dil modelleri, doğal dili işleyebilen, anlayabilen, üretebilen ve daha küçük ölçekli görevleri gerçekleştirmek için optimize edilmiş yapay zeka modelleridir. Küçük dil modellerinin büyük dil modellerine göre parametre sayısı daha düşüktür. Daha az parametreye sahip oldukları için, daha hızlı eğitilebilirler. Ayrıca bu modeller daha az hesaplama kaynağı gerektirdiği için daha hızlı çalışabilirler.

Küçük dil modelleri düşük kaynak gereksinimi, büyük dil modellerine göre daha az parametre içermesi sebebiyle mobil uygulamalarda, gömülü sistemler ile IOT cihazlarında, gerçek zamanlı metin işleme gerektiren uygulamalarda kullanılabilir. Daha büyük modeller birçok işletme için tercih edilen teknoloji olmaya devam ederken, küçük modeller hızla popülerlik kazanmaktadır.

Günümüzde popülerlik kazanan küçük dil modellerinin bazıları şunlardır (IBM 2025):

- DistilBERT
- Gemma
- GPT-4o mini
- Granite
- Phi

Bu çalışmada küçük dil modellerinin en bilinenlerinden olan Phi-3 modeli kullanılmış ve sonuçlar elde edilmiştir.

2.4.1. Phi-3

Phi-3, Microsoft tarafından geliştirilen, açık yapay zeka modeli ailesidir ve küçük dil modelleri (SLM) kategorisinde en yetenekli ve maliyet açısından en verimli seçeneklerden biridir. Phi-3 modelleri, dil anlama, mantıksal çıkarım, kodlama ve matematik gibi farklı alanlarda, kendi boyutundaki ve bir üst boyuttaki modellerden üstün performans göstermektedir (Abdin vd., 2024)

Günümüzde 3,8 milyar parametreli bir dil modeli olan Phi-3-mini, Microsoft Azure AI Studio, Hugging Face ve Ollama üzerinde kullanılabilir hale gelmiştir.

Phi-3 modeli 4K ve 128K token olmak üzere iki farklı bağlam uzunluğu sunmaktadır. 128K token bağlamı sunması sebebiyle uzun metinlerin işlenmesinde avantaj sağlamaktadır. Model, Instruction-Tuning (Kullanıcı talimatlarını anlama ve bunlara uygun yanıtlar üretme) yeteneği ile eğitilmiştir. Azure AI, Ollama, ONNX Runtime, NVIDIA NIM gibi birçok platform için kullanılabilmektedir ve uzun metin işleme, sohbet botları ve sanal asistanlar, mobil ve gömülü sistemlerde kullanılmak üzere eğitilebilirler (Microsoft Azure 2025).

2.5. Model Performans Değerlendirme Metrikleri

2.5.1. BLEU

Doğal dil işleme alanında, modellerin performanslarını ölçmek için çeşitli metrikler kullanılmaktadır. Bu metriklerden biri olan BLEU (Bilingual Evaluation Understudy), otomatik bir çeviri ile aynı kaynak cümlenin insan tarafından oluşturulmuş referans çevirileri arasındaki farkın ölçüsüdür (Microsoft 2025).

BLEU metriği, model tarafından üretilen metin (aday metin) ile referans metin (insan tarafından yazılmış) arasındaki benzerliği ölçer. Metrik, bu aradaki benzerlikleri ölçerek n-gram prensibine göre değerlendirir. Aday metindeki n-gramların referans metindeki n-gramlarla eşleşme oranı precision (kesinlik) olarak adlandırılır. Bu oran modelin kadar doğru çıktı ürettiğini gösterir. Bazı durumlarda aday metin, referans metinden daha kısa olabilir, bu durum brevity penalty (kısalık cezası) olarak adlandırılır. Bu şekilde yapılarak modelin çıktısının yetersiz bilgi içermesinin önüne geçmek hedeflenmektedir (Papineni vd., 2002). BLEU skorunun puanlaması 0 ile 1 arasında değişen bir değerdir ve 1' e ne kadar yakınsa, modelin performansı o kadar yüksek kabul edilmektedir.

2.5.2. ROUGE

ROUGE (Recall-Oriented Understudy for Gisting Evaluation), özetlerle karşılaştırma yaparak, bir özetin kalitesini otomatik olarak belirlemeye çalışır. ROUGE metriği referans metin ile aday metin arasındaki n-gram benzerliği, kelime dizileri ve kelime çiftleri gibi birimlerin sayısını hesaplar (Lin, 2004, s.1).

ROUGE metriği temel prensip olarak üretilen özet ile referans özet arasındaki n-gram eşleşmeleri arasındaki eşleşmeleri sayar. ROUGE modeli referans özetlerde bulunan n-gramların ne kadarının aday özet tarafından yakalandığını Recall (Geri çağırma) ile özet bilgilerinin ne kadar iyi kapsandığını gösterir.

ROUGE metriği metin özetleme, makine çevirisi ve metin üretimi görevlerinde hızlı ve otomatik bir değerlendirme yöntemi sunmakta, doğal dil işleme görevlerinde yaygın bir şekilde kullanılmaktadır.

ROUGE metriğinin 1'e yakın sonuçlar üretmesi, model çıktısının referans metinle yüksek kelime düzeyinde örtüştüğünü göstermektedir.

2.5.3. METEOR

Meteor metriği ilk olarak 2004 yılında geliştirilmiş ve yayımlanmıştır. Makine çevirisi çıktılarının cümle düzeyinde insan yargılarıyla yüksek düzeyde korelasyon göstermesini amaçlayan bir değerlendirme metriği olarak ön plana çıkmaktadır. METEOR, bir makine çevirisi çıktısı ile aynı kaynak cümleye ait referans çeviriler arasındaki sözcüksel benzerliği ölçmeye dayalıdır. Meteor metrik değerlendirme yaparken her bir hipotez cümle ile bir veya birden fazla referans cümle arasında bir sözcük eşleştirmesi kurar. (Lavie & Denkowski, 2009)

Meteor metriği, yalnızca n-gram eşleşmelerini değil, aynı zamanda kök, eş anlamlılık ve kelime sırasını da dikkate alarak daha dilsel temelli bir değerlendirme sunmaktadır.

Metrik değerleri 0 ile 1 arasında olup, skor 1'e yaklaştıkça, model çıktısı referans metne daha benzer ve anlamlı kabul edilir.

2.5.4. BERTScore

BERTScore, iki cümle arasındaki benzerliği, bu cümlelerdeki kelimelerin bağlamsal gömme (embedding) arasındaki kosinüs benzerliklerinin toplamı üzerinden hesaplayan bir değerlendirme metriğidir (Zhang vd., 2020). BERTScore metriği, her bir token ile hem referans hem de aday metinlerde, önceden eğitilmiş bir bağlamsal yerleştirme modeli benimser. Algoritma, referans ve aday metinlerdeki tüm tokenler arasında çift yönlü kosinüs benzerliklerini hesaplayarak bir matris oluşturur. Bu benzerlik puanları, aday metnin referans metinle ne kadar iyi eşleştiği ölçmek için kullanır. Metrik değeri minimum 0 değeri almakta olup, 1' e yaklaştıkça, model çıktısı ile referans cümle arasındaki anlamsal uyumluluk artmaktadır.

2.5.5. BLEURT

Metin üretimi sistemlerinin değerlendirilmesinde kullanılan, önceden eğitilmiş ve daha sonra insan değerlendirme verileriyle ince ayar yapılmış bir regresyon modelidir.

BLEURT, BERT tabanlı öğrenilmiş bir değerlendirme metriğidir, bu yaklaşımın temel bileşenlerinden biri, modelin genelleme yeteneği arttırmak amacıyla sentetik örnek kullanılarak geliştirilen özgün bir ön eğitim şemasıdır (Sellam, Das, & Parikh, 2020).

BLEURT skorunun değer aralığı -1 ile 1 arasında değişmektedir. Modelin performans metrik sonuçlarının 1' e yakın olması ile model çıktısıyla referans cümle arasında benzer sonuç ilişkisinin arttığını göstermektedir.

2.5.6. Perplexity

Bu metrik sezgisel olarak, dilin öngörülebilirliğini ölçerek, bir konuşma tanıma tanıma sisteminin karşılacağı zorluk seviyesini belirlemek için güçlü ve temel mektriklerden biridir (Jelinek vd., 1977).

Düşük perplexity, modelin daha doğru ve öngörülebilir çıktılar ürettiğini; yüksek perplexity ise daha fazla belirsizlik taşıyan çıktılar oluşturduğunu göstermektedir (Ankner vd., 2024).

2.5.7. Distinct-n

Distinct-n metrikleri, klasik n-gram tabanlı değerlendirme metriklerinin aksine, ifade tekrarlarına değil, model çıktılarındaki özgünlük ve sözcüksel çeşitlilik düzeyini ölçmeyi hedeflemektedir. Bu metrikler, kelime tekrarlarını değil, benzersiz n-gram oranlarını ölçerek modelin dil üretimindeki sözcüksel zenginliği yansıtmaktadır (Li vd., 2016).

Distinct-1, modelin ürettiği toplam kelime sayısı içindeki benzersiz 1-gram oranına karşılık gelir. Özellikle tekrarlı ve düşük çeşitliliğe sahip metinlerin tespitinde kullanılır. Distinct-1 değerinin yüksek çıkması modelin daha zengin ve çeşitli bir dil üretimi gerçekleştirdiğini gösterir.

Distinct-2, dil modellerinin ürettiği metinlerdeki benzersiz 2-gram oranını ölçen bir çeşitlilik metriğidir. Distinct-1 tek kelime çeşitliliği ölçerken, Distinct-2, ifade

düzeyindeki iki kelimelik kombinasyon çeşitliliğini ölçer böylece Distinct-1 metriğine kıyasla daha kapsamlı bir dil çeşitliliği değerlendirmesi sunar.

Yüksek Distinct-2 değeri, modelin üretim çıktılarında daha zengin ve tekrar içermeyen yapıların üretildiğini gösterir. Distinct-n metriğinin skor aralığı 0 ile 1 arasında değişmektedir.

BÖLÜM 3

DENEY SONUÇLARI VE PERFORMANS DEĞERLENDİRMESİ

Bu bölümde, çalışmada kullanılan veri setleri, kullanılan dil modelleri ve bu modellerin farklı istem formatları kullanılarak farklı veri setleri üzerindeki performans karşılaştırmaları ele alınmaktadır.

Sonuçların alınmasında Llama2, Llama3, Mistral7B ve Phi3 modelleri varsayılan formatları ve Türkçe'ye özgü 4 farklı istem formatı (TR1, TR2, TR3, TR4) ile ince ayar (fine-tuning) yaparak ayarlanmıştır. Bu modeller, elde edilen kayıp sonuçlarına göre irdelenmiştir. Daha sonra ince ayarlanan tüm modellerin ürettiği çıktılar BLEU, ROUGE, METEOR, BERTScore, BLEURT, Perplexity, Distinct-n performans değerlendirme metrikleri kullanılarak, metin tabanlı başarı ölçütleriyle değerlendirilmiştir. Bu metrikler sayesinde modellerin yalnızca eğitim sürecindeki kayıpları değil, aynı zamanda ürettikleri çıktının kalite ve anlam bütünlüğü de ölçülerek daha kapsamlı bir başarı analizi yapılmaktadır.

Modellerin eğitiminde kullanılan parametrelerin değerleri Çizelge 3.1'de verilmiştir.

Çizelge 3.1. Çalışmada Kullanılan Parametreler

	Parametre	Değer	Açıklama
1	max_steps	10000	Modelin toplam kaç adım boyunca eğitileceğini belirtir.
2	learning_rate	2.5e-5	Ağırlık güncellemelerinde kullanılan öğrenme oranıdır. Eğitim hızını ve stabilitesini belirler.
3	gradient_checkpoint	True	GPU belleğinden tasarruf sağlamak için ara adımları tekrar hesaplamaya olanak tanır. Büyük modellerin eğitilmesine olanak tanır.
4	fp16 / bf16	fp16 = True, bf16 = False	Yarı hassiyetli (mixed precision) eğitimle bellek kullanımı ve hızı optimize edilir.
5	optim	paged_adamw_8bit	8-bit optimizasyon ile bellek tüketimini azaltır.
6	LoRA(r, alpha, dropout)	r=8, alpha=32, dropout=0.1	Modelin sadece bazı katmalarında düşük-rank güncellenmesini sağlar. Daha az parametreyle etkili fine-tuning yapılmasına olanak tanır.
7	quantization_config	4bit, nf4, bfloat16	Modeli 4-bit hassasiyetle çalıştırarak GPU belleğinin yükünü azaltır.
8	group_by_length	True	Benzer uzunluktaki örnekleri gruplayarak daha verimli eğitim sağlar.
9	eval_steps	100	Kaç adımda bir modelin test verisi üzerinde değerlendirileceğini kontrol eder.
10	per_device_train_batch_size	4	GPU başına dğşen eğitim veri kümesi örneği sayısıdır.
11	gradient_accumulation_steps	1	Daha büyük bir efektif batch size için gradyanların kaç adımda bir toplanacağını belirtir.

3.1. Kullanılan İstem Formatları

Yapılan çalışmada, farklı dil modellerinin Türkçe istem formatlarına nasıl tepki verdiğini ve bu tepkilerin formatlara göre nasıl değiştiğini inceleyebilmek amacıyla, Llama2, Llama3, Phi3, Mistral gibi kullanılan ve İngilizce dilinde geliştirilen istem formatları incelenmiş ve bu istem formatlarının özellikleri incelenerek Türkçeye özgü, Türkçe'nin yapısal özelliklerine uygun olarak tasarlanmış 4 farklı istem formatı oluşturulmuştur. Bu formatlar TR1-4 şeklinde adlandırılmıştır. Modellerin varsayılan

formatları ve Türkçe için oluşturulan formatlar Çizelge 3.2'de verilmiştir. Mistral istem formatı Llama2 ile aynı olduğu için tabloda bu formata yer verilmemiştir.

Çizelge 3.2. Modellerin Kullanmış Olduğu İstem Formatları

Ad	**İstem Formatı**
Llama2	<s>[INST]$$INPUT$$[/INST] $$OUTPUT$$</s>
Llama3	<|begin_of_text|> <|start_header_id|>user<|end_header_id|>$$INPUT$$<|eot_id|> <|start_header_id|>assistant<|end_header_id|>$$OUTPUT$$<|eot_id|>
Phi3	<|user|>$$INPUT$$<|end|> <|assistant|>$$OUTPUT$$<|end|>
TR1	<soru>\n$$INPUT$$\n</soru> <cevap>\n$$OUTPUT$$\n</cevap>
TR2	[SORU] $$INPUT$$ [CEVAP]$$OUTPUT$$
TR3	<|soru|>$$INPUT$$<|son|> <|cevap|>$$OUTPUT$$<|son|>
TR4	<|baslangic|> <|baslik|><|kullanici|>$$INPUT$$<|son|><|baslik_sonu|> <|baslik|><|asistan|>$$OUTPUT$$<|son|><|baslik_sonu|> <|bitis|>

Formatların oluşturulmasında diğer İngilizce formatların basit ve karmaşık alternatifleri seçilmiş ve yapısal karmaşıklığın etkisinin incelenmesi amaçlanmıştır. Örneğin TR1 formatında xml etiketi tarzı bir yapı seçilmiştir. TR3 formatında Phi3 benzeri başlama ve bitiş etiketlerinin bulunduğu bir yapı seçilirken, TR2 formatında LLama2'de olduğu gibi yalnızca başlangıç etiketleri bulunmaktadır. TR1 ve TR3 formatı karmaşıklık açısından aynı olmakla birlikte kullanılan karakter ve etiket içerikleri farklıdır. Bu sayede karmaşıklık dışında etiketin türünün değişimi de incelenebilecektir. Son olarak TR4 formatında LLama3'te olduğu gibi bir başlangıç tanımlanmış ve ayrıca LLama3'te olmayan bir bitiş etiketi de eklenmiştir.

3.2. Kullanılan Veri Setleri

Deneysel çalışmalar kapsamında NovusResearch ve Yudum olmak üzere iki farklı Türkçe veri seti kullanılmıştır. Türkçe eğitim için bulunabilecek işlenmiş en büyük veriseti olmaları sebebi ile bu iki veriseti seçimi gerçekleştirilmiştir.

NovusResearch veri seti, Türkçe dilinde çeşitli talimatları içeren 51600 satırlık bir koleksiyondur. Her bir satırda, bir talimat (instruction) ve buna karşılık gelen çıktı (output) bulunmaktadır. Bu veri seti, dil modellerinin Türkçe talimatları anlaması ve uygun yanıtlar üretmesi için kullanılabilmektedir.

NovusResearch veri setinden örnek bir girdi ve çıktı satırı Çizelge 3.3'te verilmiştir.

Çizelge 3.3. NovusResearch Veri Seti Örneği

Instruction	Output
Yapay zeka ile ilişkili teknoloji ile ilgili terimlerin bir listesini oluşturun.	Makine öğrenimi, doğal dil işleme, bilgisayarla görme, robotik süreç otomasyonu, derin öğrenme, kuantum hesaplama, arttırılmış gerçeklik, sanal gerçeklik, sinir ağları, doğal dil üretimi, konuşma tanıma, duygu analizi, teknoloji destekli süreç otomasyonu

Yudum veri seti, Türkçe dilinde çeşitli talimatları içeren 82.4 bin satırlık bir koleksiyondur. Her bir satırda, bir talimat (instruction), girdi (input) ve çıktı (output) bulunmaktadır. Bu yapı, dil modellerinin Türkçe talimatları anlaması ve uygun yanıtlar üretmesi için zengin bir kaynak sunmaktadır.

Yudum veri setinden bir örnek Çizelge 3.4'te verilmiştir.

Çizelge 3.4. Yudum Veri Seti Örneği

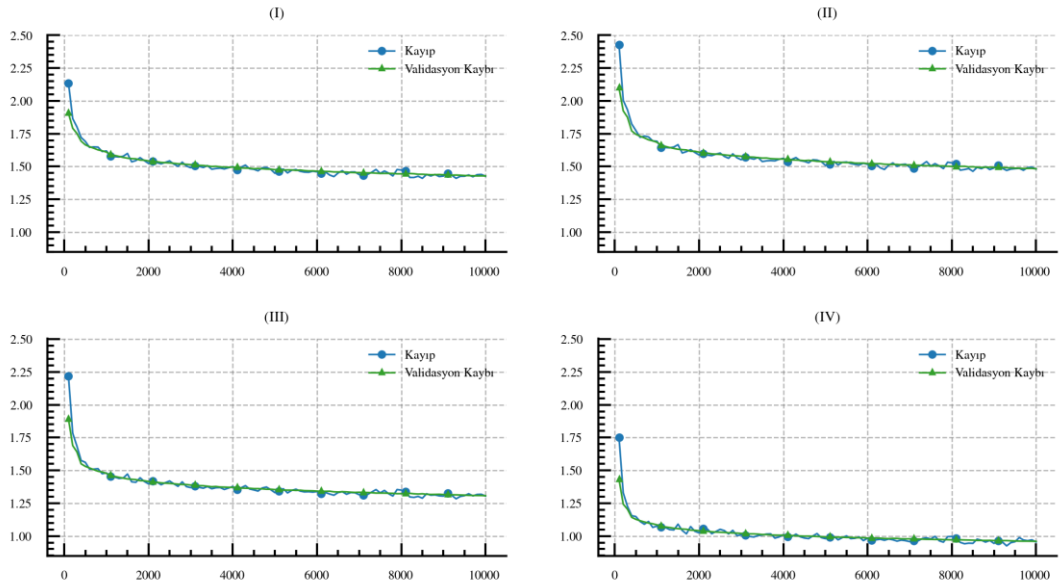
INSTRUCTION	INPUT	OUTPUT
Aşağıdaki cümleyi tekil halden çoğul hale getirin.	Kedi içiyor.	Kediler içiyor.

3.3. Dil Modellerinin Farklı İstem Formatları İle Karşılaştırılması

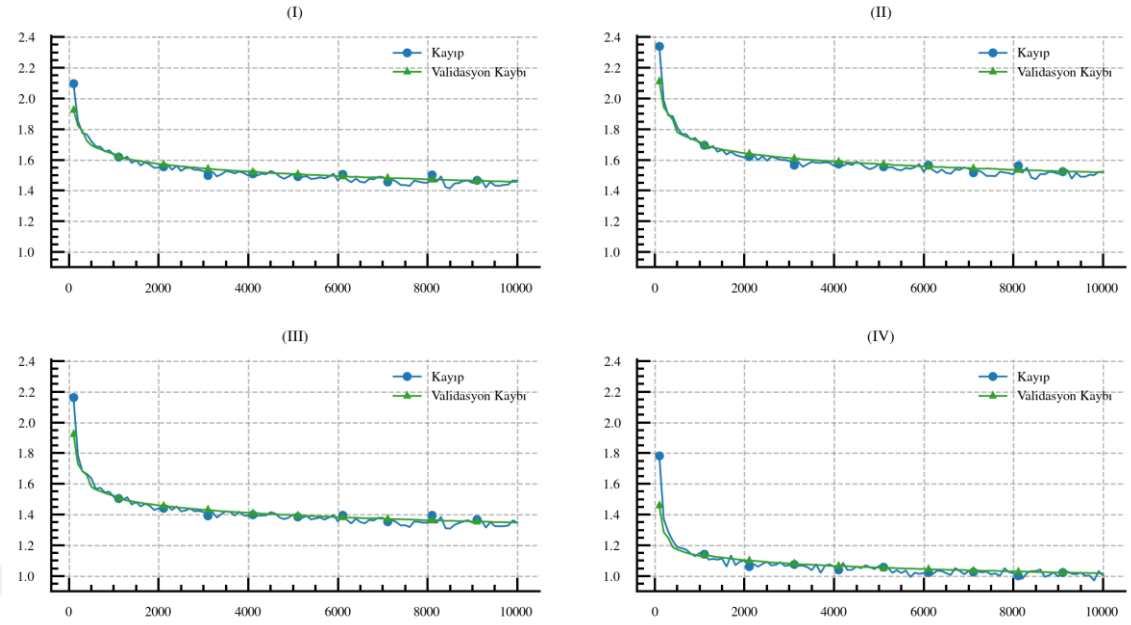
Llama2, Llama3, Phi3 ve Mistral modelleri, Çizelge 3.2'de verilen istem formatları kullanılarak NovusResearch ve Yudum veri setleri ile ince ayar (fine-tuning) yapılarak ayarlanmıştır. Bu aşamada her model varsayılan istem formatları ve Türkçe formatlar dışında birbirlerinin kullandıkları formatlar ile de eğitilmiştir. Daha sonra bu verisetleri üzerinde test verisi ile metin üretilmiş ve çalışmada verilen metrikler kullanılarak sonuçlar elde edilmiştir. Şekillerde, her bir kayıp grafiği (I)- (IV) model üzerinde farklı bir istem formatının uygulandığını temsil etmektedir. Eğitim sürecindeki kayıp (loss) değerleri, validasyon kayıpları üzerinden, modelin hem öğrenme başarımını hem de genelleme kapasitesi değerlendirilmektedir. Sonuçların sunulmasında öncelikle hangi Türkçe istem formatının en iyi sonucu elde ettiği tespit edilmiş ve daha sonra bu formatın diğer istem formatları ile karşılaştırılması ayrıca verilmiştir.

3.3.1. Llama2 Modelinin Farklı istem formatları için performans analizi

Llama2 modeli için TR istem formatlarının (TR1, TR2, TR3, TR4) farklı veri setleri üzerindeki performansları Şekil 3.1 ve Şekil 3.2'de karşılaştırılmalı olarak gösterilmektedir.



Şekil 3.1. Llama2 Modelinin NovusResearch Veri Seti Üzerinde Türkçe İstem Formatlarıyla Karşılaştırması (I: TR1, II: TR2, III: TR3, IV: TR4)



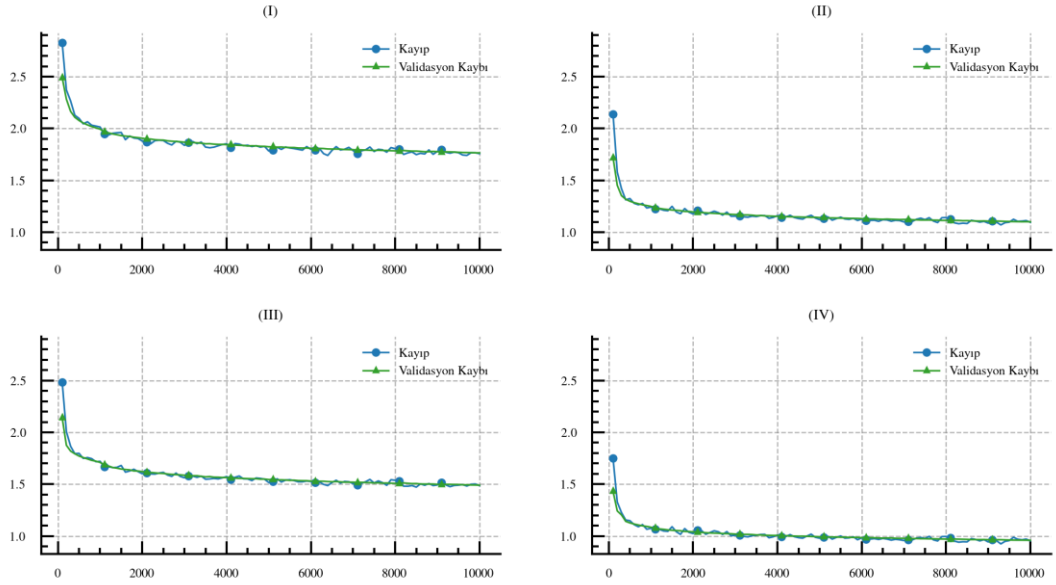
Şekil 3.2. Llama2 Modelinin Yudum Veri Seti Üzerinde Türkçe İstem Formatlarıyla Karşılaştırması (I: TR1, II: TR2, III: TR3, IV: TR4)

Grafikler incelendiğinde, IV numaralı grafiğin her iki veri setinde de diğer istem formatlarına kıyasla daha düşük bir kayıp (loss) değeriyle başladığı görülmektedir. Bu durum, modelin henüz erken aşamalarda bile daha yüksek düzeyde bir uyum sağladığını ve ilgili istem formatının, modelin öğrenme süreciyle daha uyumlu olduğu sonucunu ortaya koymaktadır. Ayrıca validasyon kaybı, eğitim kaybında olduğu gibi hızlı bir şekilde düşerek stabil hale gelmektedir. Bu da modelin, hızlı ve iyi bir şekilde genelleme yeteneğinin olduğu sonucu ortaya çıkarmakta ve aşırı öğrenme (overfitting) yaşamadığını göstermektedir.

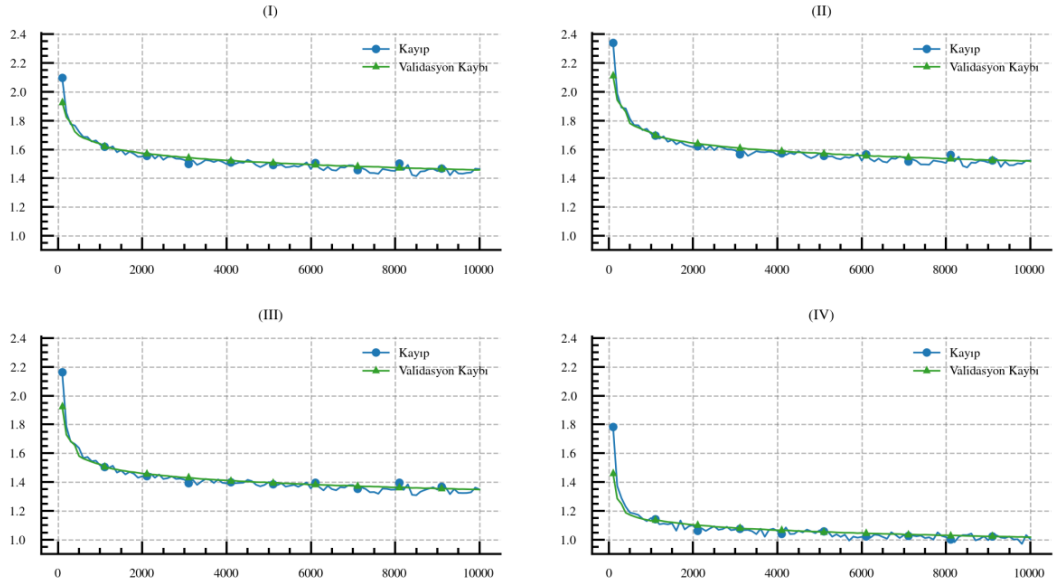
Grafikler incelendiğinde istem formatının detaylandırılmasının eğitim kaybına pozitif yönde etki ettiği her iki veri setinde de görülmektedir. TR2 formatından daha karmaşık olan TR1 ve TR3 formatları daha başarılı sonuçlar elde etmiştir. Ayrıca TR1 ve TR3 sonuçları arasında her iki veri setinde de TR3 formatı daha iyi bir sonuç elde etmiştir. Burada da aynı karmaşıklıkta iki formatın etiket seçiminin de sonuca etki ettiği görülmektedir.

Tüm bu sonuçlar doğrultusunda, IV numaralı istem formatının diğer formatlara kıyasla daha başarılı performans gösterdiği görülmektedir. Bu bağlamda TR4 istem formatı, Llama2 modeliyle birlikte diğer yaygın istem formatlarıyla karşılaştırmalı olarak

değerlendirilmiştir. TR4 formatının diğer formatlar ile karşılaştırılması Şekil 3.3 ve Şekil 3.4'te verilmiştir.



Şekil 3.3. Llama2 Modelinin NovusResearch Veri Seti Kullanılarak Farklı İstem Formatlarıyla Karşılaştırması (I: LLaMA2, II: LLaMA3, III: Phi3, IV: TR4)



Şekil 3.4. Llama2 Modelinin Yudum Veri Seti Üzerinde Farklı İstem Formatlarıyla Karşılaştırması (I: Llama2, II: Llama3, III: Phi3, IV: TR4)

Elde edilen grafikler doğrultusunda, kayıp (loss) değeri en düşük seviyede başlayan IV numaralı grafiktir. Eğitim kaybı tüm grafiklerde zamanla azalsa da en iyi değişiklik IV numaralı grafiktedir. Bu da modelin diğer modellere göre daha iyi şekilde

öğrenme sağladığını göstermektedir. TR4 istem formatının Türkçe eğitim verisi için Llama2 modeline adapte olma sürecinin daha yüksek olduğu da görülmektedir. Sonuçlar incelendiğinde, Llama2 dil modeli üzerinde farklı istem formatlarının kullanılmasıyla, NovusResearch ve Yudum veri setleri için TR4 formatının hem diğer özelleştirilmiş Türkçe istem formatlarına hem de yaygın olarak kullanılan formatlara kıyasla Türkçe veri setlerinde daha üstün performans gösterdiği görülmektedir.

Llama2 modeli üzerinde farklı veri setlerindeki istem formatlarının performans metrikleri ile birlikte skorları Çizelge 3.5 ve Çizelge 3.6'da verilmiştir.

Çizelge 3.5. Llama2 Modeli: NovusResearch Veri Setinde İstem Formatlarına Göre Performans Metrikleri

İstem Format	BLEU	R-1 (F1)	R-2 (F1)	R-L (F1)	MET	PPL	BLT	BERT (F1)	D-1	D-2
TR1	0.027	0.181	0.058	0.134	0.088	111.440	-0.835	0.445	0.133	0.504
TR2	0.034	0.191	0.061	0.140	0.088	39.607	-0.707	0.472	0.141	0.531
TR3	0.034	0.183	0.057	0.133	0.086	25.503	-0.621	0.467	0.139	0.539
TR4	**0.035**	0.192	0.064	0.144	**0.089**	**11.581**	-0.582	0.476	0.149	0.561
Llama2	0.032	**0.194**	**0.069**	**0.150**	0.086	236.580	**-0.554**	**0.484**	**0.183**	**0.602**
Llama3	0.021	0.192	0.068	0.146	0.086	46.718	-0.855	0.427	0.099	0.339
Phi3	0.015	0.143	0.032	0.105	0.046	223.421	-0.694	0.450	0.173	0.579

Çizelge 3.6. Llama2 Modeli: Yudum Veri Setinde İstem Formatlarına Göre Performans Metrikleri

İstem Format	BLEU	R-1 (F1)	R-2 (F1)	R-L (F1)	MET	PPL	BLT	BERT (F1)	D-1	D-2
TR1	0.023	0.181	0.053	0.128	0.081	110.487	-0.801	0.444	0.116	0.465
TR2	0.028	0.188	0.054	0.133	0.080	46.418	-0.675	0.466	0.113	0.474
TR3	0.026	0.183	0.051	0.128	0.079	33.453	-0.604	0.462	0.118	0.500
TR4	**0.029**	0.190	0.058	0.137	**0.084**	**12.476**	-0.612	0.474	0.128	0.518
Llama2	0.023	**0.193**	**0.061**	**0.142**	0.078	230.291	**-0.558**	**0.479**	**0.149**	**0.546**
Llama3	0.022	0.182	0.056	0.132	0.075	52.321	-0.690	0.446	0.109	0.411
Phi3	0.011	0.131	0.027	0.093	0.039	1033.51	-0.729	0.438	0.155	0.544

Yukarıda verilen Çizelge 3.5 ve Çizelge 3.6' dan çıkan sonuçlara göre, doğal dil işleme alanında kullanılan BLEU, ROUGE, METEOR, BERTScore gibi otomatik değerlendirme metrikleri, özellikle İngilizce dilinde geliştirildikleri ve bu dilin yapısal özelliklerine göre daha verimli oldukları için, Türkçe gibi eklemeli dillere uygulandığında bazı sınırlılıklar göstermekte olup, yapılan değerlendirmelerde, modelin üretmiş olduğu yanıtlar semantik olarak doğru ve bağlama uygun olsa dahi, kelime veya kelime dizilimleri farklı ifade edildiğinden, uyuşmazlıklara sebep olmaktadır. Bu uyuşmazlık sebebiyle de metrikler düşük puanlar verebilmektedir.

Çizelge 3.7'de girdi, doğru yanıt (Ground Truth) ve model çıktısı verilmiştir. Burada anlamsal olarak çıktının doğru olduğu ancak kelime eşleşmeleri tam olarak sağlanmadığından dolayı n-gram eşleşmelere bakan skor metriklerinde daha düşük skorlar elde edilmiştir.

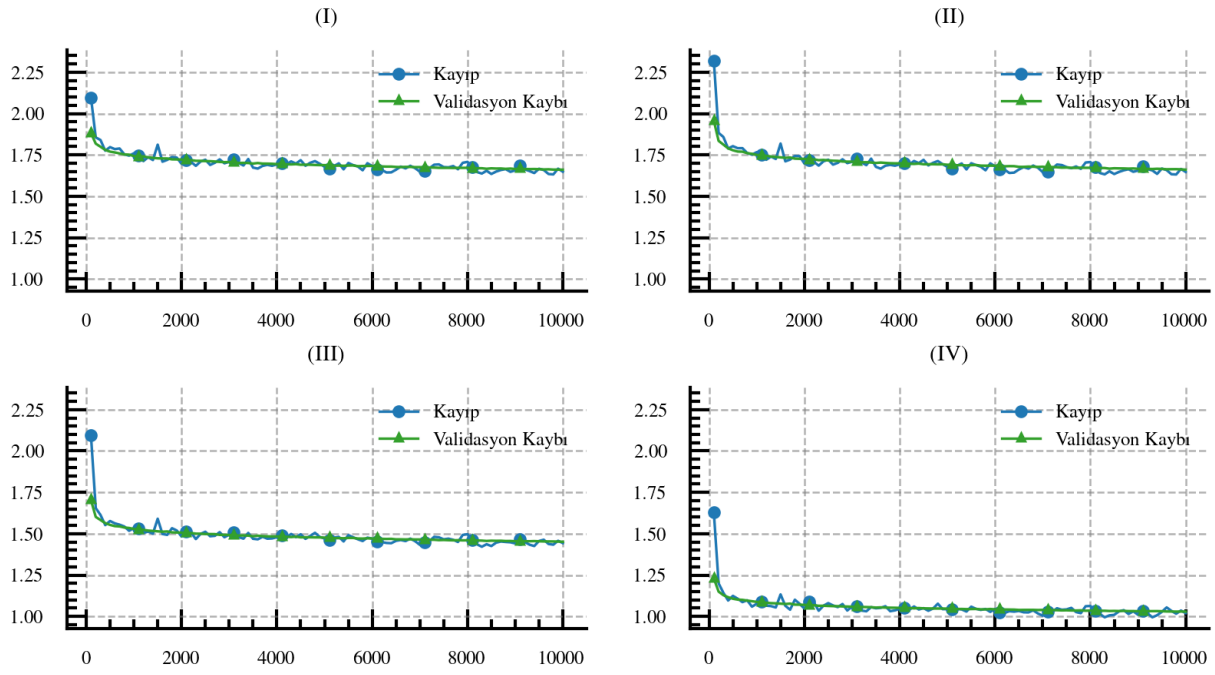
Çizelge 3.7. Gerçek Çıktı İle Üretilen Çıktı Metnin Karşılaştırılması

Girdi	Doğru Yanıt	Çıktı
<|soru|>\n 1 ile 10 arasındaki tüm sayıların toplamını hesaplayın.\n <|son|>\n	1 ile 10 arasındaki tüm sayıların toplamı 55'tir.	55

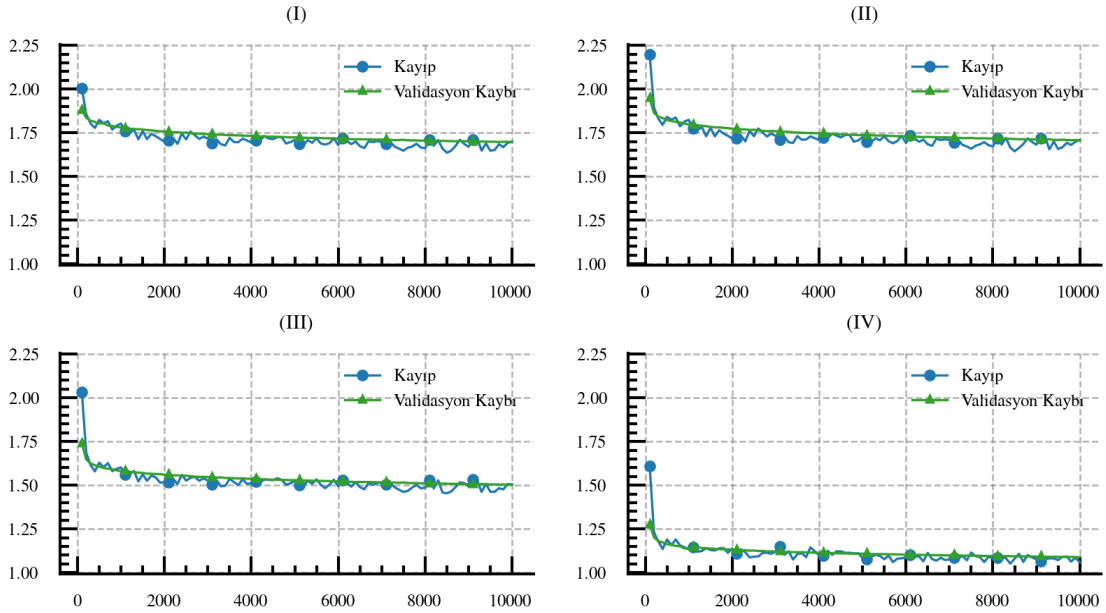
Tüm bunlara rağmen, Llama2, Llama3 ve Phi3 gibi diğer istem formatlarıyla birlikte metriklerin tümünü değerlendirecek olursak, TR (TR1, TR2, TR3, TR4) serisi olarak adlandırdığımız istem formatları, diğer istem formatlarından daha iyi sonuç vermiş veya en iyi metrik sonucuna çok yakın sonuçlar üretmiştir. Özellikle Perplexity gibi modelin bir sonraki kelimeyi tahmin etme konusundaki belirsizliğini ölçen performans metriklerinde ise TR serisi istem formatlarındaki yapının, modelin girdiyi daha doğal, açık ve yönlendirici algılamasını sağlamasından ötürü, modelin belirsizliği azalmış olup Perplexity performans metriğini TR4 istem formatında en iyi sonucuna ulaşmasını sağlamıştır. Öte yandan METEOR metriği, ROUGE, BLEU gibi yüzeysel n-gram benzerliklerden daha çok semantik tabanlı bir metrik olduğundan dolayı, TR formatlarında yalnızca yüzeysel eşleşmelere değil, aynı zamanda anlamsal doğruluğada katkı sunduğundan dolayı TR4 istem formatı en iyi sonucu vermektedir.

3.3.2.Llama3 Modelinin Farklı İstem Formatları İçin Performans Analizi

TR istem formatlarının Llama3 dil modelindeki farklı veri setleri üzerinde elde edilen kayıp grafikleri Şekil 3.5 ve Şekil 3.6'da verilmiştir.



Şekil 3.5. Llama3 Modelinin NovusResearch Veri Seti Üzerinde Türkçe İstem Formatlarıyla Karşılaştırması (I: TR1, II: TR2, III: TR3, IV: TR4)



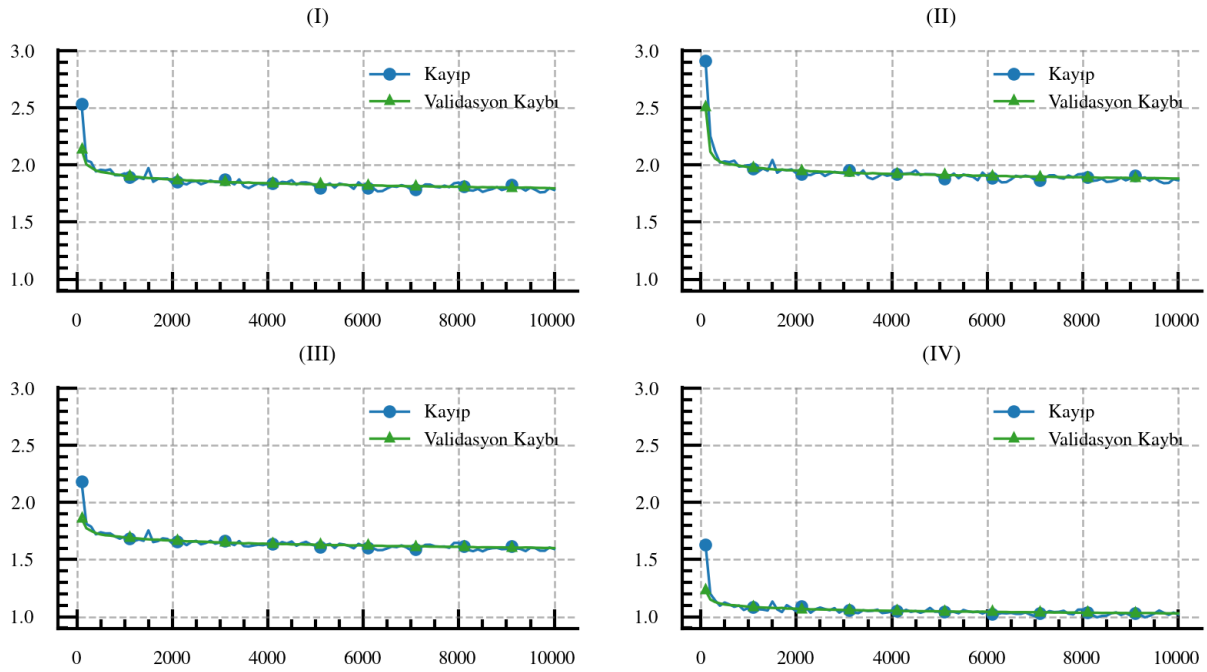
Şekil 3.6. Llama3 Modelinin Yudum Veri Seti Üzerinde Türkçe İstem Formatlarıyla Karşılaştırması (I: TR1, II: TR2, III: TR3, IV: TR4)

Özelleştirilmiş Türkçe istem formatlarına göre tüm grafikler incelendiğinde, her formatta eğitim ve validasyon kayıplarının genel olarak kararlı ve istikrarlı bir şekilde düşüş sergilediği görülmektedir.

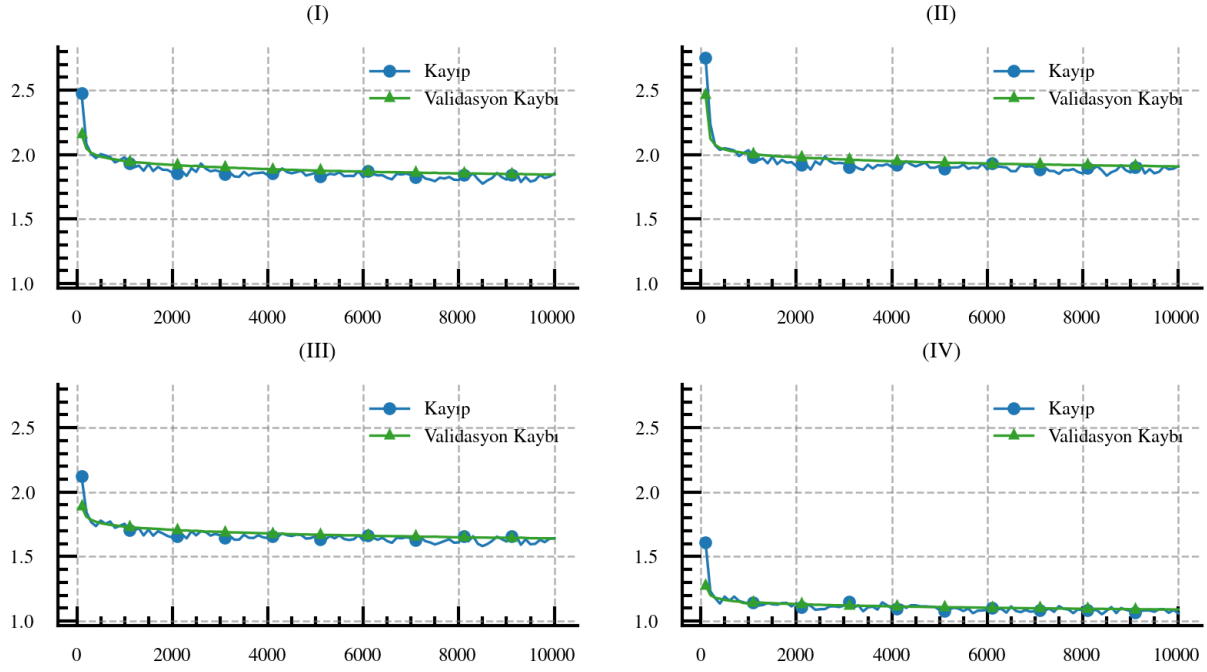
Llama2 modelinde olduğu gibi, en iyi sonuç veren istem formatı olarak, TR4 formatı diğer formatlara kıyasla daha başarılı performans sergilemektedir. TR4 modeli daha ilk aşamada diğer modellerin sonuçta ulaştığı kayıp (loss) değerinden eğitime başlamıştır. Bununla birlikte validasyon kaybı, tüm sonuçlarda eğitim grafiği ile paralellik göstermiştir.

Yine Llama2 modelinde olduğu gibi Llama3 modelinde de benzer karmaşıklıkta olan iki format arasında kayıp farkı bulunmaktadır. Farklı bir model için de bu sonuçların paralellik göstermesi, yalnızca karmaşıklığın sonuçları etkilemediğini göstermektedir.

Elde edilen bulgular doğrultusunda, TR4 istem formatının LLaMA3 modeli ile en başarılı sonucu elde eden istem formatı olduğu gözlemlenmiştir. Bu çerçevede, TR4 formatının diğer yaygın istem formatlarıyla karşılaştırılmalı performansı Şekil 3.7 ve Şekil 3.8'de verilmiştir.



Şekil 3.7. Llama3 Modelinin NovusResearch Veri Seti Kullanılarak Farklı İstem Formatlarıyla Karşılaştırması (I: Llama2, II: Llama3, III: Phi3, IV: TR4)



Şekil 3.8. Llama3 Modelinin Yudum Veri Seti Kullanılarak Farklı İstem Formatlarıyla Karşılaştırması (I: Llama2, II: Llama3, III: Phi3, IV: TR4)

Şekillerdeki sonuçlardan hareketle, TR4 istem formatının, diğer istem formatlarına göre daha düşük kayıp değerleri ile başlaması, aynı zamanda eğitim ve

validasyon kaybının en düşük seviyede olması, TR4 modelinin kayıp açısından en başarılı format olduğunu göstermektedir. Dolayısıyla TR4 formatı, diğer yaygın olarak kullanılan istem formatlarına göre NovusResearch ve Yudum Türkçe veri setlerinde daha iyi bir performans sergilemekte ve LLaMA3 modeline adapte olma ve öğrenme sürecini daha iyi yönettiğini göstermektedir.

Çizelge 3.8 ve Çizelge 3.9'da Llama3 modeli üzerinde farklı veri setlerindeki istem formatlarının performans metrik skorları görülmektedir.

Çizelge 3.8. Llama3 Modeli: NovusResearch Veri Setinde İstem Formatlarına Göre Performans Metrikleri

İstem Format	BLEU	R-1 (F1)	R-2 (F1)	R-L (F1)	MET	PPL	BLT	BERT (F1)	D-1	D-2
TR1	0.027	0.206	0.073	0.142	**0.122**	69.507	-0.817	0.473	0.114	0.458
TR2	0.033	0.204	0.068	0.137	0.119	64.635	-0.683	0.501	0.116	0.498
TR3	0.035	0.213	0.073	0.144	0.118	20.097	-0.612	0.497	0.124	0.517
TR4	**0.044**	**0.229**	**0.082**	**0.162**	0.121	**9.004**	**-0.577**	**0.522**	0.145	**0.569**
Llama2	0.019	0.185	0.051	0.121	0.088	61.397	-0.755	0.461	**0.160**	0.550
Llama3	0.031	0.197	0.065	0.128	0.118	77.237	-0.661	0.495	0.123	0.521
Phi3	0.028	0.194	0.057	0.130	0.098	64.114	-0.614	0.502	0.132	0.507

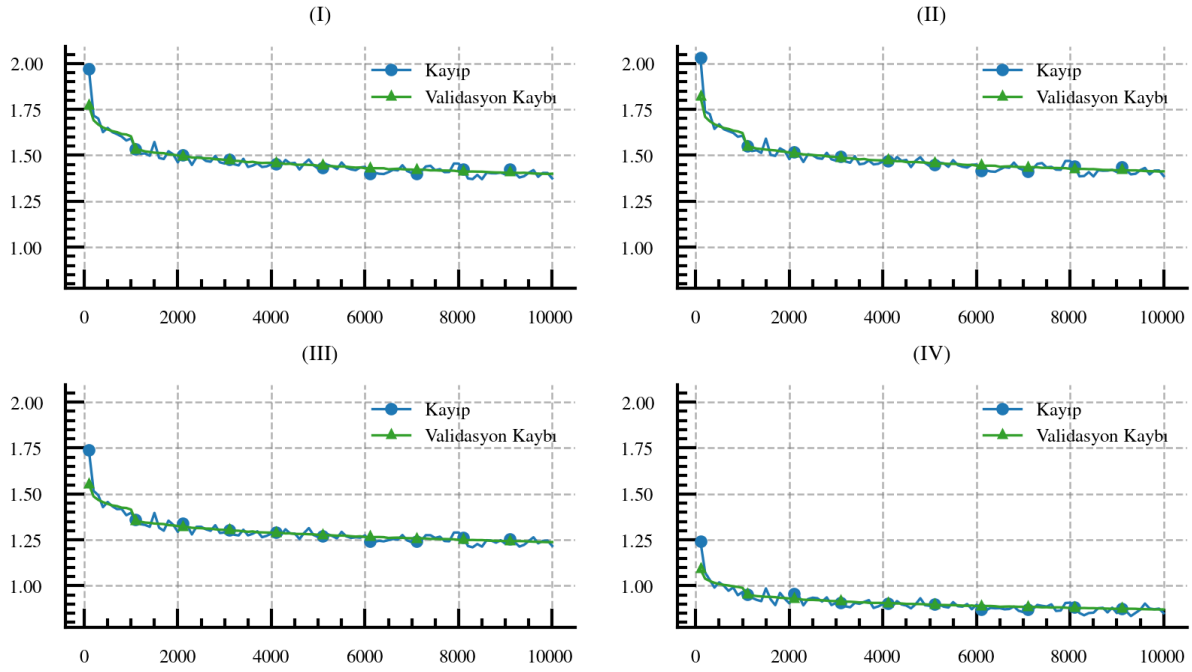
Çizelge 3.9. Llama3 Modeli: Yudum Veri Setinde İstem Formatlarına Göre Performans Metrikleri

İstem Format	BLEU	R-1 (F1)	R-2 (F1)	R-L (F1)	MET	PPL	BLT	BERT (F1)	D-1	D-2
TR1	0.029	0.224	0.072	0.148	0.124	120.996	-0.787	0.485	0.107	0.489
TR2	0.030	0.213	0.067	0.138	0.111	30.485	-0.674	0.498	0.100	0.464
TR3	0.031	0.219	0.069	0.142	**0.114**	25.818	**-0.566**	0.503	0.106	0.488
TR4	**0.037**	**0.222**	**0.074**	**0.151**	0.110	**9.685**	-0.593	**0.505**	0.119	0.513
Llama2	0.021	0.196	0.054	0.126	0.092	83.954	-0.692	0.483	0.115	0.485
Llama3	0.028	0.208	0.064	0.129	0.113	86.952	-0.635	0.495	0.104	0.479
Phi3	0.025	0.204	0.057	0.133	0.086	98.070	-0.667	0.497	**0.139**	**0.531**

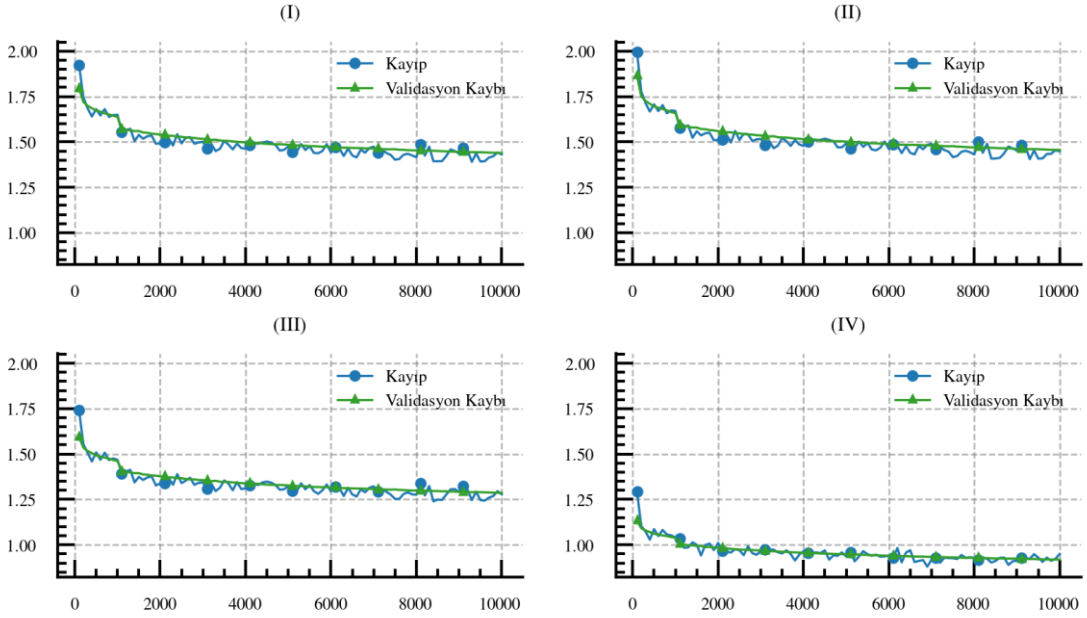
Çizelge 3.7'de de görüldüğü üzere, modelin ürettiği yanıtlar anlamsal olarak doğru ve bağlama uygun olsa dahi, değerlendirme için kullanılan BLUE, ROUGE gibi n-gram tabanlı metrikler yalnızca yüzeysel kelime ve kelime dizisi eşleşmelerine odaklandığı için sonuçları tam olarak doğru yansıtmamaktadır. N-gram tabanlı metrikler bu yapısal varyasyonları algılayamadığından dolayı, modelin verdiği doğru yanıtları da düşük skorlarla değerlendirmektedir. N-gram metriklerin sınırlılıklarına rağmen, Llama2, Llama3, Phi3 gibi istem formatlarına kıyasla TR4 istem formatı genel olarak daha başarılı sonuçlar üretmiştir.

3.3.3.Mistral Modelinin Farklı İstem Formatları İçin Performans Analizi

TR istem formatlarının Mistral dil modelindeki farklı veri setleri üzerindeki performansları Şekil 3.9 ve Şekil 3.10'da karşılaştırılmalı olarak gösterilmektedir.



Şekil 3.9. Mistral Modelinin NovusResearch Veri Seti Üzerinde Türkçe İstem Formatlarıyla Karşılaştırması (I: TR1, II: TR2, III: TR3, IV: TR4)

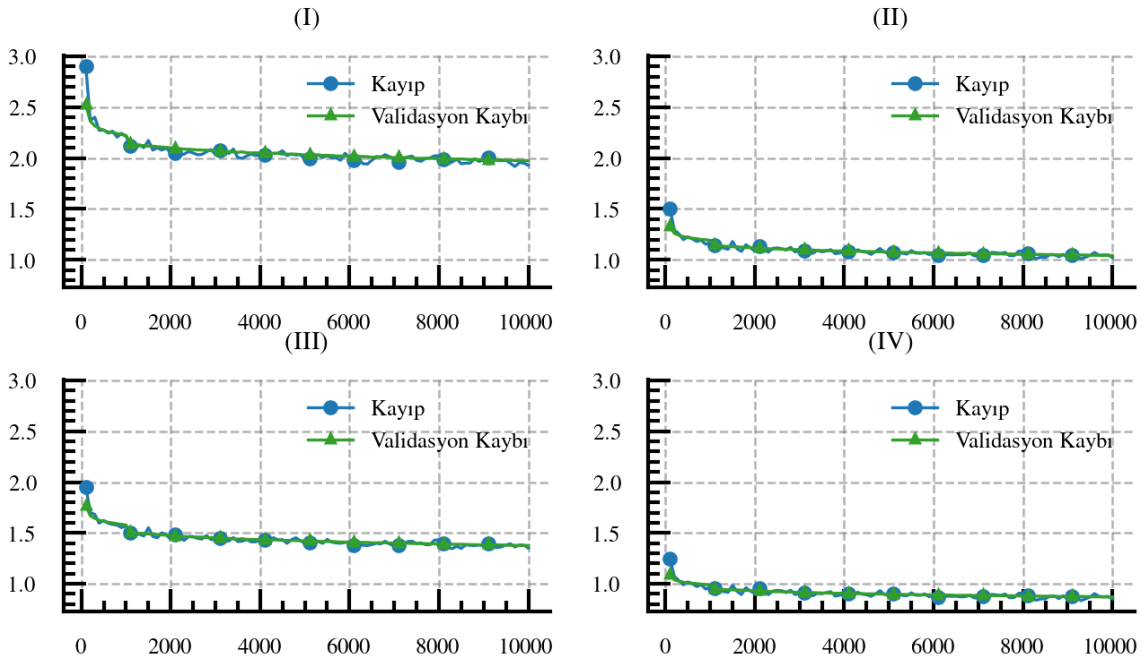


Şekil 3.10. Mistral Modelinin Yudum Veri Seti Üzerinde Türkçe İstem Formatlarıyla Karşılaştırması (I: TR1, II: TR2, III: TR3, IV: TR4)

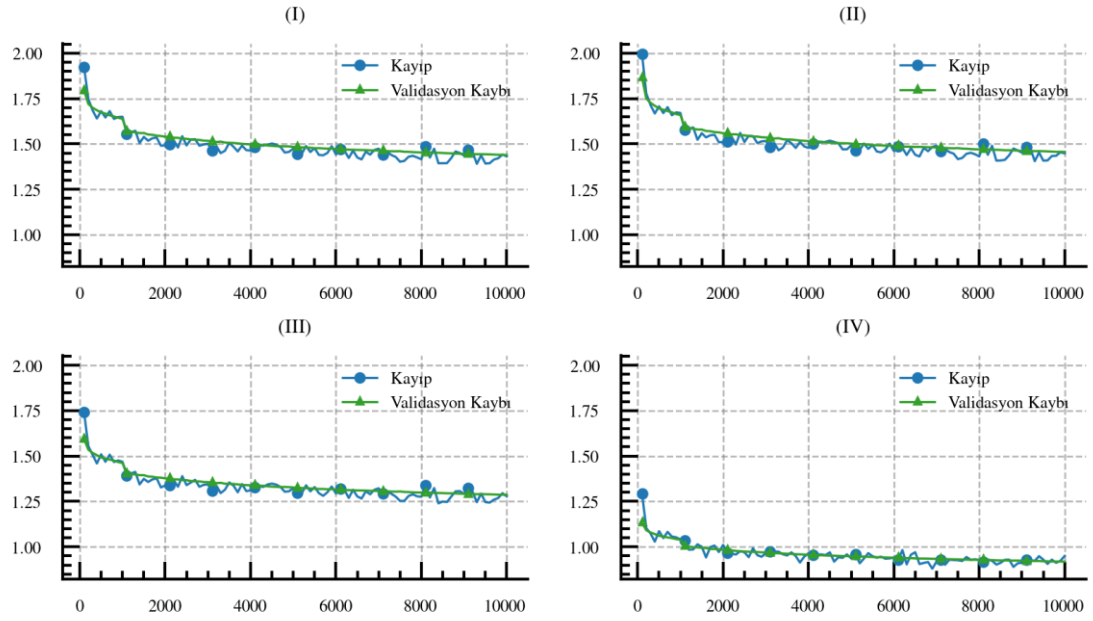
Özelleştirilmiş Türkçe istem formatlarına göre yapılan analizde, Mistral dil modeliyle gerçekleştirilen eğitim süreçlerinde her bir format için eğitim ve validasyon kayıplarının zamanla tutarlı bir şekilde azaldığı grafiksel sonuçlardan çıkarılmaktadır.

Ancak grafiklere dayanarak yapılan analizlerde, IV numaralı grafik ile temsil edilen TR4 formatının kayıp değerinin diğer istem formatlarına kıyasla daha düşük bir değerden başlaması ve ilerleyen adımlarda, diğer formatların ulaştığı minimum kayıplardan daha düşük değere ulaşması sebebiyle TR4 istem formatının diğer formatlara göre performansının daha iyi olduğunu analiz edilmektedir. Tüm bunlarla birlikte, validasyon kaybının da hızlı bir şekilde düşüş göstermesi ve kısa sürede stabil noktaya ulaşması, modelin yalnızca etkili bir biçimde öğrenmediğini, aynı zamanda yüksek bir genelleme yeteneğine sahip olduğunu da göstermektedir.

Bu bulgular doğrultusunda, TR4 istem formatı Türkçe diğer istem formatları arasında daha iyi bir performansa sahiptir. Elde edilen sonuçlara istinaden TR4 istem formatı ile eğitilen Mistral modelinin, farklı Türkçe veri setleri üzerindeki öğrenme ve genelleme performansları Şekil 3.11 ve Şekil 3.12'de karşılaştırılarak gösterilmektedir.



Şekil 3.11. Mistral Modelinin NovusResearch Veri Seti Kullanılarak Farklı İstem Formatlarıyla Karşılaştırması (I: Llama2, II: Llama3, III: Phi3, IV: TR4)



Şekil 3.12. Mistral Modelinin Yudum Veri Seti Kullanılarak Farklı İstem Formatlarıyla Karşılaştırması (I: LLaMA2, II: LLaMA3, III: Phi3, IV: TR4)

Şekillerden elde edilen sonuçlara göre, TR4 istem formatının, diğer formatlara kıyasla daha düşük kayıp değerleriyle eğitime başladığı ve eğitim sürecinin sonunda hem eğitim hem de validasyon kayıplarının en düşük seviyede stabil hale geldiği

gözlemlenmektedir. Bu durum, TR4 formatının Mistral dil modeli ile yüksek bir uyum içerisinde çalıştığını ve modelin öğrenme sürecini daha verimli yönettiğini göstermektedir.

Sonuç olarak, Mistral dil modeli farklı istem formatlarıyla eğitildiğinde en başarılı performansı Türkçeye özgü biçimlendirilmiş istem formatıyla olduğu açıkça ortaya konulmaktadır. Çizelge 3.10 ve Çizelge 3.11'de Mistral modeli üzerinde farklı veri setlerindeki istem formatlarının performans metrik skorları görülmektedir.

Çizelge 3.10. Mistral Modeli: NovusResearch Veri Setinde İstem Formatlarına Göre Performans Metrikleri

İstem Format	BLEU	R-1 (F1)	R-2 (F1)	R-L (F1)	MET	PPL	BLT	BERT (F1)	D-1	D-2
TR1	0.049	0.244	0.093	0.175	0.131	125.344	-0.786	0.510	0.136	0.533
TR2	0.064	**0.258**	**0.100**	**0.187**	0.135	47.381	-0.621	0.549	0.149	0.575
TR3	**0.066**	**0.258**	0.098	**0.187**	**0.138**	29.796	-0.503	0.552	0.148	0.575
TR4	0.057	0.245	0.091	0.175	0.131	**10.560**	-0.541	0.538	0.144	0.563
Llama2	**0.066**	0.257	0.098	0.186	0.132	204.695	**-0.500**	0.552	0.153	0.576
Llama3	0.060	0.246	0.093	0.178	0.127	56.106	-0.630	0.525	0.156	0.579
Phi3	0.049	0.248	0.087	0.186	0.106	307.662	-0.476	**0.567**	**0.168**	**0.590**

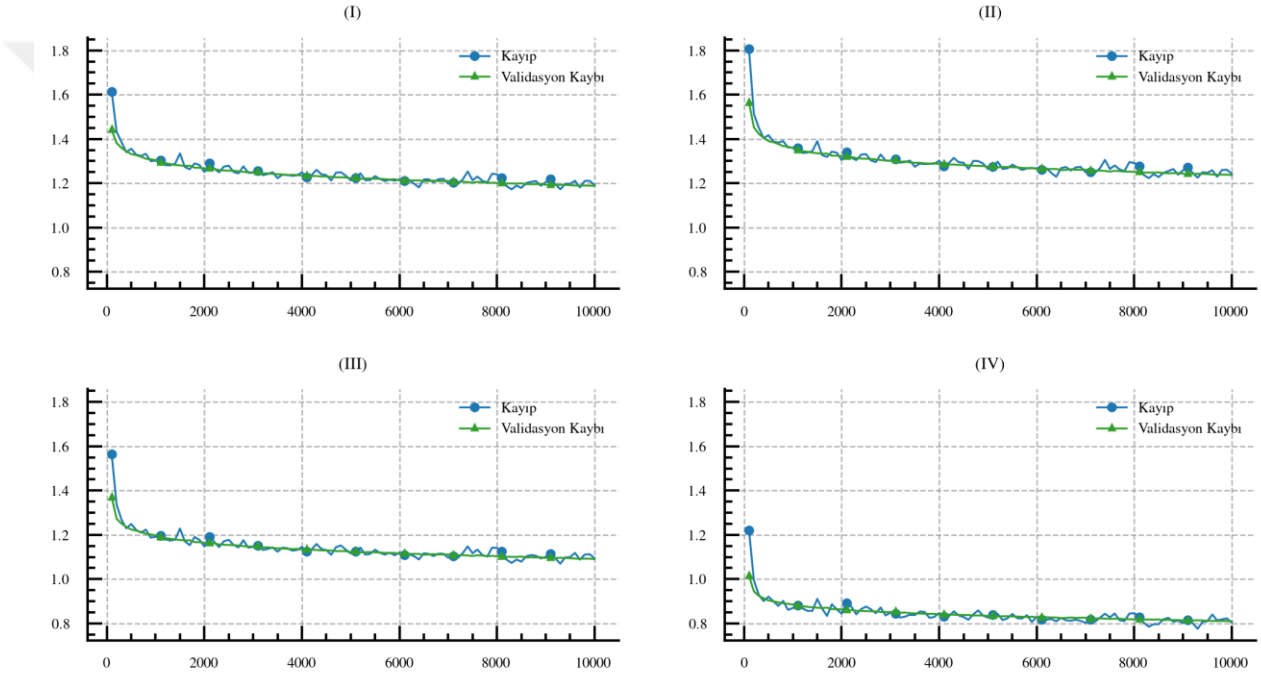
Çizelge 3.11. Mistral Modeli: Yudum Veri Setinde İstem Formatlarına Göre Performans Metrikleri

İstem Format	BLEU	R-1 (F1)	R-2 (F1)	R-L (F1)	MET	PPL	BLT	BERT (F1)	D-1	D-2
TR1	0.041	0.246	0.086	0.168	0.124	121.228	-0.757	0.509	0.110	0.482
TR2	0.045	0.245	0.085	0.168	0.119	52.242	-0.672	0.526	0.117	0.505
TR3	0.047	0.254	0.090	0.174	0.125	31.697	-0.518	0.537	0.114	0.500
TR4	0.045	0.230	0.078	0.157	0.116	**9.885**	-0.578	0.515	0.115	0.500
Llama2	**0.053**	**0.259**	**0.091**	**0.178**	**0.127**	177.929	**-0.506**	**0.545**	0.118	0.512
Llama3	0.037	0.214	0.079	0.155	0.098	56.064	-0.679	0.496	0.131	0.523
Phi3	0.040	0.231	0.071	0.160	0.095	185.312	-0.535	0.536	**0.133**	**0.529**

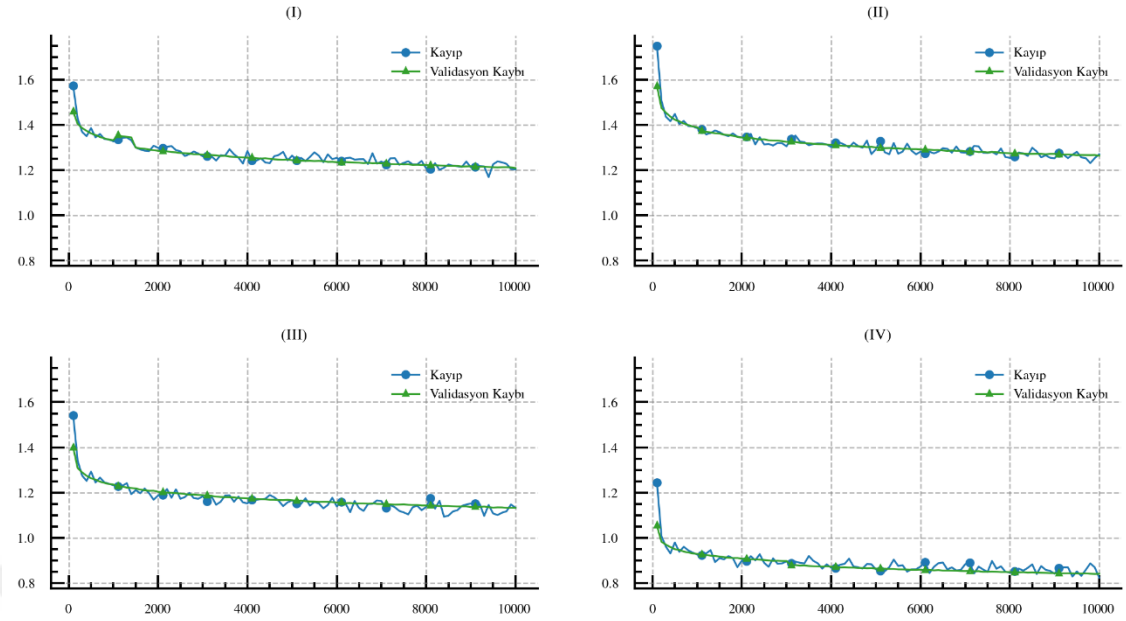
Üst bölümlerde de ortaya konulduğu üzere, n-gram temelli otomatik metrikler, yalnızca yüzey düzeyindeki kelime ve ifade eşleşmelerini baz aldığından dolayı performans metriklerinde istenen sonuçlara ulaşılamamakta olup model çıktılarının gerçek anlamda ne ölçüde doğru olduğunu tam olarak yansıtamamaktadır.

3.3.4. Phi3 Modelinin Farklı İstem Formatları İçin Performans Analizi

Büyük dil modellerinde olduğu gibi küçük bir dil modelinde de farklı istem formatları denenmiş ve sonuçlar elde edilmiştir. Phi3 modeli için bu sonuçlar Şekil 3.13 ve Şekil 3.14'te verilmiştir.



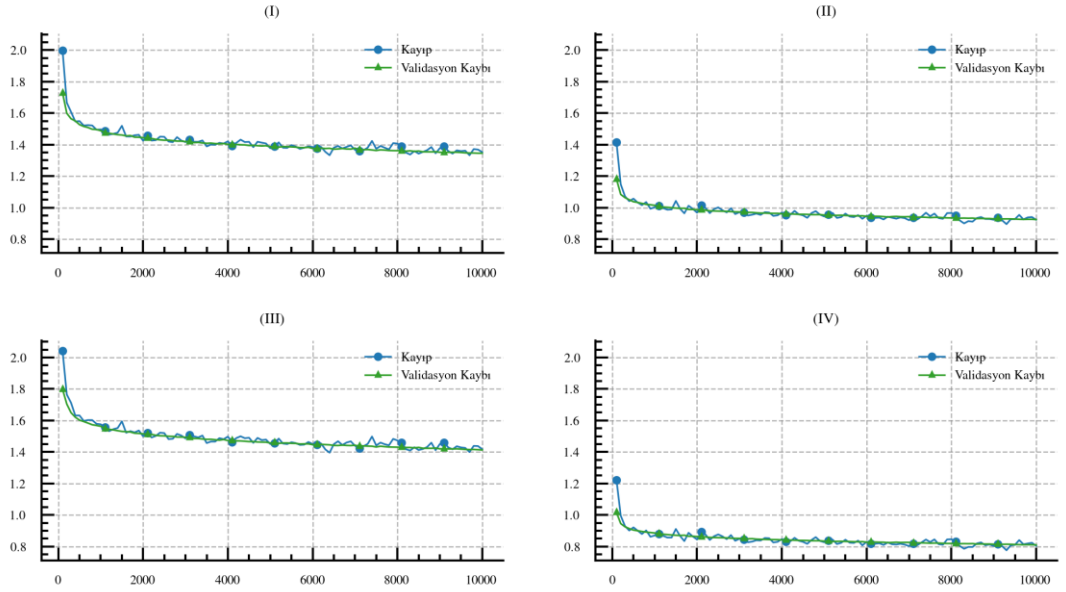
Şekil 3.13. Phi3 Modelinin NovusResearch Veri Seti Üzerinde Türkçe İstem Formatlarıyla Karşılaştırması (I: TR1, II: TR2, III: TR3, IV: TR4)



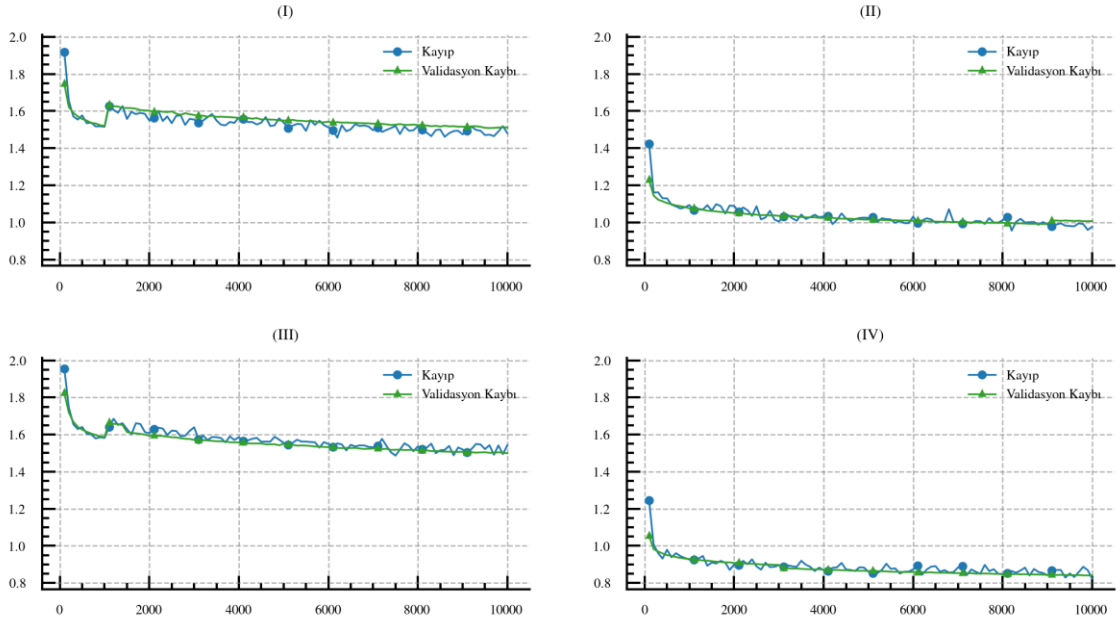
Şekil 3.14. Phi3 Modelinin Yudum Veri Seti Üzerinde Türkçe İstem Formatlarıyla Karşılaştırması (I: TR1, II: TR2, III: TR3, IV: TR4)

Dil modelinin küçük olması farkı formatlar arasındaki başarı dengesini değiştirmemiştir. Büyük dil modellerinde olduğu gibi küçük dil modelinde de formatların kayıp sıralaması benzer oranları korumaktadır. TR4 formatı en iyi sonuçları elde etmiş ve diğer formatların kayıpları da önceki sonuçlar ile paralellik göstermiştir.

TR4 istem formatı ile eğitilen Phi3 modelinin, farklı Türkçe veri setleri üzerindeki öğrenme ve genelleme performansları Şekil 3.15 ve Şekil 3.16'da karşılaştırmalı olarak gösterilmektedir.



Şekil 3.15. Phi3 Modelinin NovusResearch Veri Seti Kullanılarak Farklı İstem Formatlarıyla Karşılaştırması (I: Llama2, II: Llama3, III: Phi3, IV: TR4)



Şekil 3.16. Phi3 Modelinin Yudum Veri Seti Kullanılarak Farklı İstem Formatlarıyla Karşılaştırması (I: Llama2, II: Llama3, III: Phi3, IV: TR4)

İncelenen sonuçlarda yalnızca Türkçe olan formatın değil LLama3 formatının da phi3 modelinin varsayılan formatına göre daha iyi sonuçlar elde ettiğini göstermektedir. Bu da etiketlerin dilinin dışında karmaşıklığının da modeller üzerinde etkisini göstermektedir. Her ne kadar en karmaşık olan Llama3 formatı Phi3 için en iyi kayıp

değerlerini elde etse de yine TR4 modeli en iyi kayıp sonuçlarını göstermiştir. Bu da etiketlerin Türkçe belirlenmesinin de etkisinin bulunduğunu ispatlar niteliktedir.

Çizelge 3.12 ve Çizelge 3.13'de Phi3 modeli üzerinde farklı veri setlerindeki istem formatlarının BLEU ve ROUGE skorları görülmektedir.

Çizelge 3.12. Phi3 Modeli: NovusResearch Veri Setinde İstem Formatlarına Göre Performans Metrikleri

İstem Format	BLEU	R-1 (F1)	R-2 (F1)	R-L (F1)	MET	PPL	BLT	BERT (F1)	D-1	D-2
TR1	0.006	0.078	0.006	0.055	0.021	382.549	-0.966	0.375	0.432	0.880
TR2	0.003	0.080	0.005	0.055	0.020	422.239	-0.833	0.377	0.422	0.877
TR3	**0.007**	0.083	0.006	0.057	0.024	441.747	-0.790	0.388	0.426	0.882
TR4	0.005	0.082	0.006	0.057	0.022	452.529	-0.810	0.381	0.429	0.877
Llama2	0.005	**0.090**	**0.007**	**0.060**	**0.026**	**408.371**	-0.724	0.389	0.424	0.882
Llama3	0.005	0.079	0.006	0.054	0.023	432.846	-0.845	0.378	0.406	0.869
Phi3	0.009	0.106	0.012	0.073	0.033	389.570	**-0.695**	**0.407s**	**0.430**	**0.886**

Çizelge 3.13. Phi3 Modeli: Yudum Veri Setinde İstem Formatlarına Göre Performans Metrikleri

İstem Format	BLEU	R-1 (F1)	R-2 (F1)	R-L (F1)	MET	PPL	BLT	BERT (F1)	D-1	D-2
TR1	0.004	0.086	0.006	0.058	0.022	395.912	-0.937	0.378	0.436	0.877
TR2	0.006	0.087	0.006	0.058	0.022	427.317	-0.790	0.377	0.421	0.877
TR3	0.006	0.092	0.007	0.063	0.025	454.186	-0.736	0.393	0.425	**0.878**
TR4	0.004	0.087	0.006	0.059	0.023	451.982	-0.780	0.385	**0.435**	0.877
Llama2	0.005	0.096	0.008	0.065	0.027	427.402	-0.692	0.393	0.429	0.884
Llama3	0.005	0.082	0.005	0.055	0.022	437.143	-0.823	0.379	0.404	0.865
Phi3	**0.008**	**0.108**	**0.012**	**0.072**	**0.033**	**378.890**	**-0.679**	**0.409**	0.254	0.757

Otomatik değerlendirme metriklerinin yapısal sınırlılıkları, Türkçe dilindeki söz dizilimlerine adapte olurken zorlandığı görülmektedir. Üst bölümlerde de değinildiği

üzere, üretilen yanıtlar anlam açısından doğru olsa da bu metrikler yüzeysel benzerliklere dayalı ölçüm yöntemleri sebebiyle modelin gerçek başarımını ortaya koymakta yetersiz kalmaktadır. Özellikle daha küçük ölçekli bir dil modeli olan Phi3'ün, sınırlı parametre sayısı ve geniş olmayan öğrenme kapasitesi nedeniyle TR serilerinde daha düşük sonuç gösterdiği izlenmektedir. Bu nedenle, metrik değerlerinin yorumlanmasında yalnızca nicel sonuçlara değil, aynı zamanda niteliksel değerlerde göz önünde bulundurulmalıdır.

BÖLÜM 4

SONUÇ ve TARTIŞMA

Derin öğrenme tabanlı yaklaşımlar, özellikle doğal dil işleme (NLP) alanında son yıllarda büyük ilerlemeler sağlamıştır. Çok katmanlı yapay sinir ağları sayesinde, dilin yapısal ve anlamsal özellikleri daha derin ve etkili biçimde öğrenilebilmekte, böylece geleneksel yöntemlerin sınırlı kaldığı birçok görevde önemli ilerlemeler elde edilmektedir. Bu ilerlemeler neticesinde, metin anlama, özetleme, soru-cevap, makine çevirisi gibi birçok karmaşık görev, artık derin öğrenme tabanlı sistemler aracılığıyla daha başarılı şekilde gerçekleştirilmektedir. Derin öğrenme tabanlı yaklaşımlarda dönüm noktası sayılabilecek transformer tabanlı mimarilerin geliştirilmesiyle birlikte, dil modelleri çok daha geniş veri kümeleri üzerinde eğitilebilmiş ve bu sayede genel dil becerilerinde kayda değer bir artış sağlamıştır.

Derin öğrenme, günümüzde yalnızca akademik alanda değil çok sayıda çeşitli görevde etkin bir biçimde kullanılmakta ve birçok sektörde yenilikçi uygulamaların geliştirilmesinin önünü açmaktadır. Bu sektörlerin başında sağlık, finans ve bankacılık, hukuk, e-ticaret, kamu hizmetleri, eğitim teknolojileri, oyun ve eğlence gibi sektörlerde etkin biçimde kullanılmaktadır. Elektronik sağlık kayıtlarının analizi, müşteri hizmetleri otomasyonu, sözleşme inceleme, otomatik değerlendirme sistemleri gibi uygulamalar doğal dil işlemenin sunduğu temel katkılar arasında yer almaktadır. Tüm bu çalışmaların temelinde, büyük veri kümeleri üzerinde eğitilmiş, dil modelleri yer almaktadır.

Dil modelleri, belirli görevlerde daha yüksek başarı elde edebilmek amacıyla özelleştirilmekte ve ince ayar (fine-tuning) yöntemi ile eğitilmektedir. Bu süreçte, önceden eğitilmiş bir model, hedeflenen görev alanına özgü veri setleriyle yeniden

yapılandırılarak göreve daha uygun hale getirilmektedir. Bir modelin belirli bir görevi başarıyla yerine getirebilmesi için, düşük kayıp (loss) değerlerinin düşüklüğüyle hem de değerlendirme metrikleri üzerinden elde edilen yüksek skorlarla ölçülmektedir. Başarılı bir görev performansı için, modelin yalnızca genel dil bilgisini öğrenmesi yeterli olmayıp, aynı zamanda görev bağlamına uygun yanıtlar üretme yeteneğini de kazanması gerekmektedir.

Model eğitimi sürecinde istem (prompt) formatının önemi, kullanılan veri kümesi veya ön işleme adımlara kadar kritiktir. Her model, varsayılan bir istem formatıyla birlikte sunulsa da çalışmalar göstermiştir ki; aynı veri seti, aynı model mimarisi ve benzer parametreler kullanılsa bile, yalnızca istem formatındaki farklılık nedeniyle modelin öğrenme süreci, çıktı kalitesi ve genel performansı anlamlı ölçüde değişebilmektedir.

Tüm bunlara ek olarak, öncelikle özelleştirilmiş dört Türkçe istem formatı kendi arasında test edilmiş ve bu formatlar arasında en iyi performansı sergileyen TR4 formatı belirlenmiştir. Daha sonrasında TR4 istem formatı diğer yaygın istem formatlarıyla, farklı dil modelleri kullanılarak performansı test edilmiş ve tekrar TR4 formatı diğer varsayılan istem formatlarına göre daha düşük kayıp değerleriyle birlikte daha başarılı sonuçlar elde etmiştir. TR4 istem formatının en başarılı sonuçları vermesinin temel nedeni, diğer özelleştirilmiş Türkçe formatlardan farklı olarak, istem yapısında modele daha fazla yönlendirme ve bağlam sunan detaylı bir yapı içermektedir.

Çalışma kapsamında, Türkçe metinlerin Türkçe istem etiketleri ile eğitilmesinin, diğer istem yapılarına kıyasla daha iyi performans sağladığını göstermiştir. Aynı veri kümesiyle modelin Türkçe istem formatı kullanılarak eğitilmesi, modelin yanıt doğruluğunu arttırmıştır. Özellikle düşük kaynaklı dillerdeki sınırlı veri olanakları göz önünde bulundurulduğunda, özelleştirilmiş istem formatlarının kullanılması, modellerin daha etkili çalışmasına ve daha yüksek performans elde etmesine katkı sağlayabilecek önemli bir avantaj olarak değerlendirilebilir. Bu kapsamda, değerlendirme sürecinde BLEU, ROUGE (1,2,L), METEOR, BERTScore, Perplexity, BLEURT ve Distinct-n gibi otomatik metrikler kullanılmıştır. Ancak bu metriklerin büyük ölçüde İngilizce diline özgü yapılar dikkate alınarak geliştirilmiş olması, Türkçe gibi sondan eklemeli ve serbest sözdizimine sahip dillerde sınırlı doğrulukla sonuçlar vermesine yol açmıştır. Nitekim

yapılan değerlendirmelerde, modelin üretmiş olduğu yanıtların anlamsal olarak doğru ve bağlama uygun olmasına rağmen, yüzeysel benzerliklere dayanan n-gram eşleşmelerinin düşük olması sebebiyle performans metrik sonuçları beklenenin altında kaldığı görülmektedir. Bu bulgu, değerlendirme metriklerinin özellikle dil yapısına duyarlı olduğu durumlarda yorumlanmasının dikkatle yapılması gerekliliğini ortaya koymaktadır. Tüm bu sonuçlara rağmen Perplexity performans metriği, büyük dil modellerinde en düşük değeri TR4 istem formatı ile birlikte vermiştir, bu da metnin gelecek kelimeyi tahmin etmesinde istem formatının detayı arttıkça, daha başarılı sonuçlar ürettiğini ortaya koymaktadır.

Gelecek çalışmalarda, Türkçe, İngilizce ve çok dilli modeller için özelleştirilmiş istem formatları araştırılacaktır. Bu özel istem yapılandırmaları aracılığıyla, Türkçe, İngilizce ve çok dilli dil işleme görevlerinde daha yüksek performans elde edilmesi hedeflenmektedir.

KAYNAKLAR

A. Vaswani, “Attention is all you need,” *arXiv preprint arXiv:1706.03762*, 2017.

Abdin, M., Aneja, J., Awadalla, H., Awadallah, A., Awan, A. A., Bach, N., ... & Zhou, X. (2024). Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Amazon Web Services (AWS). (2024, November 9). *LLM nedir? - Büyük dil modellerine ayrıntılı bakış*. AWS. https://aws.amazon.com/tr/what-is/large-language-models/

Ankner, Z., Blakeney, C., Sreenivasan, K., Marion, M., Leavitt, M. L., & Paul, M. (2024). Perplexed by perplexity: Perplexity-based data pruning with small reference models. *arXiv*. https://doi.org/10.48550/arXiv.2405.20541

Auriemma Citarella, A., Barbella, M., Ciobanu, M. G., De Marco, F., Di Biasi, L., & Tortora, G. (n.d.). Rouge metric evaluation for text summarization techniques. *SSRN*. https://doi.org/10.2139/ssrn.4753220

Austin, J., Odena, A., Nye, M., Bosma, M., Michalewski, H., Dohan, D., Jiang, E., Chen, X., Jain, S., & Abid, A. (2021). Program synthesis with large language models. *arXiv*. https://arxiv.org/abs/2108.07732

Bengio, Y., Goodfellow, I., & Courville, A. (2017). *Deep learning* (Vol. 1). MIT Press.

Dang, N. C., Moreno-García, M. N., & de la Prieta, F. (2020). Sentiment analysis based on deep learning: A comparative study. *Electronics*, *9*(3), 483. https://doi.org/10.3390/electronics9030483

Erdi, B., Şahin, E. A., Toydemir, M. S., & Dökeroğlu, T. (2021). Makine öğrenmesi algoritmaları ile trol hesapların tespiti. *Düzce Üniversitesi Bilim ve Teknoloji Dergisi*, *9*(1), 430–442.

Esmer, S., Uçar, M. K., Çil, İ., & Bozkurt, M. R. (2020). Parkinson hastalığı teşhisi için makine öğrenmesi tabanlı yeni bir yöntem. *Düzce Üniversitesi Bilim ve Teknoloji Dergisi*, *8*(3), 1877–1893.

GitHub. (2025, February 1). *mistralai/mistral-inference: Official inference library for Mistral models*. GitHub. https://github.com/mistralai/mistral-inference

Goyal, M., Tatwawadi, K., Chandak, S., & Ochoa, I. (2021). DZip: Improved general-purpose lossless compression based on novel neural network modeling. In *2021 Data Compression Conference (DCC)* (pp. 153–162). https://doi.org/10.1109/DCC50243.2021.00026

IBM. (2024, November 11). *What is NLP (Natural Language Processing)?* IBM. https://www.ibm.com/topics/natural-language-processing

IBM. (2025, February 4). *What are small language models (SLM)?* IBM. https://www.ibm.com/topics/small-language-models

J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

Jelinek, F., Mercer, R. L., Bahl, L. R., & Baker, J. K. (1977). Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, *62*(S1), S63. https://doi.org/10.1121/1.2016580

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Le Scao, T., Lavril, T., Wang, T., Lacroix, T., & El Sayed, W. (2023). Mistral 7B v0.1: A 7-billion-parameter language model engineered for superior performance and efficiency. *arXiv*. https://doi.org/10.48550/arXiv.2310.06825

Karri, S. P. R., & Kumar, B. S. (2020). Deep learning techniques for implementation of chatbots. In *2020 International Conference on Computer Communication and Informatics (ICCCI)* (pp. 1–5). https://doi.org/10.1109/ICCCI48352.2020.9104142

Lavie, A., & Denkowski, M. J. (2009). The METEOR metric for automatic evaluation of machine translation. *Machine Translation*, *23*, 105–115. https://doi.org/10.1007/s10590-009-9059-4

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Li, J., Galley, M., Brockett, C., Gao, J., & Dolan, B. (2016). A diversity-promoting objective function for neural conversation models. In K. Knight, A. Nenkova, & O. Rambow (Eds.), *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 110–119). Association for Computational Linguistics. https://doi.org/10.18653/v1/N16-1014

Li, J., Yang, Y., Bai, Y., Zhou, X., Li, Y., Sun, H., Liu, Y., Si, X., Ye, Y., Wu, Y., Lin, Y., Xu, B., Ren, B., Feng, C., Gao, Y., & Huang, H. (2024). Fundamental capabilities of large language models and their applications in domain scenarios: A survey. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 11116–11141). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.599

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out* (pp. 74–81). Association for Computational Linguistics. https://aclanthology.org/W04-1013/

Microsoft Azure Blog. (2025, February 4). *Introducing Phi-3: Redefining what's possible with SLMs*. Microsoft. https://azure.microsoft.com/en-us/blog/introducing-phi-3-redefining-whats-possible-with-slms/

Microsoft Learn. (2025, February 4). *What is a BLEU score? - Azure AI Custom Translator - Azure AI services*. Microsoft. https://learn.microsoft.com/en-us/azure/ai-services/custom-translator/bleu-score

Mosavi, A., Ardabili, S. F., & Várkonyi-Kóczy, A. R. (2019). List of deep learning models. *Preprints*. https://doi.org/10.20944/preprints201908.0152.v1

Myers, D., Mohawesh, R., Chellaboina, V. I., et al. (2024). Foundation and large language models: Fundamentals, challenges, opportunities, and social impacts. *Cluster Computing, 27*, 1–26. https://doi.org/10.1007/s10586-023-04203-7

Oracle Türkiye. (2024, November 11). *Doğal dil işleme (NLP) nedir?* Oracle. https://www.oracle.com/tr/artificial-intelligence/what-is-natural-language-processing/

Öztürk, E. (2024). XCompress: LLM assisted Python-based text compression toolkit. *SoftwareX, 27*, 101847. https://doi.org/10.1016/j.softx.2024.101847

Öztürk, E., Mesut, A. Ş., & Fidan, Ö. A. (2024). A character based steganography using masked language modeling. *IEEE Access.*

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)* (pp. 311–318). Association for Computational Linguistics. https://doi.org/10.3115/1073083.1073135

S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Comput*, vol. 9, no. 8, pp. 1735–1780, 1997.

Sarker, I. H. (2021). Deep learning: A comprehensive overview on techniques, taxonomy, applications and research directions. *SN Computer Science, 2*(6), 420. https://doi.org/10.1007/s42979-021-00815-1

Sellam, T., Das, D., & Parikh, A. P. (2020). BLEURT: Learning robust metrics for text generation. *arXiv*. https://doi.org/10.48550/arXiv.2004.04696

Ślusarczyk, B. (2018). Industry 4.0: Are we ready? *Polish Journal of Management Studies, 17*(1), 232–248. https://doi.org/10.17512/pjms.2018.17.1.19

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). LLaMA: Open and efficient foundation language models. *arXiv*. https://doi.org/10.48550/arXiv.2302.13971

Tüfekci, P., & Önal, Ç. M. (2024). Kötü amaçlı yazılım tespiti için makine öğrenmesi algoritmalarının kullanımı. *Düzce Üniversitesi Bilim ve Teknoloji Dergisi, 12*(1), 307–319.

Yuan, A., Coenen, A., Reif, E., & Ippolito, D. (2022). Wordcraft: Story writing with large language models. In *Proceedings of the 27th International Conference on Intelligent User Interfaces* (pp. 841–852). https://doi.org/10.1145/3490099.3511129

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). BERTScore: Evaluating text generation with BERT. *arXiv*. https://arxiv.org/abs/1904.09675

Zhang, T., Ladhak, F., Durmus, E., Liang, P., McKeown, K., & Hashimoto, T. B. (2024). Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics, 12*, 39–57. https://doi.org/10.1162/tacl_a_00633

ÖZGEÇMİŞ

Eray Nurtekin İlkokul, ortaokul ve lise hayatına Lüleburgaz ilçesinde devam etti. 2016 yılında Trakya Üniversitesi Elektrik-Elektronik Mühendisliği Bölümü'nde lisans eğitimine başladı ve 2021 yılında mezun oldu. 2023 yılından beri Trakya Üniversitesi Veri Bilimi ve Analitiği Bölümü'nde yüksek lisans eğitimine devam etmektedir.