Kolektif Öğrenme 2024/1
Proje Önerileri

Aşağıdaki projelerden 3 tanesini seçip öncelik sıralarıyla birlikte (1-projenin adı 2-projenin adı, 3-projenin adı) **24 Aralık 2024** tarihine kadar dersin yürütücüsüne e-mail atın
ya da
kendi proje teklifinizi 1 sayfayı aşmayacak şekilde hazırlayıp dersin yürütücüsünden **24 Aralık 2024** tarihine kadar onay alın.
Not: Kendi proje önerinizde hazır veri kümeleri üzerinde hazır ensemble algoritmalarını karşılaştırmak kabul edilMEyecektir.

24 Aralığa kadar onay almayanlar dersin projesini yapa**ma**yacaklardır.
Bütünleme için yine projeler verilecektir, ancak süre mecburen daha kısa olacaktır.

Kabul süreçleri zaman alabildiğinden hızla harekete geçin ☺

Projelerde herkes tek başına çalışacaktır.

Proje Konuları:

1-R1: https://huggingface.co/datasets/Metin/WikiRAG-TR veri kümesi üzerinde her soru için 5 (1 i doğru) chunk var. Tüm chunkları (5*1000=5000) alıp bir vector db ye kaydet.
Retrieval başarısı için dense vector rep (en az 3 adet- e5, jina vb.), word ve BERT_tokenized word matching (BM25 vb.) karşılaştırma (5 temsil yöntemi), birleşimlerini (nasılı size kalmış) karşılaştırma. 1000 soru üzerinde retrieval top 1 ve top 5 başarılarının ölçülmesi.

2-R2: https://huggingface.co/datasets/Metin/WikiRAG-TR rag veri kümesi üzerinde
A) Doğru chunk ı içeren context uzadıkça (1,5,10,15) llm (cosmos dpo ve gemma2 9b it) cevaplama başarısı (doğru chunk ortada), En az 100 soru üzerinde RAGAS evaluation.
B) Doğru chunk in 15 chunk tan oluşan context teki konumunun (1:1:15) llm cevaplama başarısına etkisi, En az 50 soru üzerinde RAGAS evaluation.

3-RAG: 5 paragraftan yararlanarak cevaplanabilecek Türkçe sorular oluşturulması (en az 50 adet, sorunun cevabı 5 paragraftaki verileri de kullanmayı gerektirmelidir) ve mevcut en az 5 açık erişimli Türkçe llm in perf. ölçümü RAGAS evaluation

4-RTR: e5 ya da jina nın eğitimine uygun Türkçe verilerle devam edip Türkçe retrieval performansını iyileştirme. https://huggingface.co/datasets/Metin/WikiRAG-TR veri kümesi üzerinde her soru için 5 (1 i doğru) chunk var. Tüm chunkları (5*1000=5000) alıp bir vector db ye kaydet.
Retrieval başarısı için orijinal ve eğitilmiş modeli karşılaştırma. 1000 soru üzerinde retrieval top 1 ve top 5 başarılarını ölçülmesi.

5-MG: YSA çözümlerini birleştirmek: Farklı ilk değerlerle üretilen YSA çözümleri (en az 5 tekil öğrenici) arasında test başarısını düşürmeden yol bulmak, bu yolları görselleştirmek, bu yollar üzerindeki noktaları kullanarak ensemble oluşturmak, sadece son noktalardan oluşturulan ensemble ile karşılaştırmak, 1 metin ya da 1 görüntü veri kümesin üzerinde.

6-LLM lerle augmetation: 2 metin veri kümesi üzerinde bir ML modelini (RF, SVM, YSA vb.) temsiller ile (e5, jina vb. temsiller ile) eğit. Kullanılacak veri kümelerini belirlerken orijinal eğitim kümesi artışının test performansını anlamlı derecede yükselttiği veri kümeleri seçiniz, raporunuzda bu durumu grafikle gösteriniz. En az 3 llm in aşağıdaki augmentation süreçlerine katkılarını incele.
A) Test örneklerini augmente et llm ile. Orj test örneğinin kararını belirlerken modelin orj örnek ve aug örnekler için verdiği kararları birleştir. Sadece orj örnek için verilen kararla karşılaştır. aug sayısının 3 ve 5 değerleri için performans karşılaştır.
B) Eğitim örneklerini augment et llm ile. Temsiller ile ML modeli eğit. Eğitim kümesini büyültmenin etkisini incele. Eğitim kümesinin 2, 3, 5 katına çıkışının performansa etkisini incele.

7-LLM + aktif öğrenme: sınıf olasılığı farkı az olanlar için gerçek etiketleri, çok olanların etiketlerini llm den al. Hepsinde gerçek etiket kullanma, hepsinde LLM etiketlerini kullanma ve önerilen yaklaşımı karşılaştır.
orj eğitim kümesinin %20'si elde. sonra %10'luk artışlarla(%5 i gerçek, %5 i llm den) %100 ye kadar çıkılacak. Etiketli verilerin modellenmesinde temsiller ile ML modelleri kullanılacak. En az 2 netin veri kümesi üzerinde. Kullanılacak veri kümelerini belirlerken orijinal eğitim kümesi artışının test performansını anlamlı derecede yükselttiği veri kümeleri seçiniz, raporunuzda bu durumu grafikle gösteriniz.

8-Improved space i metin temsil yöntemleri üzerinde uygulama. 2 temsil (e5, jina vb.) al. concatenate. bunlardan imp space ile yeni boyutlar elde et. tek başlarına, birleşmiş halleri, yeni boyut eklenmiş hallerini

karşılaştır. en az 2 metin veri kümesi üzerinde ML modelleriyle eğitim ve test. imp space boyutunun orj boyutun 1.5, 2, 3, 5 katına çıkışını karşılaştır. Birleştirmede eğitimdeki dönüşümleri belirleyip (indisler ve katsayıları) test örneklerini bunlarla dönüştürmelisiniz.

9-CNL- In context learning'de ensemble: cosmos-DPO île 3 veri kümesi üzerinde (ARC, Hellaswag, GMSK8K) 10 shot sonuçlarını (400 örnek üzerinde) elde edip raporlayın. Aşağıdaki 3 yöntemle ensemble lar (ensemble size=10) üretip karşılaştırın.
A-10 örnekten 10'lu seçimler
B-10 örnekten 5'li seçimler
C-10 örneğin sıralarını değiştirmek
ensemble içindeki tekil öğrenicilerin her birinin performansını ve ensemble performansını raporlayın.
Karşılaştırma için araç: https://github.com/EleutherAI/lm-evaluation-harness

Teslim edilecekler:
1-Proje raporu: ieee konferans formatında hazırlanmış pdf
https://www.ieee.org/conferences/publishing/templates.html **(yayına hazırlık)**
raporun içinde kod linki (bir github sayfasına)

2-Sunum
Proje sunumları final tarih ve saatinde başlayacaktır.
Bu tarihe kadar online.yildiz.edu.tr ye yükleme yapabilirsiniz.

Notlandırma: Proje notu 100 üzerinden (%50 sunum, %50 rapor)
Sınıfta sunum yapmayanlar projeden notlarının en fazla %30'unu alırlar.