Teknoloji Fakültesi

Büyük Dil Modellerinde Yönerge Kullanımı Optimizasyonu ve Maliyet Etkin Sonuç Alma Yöntemleri Karşılaştırması.

BİTİRME PROJESİ

1. Ara Raporu

Bilgisayar Mühendisliği Bölümü

BİLGİSAYAR MÜHENDİSLİĞİ B

DANIŞMAN

Doç.Dr. Öğr. Üyesi Buket Doğan

MARMARA ÜNİVERSİTESİ
TEKNOLOJİ FAKÜLTESİ
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

Marmara Üniversitesi Teknoloji Fakültesi Bilgisayar Mühendisliği Öğrencileri Yusuf Ünlü ve Burak Yalçın tarafından "**Büyük Dil Modellerinde Yönerge Kullanımı Optimizasyonu ve Maliyet Etkin Sonuç Alma Yöntemleri Karşılaştırması.**" başlıklı proje çalışması, xxx tarihinde savunulmuş ve jüri üyeleri tarafından başarılı bulunmuştur.

Jüri Üyeleri

Dr. Öğr. Üyesi xxx xxx Marmara Üniversitesi	**(Danışman)**	(İMZA)……..
Prof. Dr. Xxx xxx Marmara Üniversitesi	(Üye)	(İMZA)……..
Prof. Dr. Xxx xxx Marmara Üniversitesi	(Üye)	(İMZA)……..

ÖNSÖZ

Proje çalışmamız süresince karşılaştığım bütün problemlerde, sabırla yardım ve bilgilerini esirgemeyen, tüm desteğini sonuna kadar yanımda hissettiğim değerli hocalarım, sayın Dr. Öğr. Üyesi Xxx xxx ve sayın Prof. Dr. Xxx xxx' a en içten teşekkürlerimi sunarım.

Bu proje çalışması fikrinin oluşması ve ortaya çıkmasındaki önerisi ve desteğinden dolayı değerli hocam Dr. Öğr. Üyesi Xxx xxx' a teşekkür ederim.

Proje çalışmam sırasında maddi ve manevi desteklerini esirgemeyen okul içerisinde ve okul dışında her zaman yanımda olan değerli çalışma arkadaşlarım ve hocalarım Doç. Dr. Xxx xxx ve Dr. Öğr. Üyesi ' xxx xxx a sonsuz teşekkürlerimi sunarım.

İÇİNDEKİLER

1. GİRİŞ **1**

1.1. Proje Çalışmasının Amacı ve Önemi **1**

2. BÜYÜK DİL MODELLERİ ve PROMPT MÜHENDİSLİĞİ **3**

2.1. Prompt Mühendisliği **3**

2.2. Büyük Dil Modellerinde Optimizasyon Teknikleri **4**

3. BULGULAR VE TARTIŞMA **6**

4. SONUÇLAR **7**

- **9**

ÖZET

Büyük Dil Modellerinde Yönerge Kullanımı Optimizasyonu ve Maliyet Etkin Sonuç Alma Yöntemleri Karşılaştırması

Araştırmalar devam ediyor.

Haziran, 2025 **Öğrenciler**

ABSTRACT

Optimization of Prompt Usage in Large Language Models and Comparison of Cost-Effective Retrieval Methods

Context will be added soon.

June, 2024 **Students**

1. GİRİŞ

Büyük dil modelleri, doğal dil işleme alanında devrim yaratarak, çeşitli uygulamalarda insan benzeri metin üretimi, çeviri, özetleme ve metin anlama gibi görevlerde büyük başarılar elde etmiştir. Ancak, bu modellerin yüksek işlem gücü ve büyük miktarda veri gereksinimi, kullanım maliyetlerini artırmakta ve ölçeklenebilirlik konusunda zorluklar yaratmaktadır. Özellikle, büyük dil modellerine verilen girdilerin (promptların) uzunluğu ve yapısı, modelin ürettiği çıktının doğruluğunu, tutarlılığını ve maliyetini doğrudan etkilemektedir. Bu nedenle, yönlendirici girdilerin (promptların) etkin kullanımı, hem performans optimizasyonu hem de maliyet açısından büyük önem taşımaktadır.

Bu çalışmada, büyük dil modellerinde yönerge kullanımını optimize etmek amacıyla farklı prompt sıkıştırma ve özetleme teknikleri incelenerek, bunların çeşitli modeller üzerindeki performansları ve maliyet etkinlikleri karşılaştırılacaktır. Mevcut araştırmalar, model çıktılarının kalitesini koruyarak veya artırarak, daha az kaynak kullanımıyla optimal sonuçlar elde etmenin mümkün olduğunu göstermektedir. Ancak, hangi tekniklerin hangi büyük dil modellerinde daha iyi performans gösterdiği ve maliyet açısından daha verimli olduğu konusu halen açık bir araştırma alanıdır.

Günümüzde, büyük dil modellerinin bulut tabanlı hizmetler olarak yaygınlaşması, bireysel ve kurumsal kullanıcılar için maliyet etkin çözümler üretme ihtiyacını ortaya çıkarmaktadır. Model çağrılarında kullanılan girdilerin optimize edilmesi, hem hesaplama maliyetlerini azaltmak hem de çıktının doğruluğunu ve güvenilirliğini artırmak için kritik bir adımdır. Bu bağlamda, proje kapsamında yapılan analizler, büyük dil modellerinin daha verimli kullanımına yönelik önemli katkılar sunmayı amaçlamaktadır.

1.1. Proje Çalışmasının Amacı ve Önemi

Bu projenin temel amacı, büyük dil modellerinde etkili prompt sıkıştırma ve özetleme yöntemlerini araştırarak, bu yöntemlerin farklı modellerdeki maliyet etkinliklerini karşılaştırmaktır. Doğru tasarlanmış bir yönlendirme metni (prompt), modelin daha iyi yanıtlar üretmesini sağlarken gereksiz uzunluktaki girdiler, işlem maliyetlerini artırarak

gereksiz kaynak kullanımına yol açmaktadır. Bu nedenle, optimize edilmiş prompt kullanımı, hem büyük dil modellerinin performansını artırmak hem de maliyetleri düşürmek için kritik bir konudur. Bu çalışma, farklı sıkıştırma ve özetleme yöntemlerini analiz ederek, büyük dil modellerinde hangi stratejilerin en iyi sonuçları verdiğini belirlemeyi hedeflemektedir. Yapılan karşılaştırmalar sayesinde, belirli bir modelde optimum maliyet-performans dengesini sağlayan en iyi teknikler ortaya konulacaktır. Ayrıca, büyük dil modellerinin akademik ve endüstriyel kullanımlarında daha verimli hale getirilmesine katkı sağlanacaktır.

2. BÜYÜK DİL MODELLERİ ve PROMPT MÜHENDİSLİĞİ

Büyük dil modelleri (LLM'ler), geniş çaplı veri setleriyle eğitilmiş, metin tabanlı görevleri gerçekleştirme yeteneğine sahip derin öğrenme tabanlı yapay zeka sistemleridir. Bu modeller, doğal dil işleme (NLP) alanında önemli bir devrim yaratarak, dil anlama, çeviri, özetleme, kod üretme ve metin oluşturma gibi birçok farklı görevi yüksek doğruluk oranlarıyla gerçekleştirebilmektedir. Büyük dil modelleri, genellikle milyonlarca hatta milyarlarca parametreye sahip olup, devasa veri kümeleriyle eğitildikleri için geniş bağlamları kavrayabilir ve tutarlı yanıtlar üretebilirler.

Büyük dil modellerinin en bilinen örnekleri arasında OpenAI tarafından geliştirilen GPT serisi, Google'ın Bard ve Gemini modelleri, Meta'nın LLaMA modeli ve Anthropic'in Claude serisi bulunmaktadır. Bu modellerin her biri farklı mimari ve eğitim yaklaşımlarına sahip olsa da temel amaçları, kullanıcı girdilerini (promptları) analiz ederek anlamlı ve bağlama uygun yanıtlar üretmektir.

Bununla birlikte, büyük dil modellerinin kullanımında karşılaşılan en büyük zorluklardan biri, yüksek hesaplama maliyetleridir. Bu modellerin çalıştırılması, büyük miktarda işlem gücü ve bellek gerektirdiğinden, model sorgularının optimizasyonu büyük önem taşımaktadır. Özellikle, gereksiz uzunluktaki girdiler veya yanlış yapılandırılmış promptlar, hem yanıt süresini uzatmakta hem de gereksiz maliyet artışına yol açmaktadır. Bu nedenle, büyük dil modellerinin etkili bir şekilde kullanılması için optimize edilmiş giriş metinlerinin oluşturulması gerekmektedir.

2.1. Prompt Mühendisliği

Prompt mühendisliği, büyük dil modellerinin istenen çıktıları üretmesini sağlamak için yönlendirme metinlerinin (promptların) optimize edilmesi sürecidir. Bu süreç, modelin daha doğru, tutarlı ve maliyet açısından verimli yanıtlar üretmesi için giriş metinlerinin dikkatlice tasarlanmasını içerir. Doğru yapılandırılmış bir prompt, modelin istenen bilgiyi üretmesini kolaylaştırırken, kötü tasarlanmış bir prompt gereksiz uzunluk ve belirsizlik nedeniyle modelin beklenenden farklı veya düşük kaliteli yanıtlar vermesine yol açabilir.

Prompt mühendisliği, temel olarak iki ana stratejiye dayanır:

Yönerge (instruction) optimizasyonu: Modelin daha verimli yanıtlar verebilmesi için açık ve net talimatlar verilmesi gerekmektedir. Örneğin, "Bu metni özetle" gibi basit bir yönerge yerine, "Bu metni 50 kelimeyi geçmeyecek şekilde özetleyerek ana fikirleri koru" gibi daha ayrıntılı bir talimat verilmesi, modelin daha doğru yanıtlar üretmesini sağlayabilir.

Prompt sıkıştırma ve özetleme: Büyük dil modelleri, genellikle giriş uzunluğuna bağlı olarak daha fazla kaynak tüketir. Bu nedenle, girdilerin gereksiz kelimelerden arındırılması ve en verimli şekilde yapılandırılması, maliyet etkinliği açısından büyük önem taşımaktadır. Örneğin, uzun bir metni olduğu gibi modele sunmak yerine, ön işleme aşamasında özetleyerek modelin işleyebileceği kompakt bir hale getirmek, performans optimizasyonu açısından faydalı olabilir.

Prompt mühendisliği, büyük dil modellerinin daha az kaynak tüketerek daha iyi sonuçlar üretmesine yardımcı olmanın yanı sıra, modelin yanıtlarının doğruluk ve güvenilirlik oranını da artırmaktadır. Özellikle ticari ve akademik kullanımlarda, model çağrı maliyetlerini minimize etmek ve işlem süresini optimize etmek için prompt mühendisliği stratejileri büyük önem taşımaktadır.

Bu çalışmada, farklı prompt mühendisliği teknikleri analiz edilerek, büyük dil modelleriyle etkileşimde en yüksek verimliliği sağlayan yöntemler karşılaştırılacaktır. Amaç, hangi optimizasyon stratejilerinin farklı modellerde daha iyi çalıştığını belirlemek ve böylece büyük dil modellerinin kullanımını hem maliyet hem de performans açısından daha verimli hale getirmektir.

2.2. Büyük Dil Modellerinde Optimizasyon Teknikleri

Büyük dil modellerinin yüksek hesaplama maliyetleri ve işlem süreleri, verimli kullanım için optimizasyon gerektirir. Bu süreçte model ince ayarı (fine-tuning), ön işleme, prompt mühendisliği, yanıt üretim stratejileri, önbellekleme ve model seçimi gibi teknikler uygulanır. Büyük dil modelleri (LLM'ler), doğal dil işleme (NLP) alanında güçlü yeteneklere sahip olsa da, yüksek işlem maliyetleri ve performans sınırlamaları nedeniyle dikkatli bir şekilde optimize edilmelidir. Bu optimizasyon süreci, hem modelin

yanıt kalitesini artırmayı hem de hesaplama maliyetlerini düşürmeyi hedefler. Büyük dil modellerinin verimli kullanımı için model ince ayarı (fine-tuning), kapsamlı ön işleme, prompt mühendisliği, yanıt üretim stratejileri, önbellekleme ve model seçimi gibi çeşitli teknikler uygulanmaktadır. Model ince ayarı (fine-tuning), büyük dil modellerinin belirli bir alan veya görev için daha iyi performans göstermesini sağlamak amacıyla önceden eğitilmiş modellerin ek veriyle yeniden eğitilmesi işlemidir. Fine-tuning sayesinde, genel amaçlı dil modelleri, özel sektör uygulamaları veya araştırmalar için daha hassas ve alakalı çıktılar üretebilir. Ancak, bu süreç büyük veri ve işlem gücü gerektirdiğinden her senaryoda uygulanması pratik olmayabilir. Alternatif olarak, hafifletilmiş modeller (distilled models) veya belirli görevler için optimize edilmiş daha küçük ölçekli modeller tercih edilebilir.

Büyük dil modellerinin yanıt üretiminde girdi optimizasyonu önemli bir faktördür. Girdilerin gereksiz kelimelerden ve tekrar eden bilgilerden arındırılması, modelin daha hızlı ve maliyet etkin çalışmasını sağlar. Kapsamlı ön işleme ve prompt mühendisliği sayesinde, modelin verilen girdiyi daha verimli şekilde işlemesi sağlanabilir. Metin özetleme teknikleriyle uzun girdiler kısaltılabilirken, açık ve net yönergeler içeren prompt formatları modelin istenen çıktıyı daha doğru şekilde üretmesine yardımcı olur.

Yanıt üretim sürecinde temperature parametresi, top-k ve top-p sampling gibi stratejiler, modelin çıktılarını optimize etmek için kullanılır. Temperature değeri, modelin yanıtlarının ne kadar rastgele veya kesin olacağını belirlerken, top-k ve top-p sampling yöntemleri gereksiz kelimelerin seçilmesini engelleyerek daha anlamlı çıktılar üretmeye yardımcı olur. Ayrıca, modelin çıktılarının belirli bir uzunlukta tutulması, gereksiz işlem maliyetlerinden kaçınmak için etkili bir tekniktir.

Daha düşük maliyetli ve hızlı yanıtlar elde etmek için önbellekleme (caching) teknikleri uygulanabilir. Önbellekleme, sık kullanılan yanıtların veya sorguların saklanarak tekrar işlenmesine gerek kalmadan hızlıca geri döndürülmesini sağlar. Statik yanıtlar için önceden hazırlanmış sonuçlar kullanılabilirken, dinamik önbellekleme stratejileriyle sıkça sorulan soruların yanıtları belirli bir süre boyunca güncellenebilir. Bu yöntem, özellikle müşteri destek botları, bilgi tabanları ve sık tekrar eden metin üretim süreçleri için büyük avantaj sağlar. Doğru model seçimi, maliyet ve performans dengesi açısından kritik bir rol oynar. Örneğin, GPT-4 gibi büyük modeller, yüksek doğruluk gerektiren

görevler için kullanılırken, daha basit ve maliyet etkin alternatifler olan GPT-3.5, DistilBERT veya T5 gibi modeller belirli görevler için yeterli performans sunabilir. Doğru modelin seçilmesi, hem işletim maliyetlerini düşürmeye hem de işlem hızını artırmaya yardımcı olur.

3. BULGULAR VE TARTIŞMA

Büyük dil modellerinin etkin kullanımı için optimizasyon tekniklerinin bilinçli bir şekilde uygulanması gerekmektedir. Fine-tuning, ön işleme, prompt mühendisliği, yanıt üretim stratejileri, önbellekleme ve model seçimi gibi yöntemler, bu sistemlerin daha verimli ve ekonomik hale getirilmesini sağlar. Bu çalışma kapsamında, bu tekniklerin farklı büyük dil modellerinde nasıl uygulanabileceği ve hangi yöntemlerin en iyi maliyet-performans dengesini sunduğu analiz edilecektir.

Bu çalışma kapsamında, büyük dil modellerinin (LLM'ler) verimli kullanımı için optimizasyon stratejileri üzerine detaylı bir analiz gerçekleştirdik. Öncelikle, büyük dil modellerinin genel yapısını inceledik ve bu modellerin yüksek işlem maliyetleri ile performans gereksinimlerini ele aldık. Daha sonra, model ince ayarı (fine-tuning), kapsamlı ön işleme, prompt mühendisliği, yanıt üretim stratejileri, önbellekleme ve model seçimi gibi çeşitli optimizasyon tekniklerini değerlendirerek, bu yöntemlerin büyük dil modelleri üzerindeki etkilerini teorik olarak tartıştık.

Bundan sonraki aşamada, belirlenen optimizasyon yöntemlerinin farklı büyük dil modelleri üzerindeki etkilerini karşılaştırmalı olarak test etmeye odaklanacağız.. Ayrıca bu tekniklerin yanında bu süreçte arxiv.org'dan eriştiğimiz güncel makaleler vasıtasıyla yeni teknikleri de takip edeceğiz ve uygun görülmesi halinde projeye dahil edilmesini değerlendireceğiz.

4. SONUÇLAR

•

KAYNAKLAR