YTU Cosmos DPO Modeli üzerinde 10 Atışlı Bağlam İçi Öğrenme Kullanarak 3 Farklı Topluluk Yöntemlerinin Karşılaştırılması

Comparison of 3 Different Ensemble Methods Using 10 Shot In-Context Learning (ICL) on YTU Cosmos DPO Model

Metin Uslu
Data Science and Big Data
Yildiz Technical University
Istanbul, Turkey
metin.uslu@std.yildiz.edu.tr

***Özetçe* —Büyük dil modelleri çok farklı görevlerde genelleme yetenekleri sebebiyle iyi performanslar göstermektedirler. Bu çalışmada YTU Cosmos Turkish LLama DPO modeli üzerinde ARC, HellaSwag ve GSM8K Türkçe veri setleri üzerinde 10 atışlı bağlam içi öğrenme yöntemi kullanılarak 10 büyüklüğünde 3 farklı emsemble yöntemi kullanılarak veri setleri üzerinde sabit, tekrarlı, azaltılarak ve yerleri değiştirilerek olmak üzere model üzerinde performans etkileri ölçülmüştür. Topluluk yöntemleri ile performansın iyi olduğu gözlemlenmiştir.**

Anahtar Kelimeler—yapay zeka, büyük dil modelleri, bağlamda öğrenme, üretici modeller

***Abstract*—**

***Keywords*—**

I. GİRİŞ

Günümüzde büyük dil modelleri birçok farklı görevlerde kullanılmaktadır. Bu dil modelleri verilen görevleri kendileri yapabildiği gibi bağlam içi öğrenmesi ile örnek bağlamları anlayıp ardından sonraki gelecek soruları bunlara göre cevaplayabilmektedir. Bu çalışmada ilk olarak 3 veri seti üzerinde 10 atışlı bağlam içi öğrenme yöntemi kullanılarak her bir veri setinden alınan 400'er adet örnek için cevaplar alınmıştır. Ardından 3 farklı Topluluk Öğrenmesi yöntemi 10 tekrarlı olarak 400'er adet örnek için cevaplar alınmış ve topluluk performansları hesaplanmıştır.

II. VERİ SETLERİ

A. ARC Veriseti

ARC (AI2 Reasoning Challenge), Allen Institute for AI tarafından oluşturulmuş bir veri setidir ve doğal dil işleme (NLP) araştırmalarında yaygın olarak kullanılan önemli bir benchmark veri setidir. Özellikle, makine öğrenimi modellerinin akıl yürütme yeteneklerini ölçmek ve değerlendirmek için tasarlanmıştır.

ARC veri seti, genellikle ABD ilkokul ve ortaokul müfredatına dayalı çoktan seçmeli sorulardan oluşur. Sorular, doğal dil anlayışı ve akıl yürütme gerektirir. Veri seti iki ana bölüme ayrılır:

- Easy Set (Kolay Kısım): Basit bir bilgi alma yöntemiyle çözülebilecek sorular.
- Challenge Set (Zorlayıcı Kısım): Sadece bilgi alma değil, aynı zamanda çıkarım yapma ve karmaşık akıl yürütme gerektiren sorular.

ARC veri seti formatı aşağıdaki gibidir.

Başlangıç Cümlesi: Bir hikaye veya durum.

Seçenekler: 4 seçenek (A, B, C, D) içerir. Bunlar, başlangıç cümlesine mantıklı ve bağlama uygun bir şekilde tamamlanması gereken cümlelerdir.

Doğru Cevap: En mantıklı tamamlayıcı, doğru cevap olarak belirlenmiştir.

B. HellaSwag Veriseti

HellaSwag, doğal dil işleme (NLP) araştırmalarında kullanılan bir diğer önemli veri setidir. Bu veri seti, dil modelleme ve özellikle dilsel anlamlandırma ve akıl yürütme üzerine testler yapmayı amaçlayan bir koleksiyondur. HellaSwag, özellikle koşullu dil modelleme ve doğa dışı çıkarım yapabilme gibi becerileri ölçmek için tasarlanmıştır.

HellaSwag, gerçek dünyadan alınan metinler ve mantıklı sonlandırmalar arasında ilişki kurmaya dayalı çoktan seçmeli sorulardan oluşur. Veri seti, bir başlangıç durumu ve buna bağlı olarak verilmiş dört potansiyel sonlandırma seçeneğinden oluşur. Buradaki amaç, modellerin bir olayın mantıklı bir şekilde sonlanıp sonlanmayacağını değerlendirmesidir.

C. GSM8K Veriseti

GSM8K (Grade School Math 8K), matematiksel problem çözme becerilerini test etmek için kullanılan bir başka önemli veri setidir. Bu veri seti, özellikle ilkokul seviyesindeki matematiksel soruları çözebilen modelleri değerlendirmeyi amaçlar. GSM8K, makine öğrenimi ve doğal dil işleme (NLP) araştırmalarında matematiksel akıl yürütme ve hesaplama yeteneklerini test etmek için kullanılır. Veri seti formatı aşağıdaki gibidir.

Soru: Matematiksel bir problem, genellikle dilde yazılmıştır (örneğin, "Bir çiftlikte 15 inek ve 7 koyun var. Toplamda kaç bacak var?").

Çözüm Adımları: Soruya giden çözüm adımlarını açıklayan metin. Modelin çözümü, bu adımları doğru bir şekilde takip etmesini gerektirir.

Sonuç (Doğru Cevap): Sorunun doğru cevabı.

III. Dil Modeli: YTU Cosmos Turkish Llama DPO

Yıldız Teknik Üniversitesi (YTU) Cosmos araştırma grubu tarafından geliştirilen LLaMa(Large Language Model Meta AI) tabanlı Türkçe dil modelidir. Bu model, **Türkçe diline özel optimizasyonlarla** (DPO) geliştirilmiş olup, dil işleme ve anlamlandırma konularında performansı artırmak için tasarlanmıştır.

YTU Cosmos Turkish Llama DPO dil modeli 8 milyar parametreye sahip olup, modelin uygulama alanları; Türkçe Doğal Dil İşleme Uygulamaları, Soru-cevap, Metin Oluşturma gibi görevlerde kullanılabilmektedir.

IV. Bağlam İçi Öğrenme

Bağlam içi öğrenme, büyük dil modellerinin örnek bağlamları anlayıp ardından sonraki gelecek soruları bunlara göre cevaplanmasının istenmesidir. Bağlamda öğrenme ile dil modellerine sınırlı sayıda örnekler vererek yeni görevlerde performansı gözlemlenmektedir.

A. *Sıfır Atışlı Öğrenme*

Sıfır atışlı öğrenme, model hiç görmediği bir görev veya problem üzerinde tahmin yapar. Burada modelin eğitilirken öğrenilen bilgiler ile cevaplar vermesi olarak ifade edilir.

B. *Tek Atışlı Öğrenme*

Tek atışlı öğrenme, modelin yalnızca bir örneğin gösterilerek adapte olması ve ardından hiç görmediği sorulara cevaplar vermesi olarak ifade edilebilir.

C. *Az Atışlı Öğrenme*

Az atışlı öğrenme, modelin birkaç örnek ile öğrenip, adapte olup tahmin yapabilmesi olarak ifade edilebilir.

V. Topluluk Yöntemleri

Bu çalışmada 3 farklı Ensemble yöntemi kullanılmıştır. Ensemble büyüklükleri 10'dur. Çalışma sonunda her bir ensemble yöntemi ile performanslar ölçülmüştür. Bu çalışmanın başında her bir veri setinden 10'ar adet rasgele örnek seçilmiştir. Seçilen bu 10 örnek aşağıdaki ensemble yöntemlerinde kullanılmıştır.

A. *Ensemble Yöntem-1: 10 Tekrarlı 10 Örnekten 10'lu Seçimler(Yerine Koymalı)*

Bu yöntemde her bir veri kümesi içerisinden 10 kere tekrarlanmak üzere, 10 örnek içerisinden 10 örneğin tekrarlı bir şekilde seçilmesi işlemidir. Bu işlemden sonra seçilen 10 adet 10 Atışlı örneklem Büyük Dil Modeline Bağlam İçi Örnekleme yoluyla verilmiş olup ardından Test kümesinde bulunan 400 örnek içerisindeki her bir örneğin cevaplanması sağlanmıştır. 1 test örneği için 10 adet cevap alınmıştır. Ardından bu cevaplar birleştirilerek performans ölçülmüştür.

B. *Ensemble Yöntem-2: 10 Tekrarlı 10 Örnekten 5'li Seçimler*

Bu yöntemde her bir veri kümesi içerisinden 10 kere tekrarlanmak üzere, 10 örnek içerisinden 5 tane örneğin seçilmesi işlemidir. Bu işlemden sonra seçilen 10 adet 10 Atışlı örneklem Büyük Dil Modeline Bağlam İçi Örnekleme yoluyla verilmiş olup ardından Test kümesinde bulunan 400 örnek içerisindeki her bir örneğin cevaplanması sağlanmıştır. 1 test örneği için 10 adet cevap alınmıştır. Ardından bu cevaplar birleştirilerek performans ölçülmüştür.

C. *Ensemble Yöntem-3: 10 Tekrarlı 10 Örneğin Sıralarının Değiştirilmei*

Bu yöntemde her bir veri kümesi içerisinden 10 kere tekrarlanmak üzere, 10 örnek içerisinden 10 tane örneğin sıralarının değiştirilerek seçilmesi işlemidir. Bu işlemden sonra seçilen 10 adet 10 Atışlı örneklem Büyük Dil Modeline Bağlam İçi Örnekleme yoluyla verilmiş olup ardından Test kümesinde bulunan 400 örnek içerisindeki her bir örneğin cevaplanması sağlanmıştır. 1 test örneği için 10 adet cevap alınmıştır. Ardından bu cevaplar birleştirilerek performans ölçülmüştür.

VI. Sonuçlar

Arc, HellaSwag ve GSM8K veri setleri için 10 Atış ve 3 farklı ensemble yöntemi için performans sonuçları Tablo I, Tablo II ve Tablo III'de paylaşılmıştır.

Bu çalışmaya ait tüm kaynaklara https://github.com/metinuslu/blm5109_cl_capstone_project buradan erişim sağlayabilirsiniz.

TABLO I: ARC Veri Seti Performans

ARC Veri Seti						
	Atış B.	Tekrarlı Seçim	Doğru	Yanlış	Test B.	Başarım
10 Shot	10	Yok	235	165	400	58,75%
Ensemble-A (Size=10)	10	Var	243	157	400	60,75%
Ensemble-B (Size=10)	5	Yok	235	165	400	58,75%
Ensemble-C (Size=10)	10	Yok	218	182	400	54,50%

TABLO II: HellaSwag Veri Seti Performans

HellaSwag Veri Seti						
	Atış B.	Tekrarlı Seçim	Doğru	Yanlış	Test B.	Başarım
10 Shot	10	Yok	175	225	400	43,75%
Ensemble-A (Size=10)	10	Var	189	211	400	47,25%
Ensemble-B (Size=10)	5	Yok	201	199	400	50,25%
Ensemble-C (Size=10)	10	Yok	203	197	400	50,75%

TABLO III: GSM8K Veri Seti Performans

	GSM8K					
	Atış B.	Tekrarlı Seçim	Doğru	Yanlış	Test D.	Başarım
10 Shot	10	Yok	29	371	400	7,25%
Ensemble-A (Size=10)	10	Var	28	372	400	7,00%
Ensemble-B (Size=10)	5	Yok	27	373	400	6,75%
Ensemble-C (Size=10)	10	Yok	28	372	400	7,00%

VII. KAYNAKLAR

[1] Türkçe Dil Modellerinin Performans Karşılaştırılması, https://arxiv.org/pdf/2404.17010

[2] https://huggingface.co/datasets/malhajar/arc-tr

[3] https://huggingface.co/datasets/malhajar/hellaswag-tr

[4] https://huggingface.co/datasets/malhajar/gsm8k-tr

[5] https://github.com/NirDiamant/Prompt_Engineering

[6] https://huggingface.co/ytu-ce-cosmos/Turkish-Llama-8b-DPO-v0.1

[7] https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard#/

[8] https://huggingface.co/spaces/lmarena-ai/chatbot-arena-leaderboard

[9] https://www.upstage.ai/blog/en/llm-evaluation-part1-benchmark-datasets