Journal of the Faculty of Engineering and Architecture of Gazi University 40:3 (2025) 1787-1796

Mühendislik Mimarlık Fakültesi Dergisi
Journal of The Faculty of Engineering and Architecture of Gazi University

Elektronik / Online ISSN : 1304 - 4915
Basılı / Printed ISSN : 1300 - 1884

Large-scale impact analysis on large language models for Turkish question-answering

Zekeriya Anıl Güven*
Department of Computer Engineering, Faculty of Engineering and Architecture, Izmir Bakircay University, Izmir, 35665, Türkiye

Highlights:

- Pioneering study for Turkish QA task
- The positive effect demonstration of semantic similarity for QA
- Contribution to literature with the new fine-tuned LLMs

Keywords:

- Large language model
- SQuAD
- Text generation
- Question answering
- Semantic similarity

Article Info:

Research Article
Received: 24.08.2024
Accepted:01.02.2025

DOI:

10.17341/gazimmfd.1538022

Correspondence:

Author: Zekeriya Anıl Güven
e-mail: zekeriyaanil.guven@bakircay.edu.tr
phone: +90 506 562 3572

Graphical/Tabular Abstract

Large language models (LLMs) have recently become popular in many natural language processing tasks. There are fewer studies on LLMs in low-level languages such as Turkish. Therefore, the success of BERT, ALBERT, DistilBERT, mDeBERTa, and mT5 LLMs was analyzed for the Turkish question-answering task. The Turkish version of the benchmark dataset, SQuAD, was used. As a result of training these LLMs by fine-tuning, mDeBERTa became the most successful model with 74.50% accuracy. In addition, the effect of the threshold value of the answer probability of these models and the semantic similarity between the predicted and actual answers of the LLMs were examined. This methodology was summarized in Figure A. Experiments show that best accuracies' table of Figure A have also shown the positive impact of these analysis.

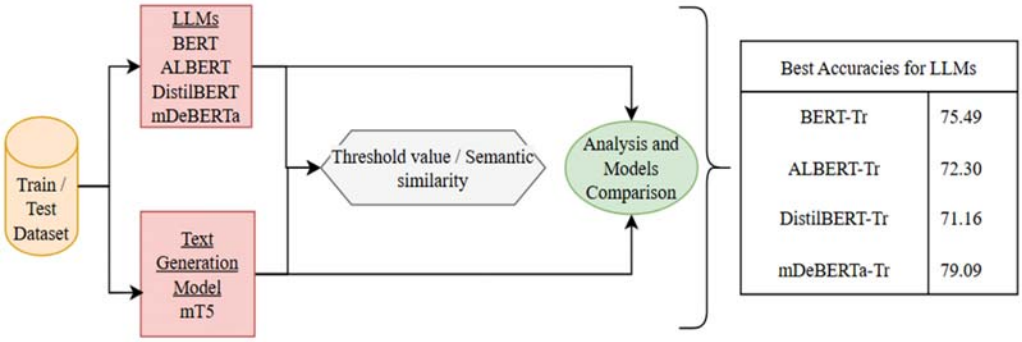


Figure A. Methodology and analysis results of this study

Purpose: It is a pioneering study comparing LLMs and text generation models for Turkish QA tasks. Contribution to the literature was made by training LLMs in Turkish QA tasks. It also analyzed which question types the models can answer better for Turkish. The effect of analyzing the predicted answers according to the threshold value and the semantic similarity between the actual and predicted answers.

Materials and Methods: Turkish SQuAD was chosen as the dataset. The original English SQuAD2.0 dataset, including titles, context paragraphs, questions, and answers, was translated into Turkish using Amazon Translate. Pre-trained BERT, ALBERT, DistilBERT, and mDeBERTa LLMs are trained and tested by finetuning them for the Turkish QA task. In addition, the text generation model, which is multilingual T5, was trained with Turkish SQuAD through the finetuning process. Additionally, a Turkish LLM was used to measure the similarity between the actual and predicted answers.

Results: BERT-Tr, DistilBERT-Tr, ALBERT-Tr, and mDeBERTa-Tr LLMs and mT5 were finetuned and analyzed on the test dataset. Analysis showed that the most successful model was mDeBERTa-Tr with 74.50% accuracy. In the next stage, the effect of the threshold value was investigated on these models when the answer's probability of the questions answered by the LLMs was examined. As a result of the analysis, there was an increase in the accuracy value of up to 0.13% in other LLMs except the BERT-Tr. In the last stage, the semantic relationship of the predicted answers produced by the models with the actual answers was examined. The value between the actual and predicted answers was analyzed for this task using the semantic similarity model. As a result of this analysis, it was observed that the accuracy value of all models increased between 0.7% and 6.59% and the most successful model was mDeBERTa-Tr with 79.09%.

Conclusion: BERT-Tr, DistilBERT-Tr, ALBERT-Tr, and mDeBERTa-Tr LLMs and mT5 were used for Turkish QA task. Many variations like threshold value, semantic similarity were performed on these models. Analysis shows that semantic similarity has a positive effect on these models and the most successful model is mDeBERTa.

Journal of the Faculty of Engineering and Architecture of Gazi University 40:3 (2025) 1787-1796

Mühendislik Mimarlık Fakültesi Dergisi
Journal of The Faculty of Engineering and Architecture of Gazi University

Elektronik / Online ISSN : 1304 - 4915
Basılı / Printed ISSN : 1300 - 1884

Türkçe soru cevaplama için büyük dil modelleri üzerinde geniş ölçekli etki analizi

Zekeriya Anıl Güven*
İzmir Bakırçay Üniversitesi, Mühendislik ve Mimarlık Fakültesi, Bilgisayar Mühendisliği Bölümü, 35665, İzmir, Türkiye

ÖNEÇIKANLAR

- Büyük dil modelleri ile beraber Türkçe soru cevaplama görevi için öncü bir çalışma
- Soru cevaplama için anlamsal benzerliğin olumlu etkisi gösterimi
- Yeni ince ayar yapılmış büyük dil modelleri ile beraber literatüre katkı sağlama

Makale Bilgileri
Araştırma Makalesi
Geliş: 24.08.2024
Kabul: 01.02.2025

DOI:
10.17341/gazimmfd.1538022

Anahtar Kelimeler:
Büyük dil modeli,
SQuAD,
metin üretme,
soru cevaplama,
anlamsal benzerlik

ÖZ

Son zamanlarda, büyük dil modelleri (LLM) birçok doğal dil işleme görevinde oldukça popüler hale gelmiştir. Türkçe gibi düşük seviyeli dillerde kullanımın yaygınlaşması açısından LLM'lere katkıda bulunmak oldukça önemlidir. Bu nedenle çalışmada, Türkçe soru-cevap görevi için BERT, ALBERT, DistilBERT, mDeBERTa ve mT5 LLM'lerinin başarısı analiz edilmiştir. SQuAD veri kümesinin Türkçe versiyonu veri seti olarak kullanılmıştır. Bu LLM'lerin ince ayar yapılarak eğitilmesi sonucunda, mDeBERTa %74.50 doğruluk ile en başarılı model olmuştur. Ayrıca, bu modellerin tahmin ettiği cevapların olasılığına eşik değerinin etkisi ve LLM'lerin tahmin edilen ve gerçek cevapları arasındaki anlamsal benzerliğin etkisi araştırılmıştır. Eşik değerinin etkisi analiz edildiğinde, LLM'lerin doğruluk değerinde %0.13'e kadar bir doğruluk artışı gözlenmiştir. Anlamsal benzerliğin LLM'ler üzerindeki etkisi analiz edildiğinde ise doğruluk değerinin %0.7 ile %6.59 arasında artış sağlanırken, en başarılı model %79.09 ile mDeBERTa olmuştur. Sonuç olarak, LLM'ler için eşik değeri ve anlamsal benzerliğin analiz edilmesinin olumlu bir etkiye sahip olduğunu göstermektedir.

Large-scale impact analysis on large language models for Turkish question-answering

HIGHLIGHTS

- Pioneering study with LLMs for Turkish QA task
- The positive effect demonstration of semantic similarity for QA
- Contribution to literature with the new fine-tuned LLMs

Article Info
Research Article
Received: 24.08.2024
Accepted:01.02.2025

DOI:
10.17341/gazimmfd.1538022

Keywords:
Large language model,
SQuAD,
text generation,
question answering,
semantic similarity

ABSTRACT

Large language models (LLMs) have recently become popular in many natural language processing tasks. It is very important to contribute to LLMs in order to increase the use of low-level languages such as Turkish. Therefore, the success of BERT, ALBERT, DistilBERT, mDeBERTa, and mT5 LLMs was analyzed for the Turkish question-answering task in this study. The Turkish version of the benchmark dataset, SQuAD, was used as the dataset. As a result of training these LLMs by fine-tuning, mDeBERTa became the most successful model with 74.50% accuracy. In addition, the effect of the threshold value of the predicted answer probability of these models and the semantic similarity between the predicted and actual answers of the LLMs were examined. When the effect of the threshold value was analyzed, an accuracy increase of up to 0.13% was observed in the accuracy value of LLMs. Analyzing the effect of semantic similarity on LLMs showed that the accuracy value increased between 0.7% and 6.59% and the most successful model was mDeBERTa with 79.09%. The results show that analyzing LLMs' threshold value and semantic similarity had a positive effect.

*Sorumlu Yazar/Yazarlar / *Corresponding Author/Authors: *zekeriyaanil.guven@bakircay.edu.tr, anilguven1055@gmail.com / Tel: +90 506 562 3572

1. Giriş (Introduction)

Sosyal medya platformları, haber ve alışveriş siteleri günümüzde önemli bir konuma sahiptir. Bu platformlar ve sitelerdeki birçok kullanıcının verilerinden yararlanarak doğal dil işleme (NLP) teknikleri ile öngörüler veya yorumlar elde edebilmektedir. NLP, metinleri işleyen ve metinlerin bilgisayar tarafından anlaşılmasını sağlayan yapay zeka alt alanıdır. NLP'nin Doğal Dil Anlama ve Doğal Dil Oluşturma olmak üzere iki ana alt görevi bulunmaktadır. Doğal dil anlama, algoritmalara dayanarak bilgisayarların insan dilini ve insan duygularını anlamasına yardımcı olmaktadır ve duygu analizi, makine çevirisi, sohbet robotu ve belge analizi görevlerini içermektedir. Doğal dil oluşturma ise soru cevaplama, metin özetleme ve metin oluşturma dahil olmak üzere metin içeriği oluşturmaya odaklanmaktadır [1].

Soru Cevaplama (QA), doğal dildeki sorulara doğru cevaplar vermeyi amaçlayan bir NLP görevi olup insan-makine etkileşimine yardımcı olmaktadır. Bir QA sisteminin doğal dil metnini, dil bilimini ve paylaşılan bilgiyi anlaması gerekmektedir [2]. QA sistemleri iki ana kategoriye ayrılmaktadır: açık alan ve kapalı alan. Kapalı alan QA sistemi, eğitim, sağlık ve spor gibi sınırlı alanlardaki soruları cevaplamaya odaklanmaktadır [3]. Buna karşılık, açık alan QA sistemi ise farklı alanlardaki soruları cevaplamayı amaçlamaktadır. Böyle bir durumda, farklı alanlardaki soruları cevaplamak için dünya bilgisine erişmek gerekmektedir [4]. Ayrıca QA sistemi için, soru türüne bağlı olarak, evet/hayır veya normal metin şeklinde cevaplar yer almaktadır [5].

NLP'de soru cevaplamada büyük dil modelleri (LLM) kullanımı günümüzde oldukça yaygındır. LLM'ler, bir kodlayıcı, kod çözücü veya her ikisinden oluşan derin öğrenmeye dayalı modellerdir. 2017'den önce, çoğu NLP modeli yalnızca belirli görevler için gözetimli öğrenme ile eğitilip kullanılabilirken, bu kısıtlamayı gidermek için Transformers [6] adlı bir öz-dikkat ağı mimarisine sahip model geliştirilmiştir. Transformers modeli ile beraber, farklı NLP görevleri için LLM'lerin oluşturulması gerçekleştirilmiştir. Örneğin Transformers'dan Çift Yönlü Kodlayıcı Temsilleri (BERT) [7] ve Üretken Önceden Eğitilmiş Transformatör (GPT) [8] modelleri farklı NLP görevleri için kullanılmaktadır. Her iki model de yarı gözetimli yaklaşımlarla üstün genelleme yetenekleri göstermektedir. LLM'lerde geniş bir veri kümesi üzerinde (genellikle etiketlenmemiş) genel bilgi öğrenmek için modelin eğitilmesi ön-eğitim aşaması iken, önceden eğitilmiş modelin belirli bir görev veya belirli bir alan için daha küçük, etiketlenmiş bir veri kümesi ile eğitilmesi ince ayar aşaması olarak tanımlanmaktadır [9].

İngilizce, Çince gibi birçok dilde QA dahil farklı NLP görevleri için LLM'lerin kullanımı mevcuttur. Türkçe dili için LLM'lerin kullanımına katkı sağlamak için bu çalışmada Türkçe QA görevine odaklanılmaktadır. BERT, DistilBERT, ALBERT ve mDeBERTa LLM'leri, SQuAD 2.0 veri kümesinin Türkçe sürümü (https://github.com/boun-tabi/SQuAD-TR) kullanılarak ince ayar ile eğitilmektedir. Eğitilen tüm LLM'ler test veri kümesi ile analiz edilmektedir. Ayrıca çalışmada, LLM'lerin tahmin ettiği cevaplardaki skoru için eşik değerinin etkisi, LLM'lerin tahmin ettiği ve gerçek cevaplar arasındaki anlamsal benzerliğin modellerin performansına etkisi araştırılmaktadır. Son aşamada ise ince ayar ile çok dilli T5(mT5) metin oluşturma modelinin eğitilerek bu modelin QA görevi için başarısı da analiz edilmektedir.

Bu çalışmanın temel katkıları şunlardır:

- Türkçe QA görevi için LLM'leri ve metin üretim modellerini karşılaştıran öncü bir çalışmadır.
- Literatüre Türkçe dili için LLM'lerin geliştirilmesi ile beraber literatüre katkı sağlamaktadır.
- LLM'lerin hangi soru tiplerini daha iyi cevaplayabildiği analiz edilmektedir.
- LLM'lerin tahmin ettiği cevaplardaki skorun, eşik değerine göre analiz edilmesiyle beraber LLM'lere olumlu etkisi gösterilmektedir.
- LLM'lerin tahmin ettiği ve gerçek cevap arasındaki anlamsal benzerliğin olumlu etkisi analizlerle desteklenmektedir.

Bu çalışmanın özet akışı şu şekildedir. Literatürdeki benzer çalışmalar "İlgili Çalışmalar" bölümünde belirtilmiştir. "Metodoloji" bölümü bu çalışmanın metodolojisini, kullanılan veri setini, LLM'leri ve yöntemleri açıklamaktadır. Dördüncü bölüm LLM'ler ve metin oluşturma modelleri üzerine analizler ve sonuçları içermektedir. Modellerin performansı "Tartışma" bölümünde incelenmektedir. Son bölümde ise bu çalışmanın sonuçları ve gelecekteki çalışmalar belirtilmektedir.

2. İlgili Çalışmalar (Related Works)

Alzubi vd. [10], araştırmacıların ve klinik personelin doğru bilimsel bilgilere erişmesini sağlamak için COBERT QA sistemini önermişlerdir. Bu sistemde COVID-19'un zorlukları ele alınmıştır. Önerdikleri model, Retrieval, Reader ve Ranker'dan oluşan üç bileşenli bir mimari içermiştir. Analiz sonucunda, önerdikleri DistilBERT modelinin en iyi performansı gösterdiğini belirtmişlerdir. Kierszbaum ve Lapasset [11], uluslararası havacılık güvenliği bağlamındaki hadiseleri tanımlayan Havacılık Güvenliği Raporlama Sistemi (ASRS) veri setini, serbest metin raporlarında QA görevi için BERT modeli ile analiz etmişlerdir. ASRS veri setinin yapılandırılmış kısmıyla doğrudan çalışmak yerine doğal dil anlatımlarındaki bilgileri değerlendirmek için NLP algoritmalarını entegre etmişlerdir. Önerdikleri yaklaşım ile yaklaşık %70'lik bir doğruluk elde etmişlerdir. LLM'ler klinik bilgileri değerlendirmek için genellikle sınırlı ölçütlere güvendiklerinden, Singhal vd. [12], profesyonel tıp, araştırma sorguları ve çevrimiçi aramalar sağlayan HealthSearchQA içeren altı tıbbi QA veri kümesini birleştiren MultiMedQA'yı önermişlerdir. Analizleri sonucunda, önerilen modelleri %67.6'lık bir başarı oranına ulaşmıştır. Abdelhay vd. [13], Arapça'daki sağlık QA botu güvenlik açığını ortadan kaldırmak için bir tıbbi bot önermişlerdir. QA veri kümesini Transformers ve LSTM'den türetilen modellerle analiz etmişler ve Transformers modelinin daha başarılı olduğunu göstermişlerdir. Wang vd. [14], aynı sorunun tüm pasajları için cevap puanlarını genel olarak normalleştiren çok pasajlı bir BERT modeli önermişlerdir. Önerilen modelleri ile daha fazla pasaj kullanarak cevapları daha iyi tespit ettiğini göstermişlerir. Qu vd. [15], QA görevi için BERT modelinde geçmiş cevap yerleştirme adı verilen bir yaklaşım önermişlerdir. Bunu BERT modeline entegre etmek için konuşma geçmişini kullanmışlardır. İlk olarak yaklaşımlarını bir konuşma arama ortamı olarak tanımlamışlar ve ardından QA görevi için genel bir çerçeve sunmuşlardır.

Önceki çalışmalarda SQuAD veri kümesi için birçok LLM kullanılmıştır. Devlin vd. [7], dikkat mekanizmasını kullanan ve çift yönlü bağlamı inceleyen Transformatör tabanlı BERT modelini önermişlerdir. Önerilen model sınıflandırma, QA vb. gibi birçok görev için kullanılabilmektedir. BERT modeli QA görevi için SQuAD ile analiz edilmiştir. Zhang vd. [16], grameri mümkün olduğunca etkili bir şekilde kullanmak için daha iyi dilsel kelime temsili mekanizmasına açık sözdizimsel kısıtlamalar eklemiştir. Yaklaşımın etkisini analiz etmek için, önerilen sözdizimi odaklı ağını (SG-Net) önceden eğitilmiş BERT modeline uygulamışlardır. QA görevinde SQuAD'a uygulanan analizler sonucunda, yaklaşımları BERT modelinden daha iyi sonuç vermiştir. Zhang vd. [17], açık bağlamsal semantiği içeren Semantik-aware BERT (SemBERT) dil temsil modelini geliştirmişlerdir. Modeli ince ayarlamak için BERT'i kullanmışlardır. SQuAD için yapılan analizler sonucunda SemBERT,

BERT modelinden daha iyi performans göstermiştir. Lan vd. [18], BERT modelinin bellek sınırı ve eğitim süresi sorunu nedeniyle daha az parametreli BERT tabanlı ALBERT modelini önermişlerdir. Analiz sonucunda, BERT'den çok daha hızlı eğitilen ve daha az bellek kullanan bu modelle SQuAD için etkileyici sonuçlar elde etmişlerdir. Türkçe QA görevi için LLM çalışmalarının sayısı nispeten düşüktür. Soygazi vd. [19], tarihsel QA veri kümesi olan TQuAD ile soruyu pasajı anlayarak cevapladıkları LLM modellerini analiz etmişlerdir. Önceden eğitilmiş Türkçe BERT, ELECTRA ve ALBERT dil modellerini ince ayarlayarak modelleri test etmişlerdir. Akyon vd. [20], QA ve soru oluşturma için çoklu görev yapacak çok dilli bir T5 (mT5) dönüştürücüyü ince ayarlamışlardır. Bu işlem için TQuAD geçmiş veri kümesini kullanmışlardır. Analizleri sonucunda, önerdikleri model soruları yanıtlamada iyi bir performans göstermiştir. İncidelen ve Aydoğan [21], Türkiye'deki Yükseköğretim Kurulu Tez Merkezi'nde bulunan Türkçe Vikipedi ve tıp tezleri kullanarak bir tıp Türkçe QA veri seti sunmuşlardır. Toplam 8200 soru-cevap çifti içeren bu veri seti üzerinde BERTurk modelini ince ayar ile eğiterek analiz etmişlerdir. Tutar ve Yildiz [22], Microsoft Translator kullanarak kapsamlı İngilizce SQuAD 1.1 veri kümesinden bir Türkçe QA veri kümesi elde etmişler ve bu verilerin yaklaşık %60'ını kullanarak BERTurk modelini analiz etmişlerdir. Türkçe için büyük dil modelleri üzerine farklı çalışmalar da mevcuttur. Türkçe patent sınıflandırması için Kahraman vd uyguladığı çalışma [23] ve Türkçe ses kayıtları ile duygu analizine yönelik Tepecik ve Demir'in çalışması [24] örnek olarak verilebilir.

3. Metodoloji (Methodology)

Bu çalışmada, Türkçe QA görevinde LLM'lerin kullanımına katkı sağlamak için LLM'ler ince ayar ile eğitilerek analiz edilmektedir. Türkçe için önceden eğitilmiş BERT, ALBERT ve DistilBERT ve çoklu diller için önceden eğitilmiş mDeBERTa LLM'leri bu aşamada kullanılmaktadır. Bu LLM'ler QA görevi için SQuAD veri kümesinin Türkçe versiyonu ile ince ayar yapılarak eğitilmektedir. Model analizinde, farklı eşik değerlerinin cevapları tespit etmedeki etkisi ve gerçek-tahmini cevaplar arasındaki anlamsal benzerliğin etkisi de araştırılmaktadır. Son aşamada, bir metin üretim modeli olan mT5'in QA veri kümesiyle ince ayar uygulanması ile bu modelin doğruluğu analiz edilmektedir. Tüm analizler sonucunda, LLM'lerin doğruluk değerleri karşılaştırılmaktadır. Bu çalışmanın metodolojisi Şekil 1'de gösterilmektedir.

3.1. Veri Seti (Dataset)

Başlıklar, bağlam paragrafları, sorular ve cevaplar dahil olmak üzere orijinal İngilizce SQuAD2.0 veri seti, Budur vd. [25] tarafından Amazon Translate kullanılarak Türkçe'ye çevrilmiştir. Cevap aralıklarının başlangıç pozisyonları, otomatik çeviri nedeniyle pozisyonlarında meydana gelen değişiklikler nedeniyle yeniden güncellenmiştir [25]. Çevrilen veri setinin istatistikleri Tablo 1'de verilmektedir. Bu veri seti hem cevaplı hem de cevapsız sorulardan oluşmaktadır. Ayrıca Türkçe SQuAD'ın eğitim ve test veri setlerine ait istatistikler de tabloda belirtilmektedir.

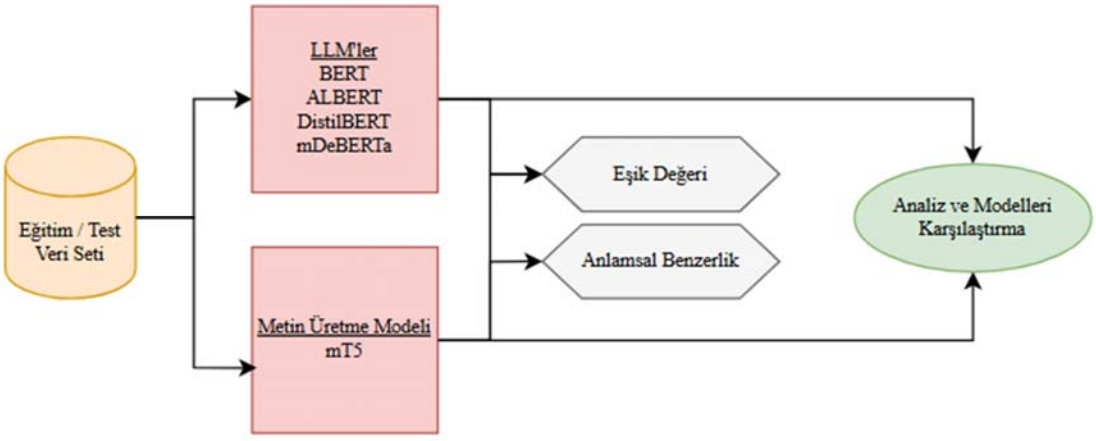
Türkçe SQuAD'ına bir örnek olarak paragraf (bağlam), soru ve ilgili cevap Şekil 2'de verilmektedir. Model paragraf ve soruyu girdi olarak alarak cevabı üretmektedir.

3.1. Büyük Dil Modelleri (Large Language Models)

Önceden eğitilmiş BERT, ALBERT, DistilBERT ve mDeBERTa LLM'leri, Türkçe QA görevi için Türkçe SQuAD ile ince ayar yapılarak eğitilmektedir. Ardından eğitilen modeller test verileri ile test edilmektedir. Alt bölümler, bu LLM'lerin genel mimarisini açıklamaktadır.

3.1.1. BERT

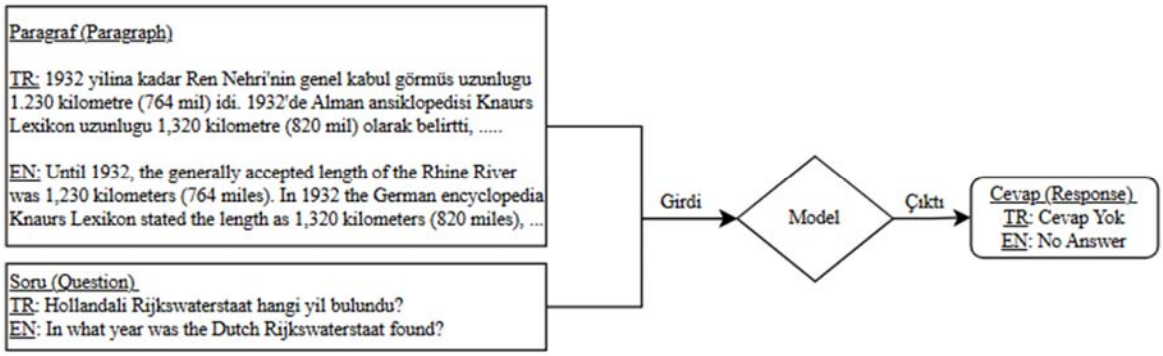
Transformatörlerden Çift Yönlü Kodlayıcı Temsilleri anlamına gelen BERT, her görev için bir kelimenin hem sol hem de sağ bağlamını inceleyen bir modeldir. Birden fazla bağlamsal kelime temsili üretmektedir. Bu model maskeleme, sonraki cümle tahmini, QA ve metin sınıflandırması gibi çeşitli NLP görevlerinde kullanılmaktadır [7]. BERT modelinin QA görevi için süreci Şekil 3'te gösterilmektedir. Bu şekil, modelin paragrafları (bağlam) ve soruları girdi olarak aldığını ve her iki girdiyi belirteçlerle birleştirip modeli besleyerek bir cevap ürettiğini göstermektedir.



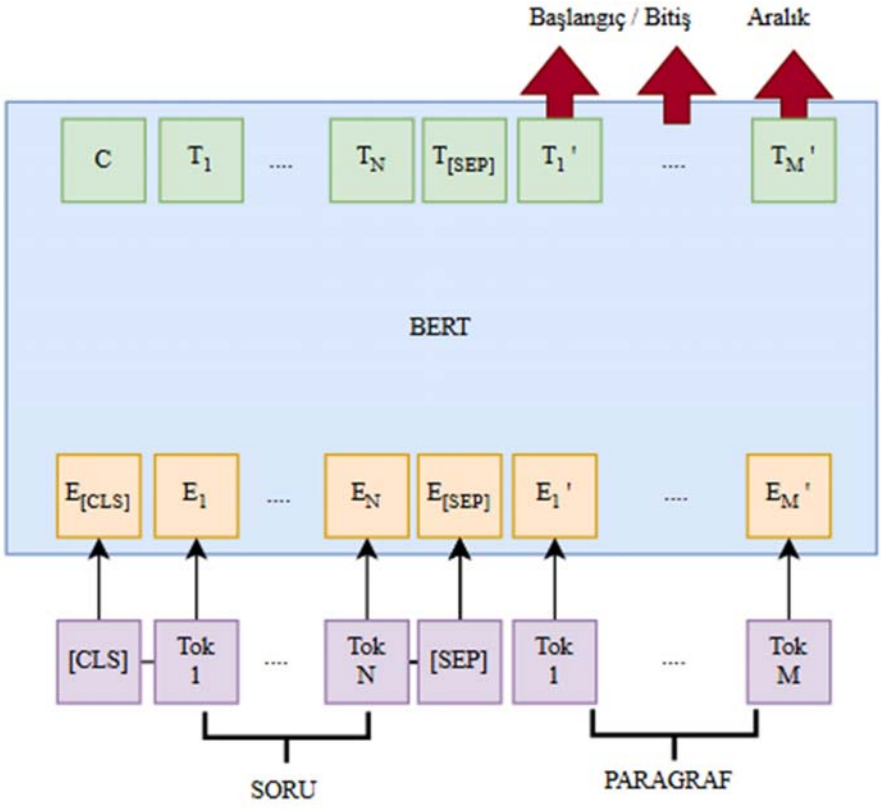
Şekil 1. Çalışmanın metodolojisi (Methodology of this study)

Tablo 1. Türkçe SQuAD'ın istatistikleri [25] (Statistics of Turkish SQuAD)

	Makale	Paragraf	Cevaplanabilir Sorular	Cevap İçermeyen Sorular	Toplam
Eğitim (train)	442	18776	61293	43498	104791
Test (test)	35	1204	2346	5945	8291



Şekil 2. Türkçe SQuAD için örnek senaryo (Example scenario for Turkish SQuAD)



Şekil 3. QA görevi için BERT modelinin süreci [10] (Processes of BERT model for QA task)

BERT modeli yapısında ön eğitim ve ince ayar aşamalarını içermektedir. Ön eğitim aşamasında, model etiketlenmemiş verilerle eğitilirken, ince ayar aşamasında önceden eğitilmiş parametrelerle başlatılan model herhangi bir görev için etiketlenmiş verilerle ince ayarlanmaktadır [7].

3.2.2. ALBERT

BERT modelinin bellek sınırlamaları ve iletişim yükü problemleri bulunmaktadır. Bu sorunu ele almak için daha az parametreli ALBERT modeli önerilmiştir. Parametre sayısını azaltmak için faktörlü yerleştirme parametrelendirmesi ve katmanlar arasında parametre paylaşımı tekniklerini kullanmaktadır. Böylece ağ derinliği arttıkça parametrenin artmasını önlemektedir [18]. BERT tabanlı bir model olduğundan QA dahil olmak üzere birçok görevde kullanılabilmektedir.

3.2.3. DistilBERT

BERT'den daha küçük ve daha hızlı bir model olan DistilBERT, yalnızca ham metinler kullanılarak otomatik bir işlemle önceden eğitilerek oluşturulmuştur. Bu model, damıtma kaybı, maskelenmiş dil modellemesi ve kosinüs yerleştirme kaybı ile önceden eğitilmiştir. Damıtma kaybı, BERT ile tam olasılıklarla eşleşmeyi sağlarken, kosinüs yerleştirme kaybı gizli durumları yakalamak için kullanılmaktadır. Maskelenmiş dil modelleme ise BERT gibi girdideki kelimeleri maskelemek için kullanılmaktadır [26]. Bu model QA görevi için eğitilebilmektedir.

3.2.4. mDeBERTa

BERT ve RoBERTa modellerini iki yöntemle birleştiren DeBERTa (Dağınık dikkat ile kod çözme-geliştirilmiş BERT) modeli

önerilmiştir. Dağınık dikkat mekanizması ilk yöntem olarak kullanılmaktadır. Bu yöntemde her kelime sırasıyla içeriğini ve konumunu kodlayan iki vektörle temsil edilmektedir. Ek olarak, kelimeler arasındaki dikkat ağırlıkları, içerikleri ve göreli konumları üzerinde ayrıştırılmış matrisler kullanılarak hesaplanmaktadır. İkinci yöntem olarak geliştirilmiş bir maske kod çözücü bulunmaktadır. Bu yöntem, kod çözme katmanındaki mutlak konumları birleştirerek ön eğitiminde maskelenen belirteçleri tahmin etmektedir [27]. Bu model çok dilli DeBERTa (mDeBERTa) olarak önceden eğitilmiş olarak kullanılabilmektedir. LLM'lerin genel versiyonlarının tüm özellikleri Tablo 2'de verilmektedir. Bu çalışma, bu modellerin önceden eğitilmiş LLM'lerini Türkçe QA görevi için kullanmaktadır.

3.3. Metin Üretme Modeli: T5 (Text Generation Model: T5)

T5, ana özelliği "metinden metne" biçimini kullanmak olan önceden eğitilmiş bir dil modelidir. Bu yaklaşım, görev biçiminin modelin bazı girdilere dayalı olarak metin oluşturmasına izin verdiği üretken görevler için kullanılmaktadır. mT5 modeli, 101 dili kapsayan yeni bir Ortak Arama tabanlı veri kümesi üzerinde önceden eğitilmiş, T5'in çok dilli bir sürümüdür [28].

QA görevi için, çok dilli T5 olan metin oluşturma modeli, ince ayar süreci boyunca Türkçe SQuAD ile eğitilmektedir. Daha sonra, bu eğitilmiş modele istemler (prompt) gönderilmektedir. İstem bir paragraf (bağlam) ve bir soru içermektedir.

4. Deneysel Çalışmalar (Experimental Results)

Deneysel çalışmalar iki kısımda incelenmektedir. İlk aşamada kullanılan veri seti üzerinde analizler gerçekleştirilirken, diğer aşamada ise LLM'lerin analizi üzerinde durulmaktadır. Analizler iki ayrı alt başlık altında detaylı anlatılmaktadır.

4.1. Türkçe SQuAD'ın Analizi (Analysis of Turkish SQuAD)

Veri setinde, çeşitli soru zamirleri aranan cevapları yönlendirmektedir (örneğin, "kim" bir kişinin arandığını ifade eder). Bu soru zamirleri hakkında istatistikler çıkarılmıştır. İstatistiklere ait sonuçlar Tablo 3'te sunulmaktadır. Tablo, "ne" soru zamirinin eğitim ve test setlerinde yaygın olarak bulunduğunu göstermektedir.

4.2. LLM'lerin *Analizi (Analysis of LLMs)*

Huggingface gibi platformlarda birçok görev için büyük modeller bulunmaktadır. Bu çalışmada, Türkçe soruları cevaplamak için bu platformdaki ön-eğitilmiş LLM'ler kullanılmıştır. Kullanılan Türkçe LLM modelleri:

- BERT-TrSQuAD (https://huggingface.co/anilguven/bert_tr_qa_turkish_squad),
- ALBERT-TrSQuAD (https://huggingface.co/anilguven/albert_tr_qa_turkish_squad),
- DistilBERT-TrSQuAD (https://huggingface.co/anilguven/distilbert_tr_qa_turkish_squad)
- mDeBERTa-TrSQuAD (https://huggingface.co/anilguven/mdeberta_tr_qa_turkish_squad)

Bu LLM'ler, Türkçe SQuAD üzerinde ince ayar yapılarak eğitilmiş ve test edilmiştir. Bu LLM'lerin analiz sonucunda doğruluk değerleri Tablo 4'te verilmiştir. Cevaplanabilir ve cevap içermeyen soruların sayısı dengesiz olduğundan (Tablo 1), LLM'lerin bu soru türlerine ait

Tablo 2. LLM'lerin istatistikleri (Statistics of LLMs)

	Parametre Sayısı (Parameter Count)	Katman (Layer)	Gizli Boyut (Hidden Size)
BERT-base	110M	12	768
BERT-large	340M	24	1024
ALBERT-base	12M	12	768
ALBERT-large	18M	24	1024
DistilBERT	66M	6	768
mDeBERTa-base	276M	12	768

Tablo 3. Soru zamirlerinin istatistikleri (Statistics of question pronouns)

Soru Zamirleri (Pronouns)	Eğitim Kümesi (Train set)	Test Kümesi (Test set)
Ne (what)	41500	3830
Hangi (which)	29911	1894
Kim (who)	9808	664
Ne zaman (when)	7302	576
Kaç (how many/much)	5805	341
Nerede (where)	4268	336
Neden (why)	2420	239
Nasıl (how)	1752	230
Diğer – mı, mi, vb. (others)	2025	181
Total Count	104791	8291

Tablo 4. LLM'lerin doğruluk değerleri (Accuracies of LLMs)

LLM'ler (LLMs)	Cevaplanabilir (Answerable) (%)	Cevabı Olmayan (No-Answer) (%)	Genel Doğruluk (General Accuracy) (%)
BERT-TrSQuAD	34.27	87.42	72.37
ALBERT-TrSQuAD	14.19	94.23	71.58
DistilBERT-TrSQuAD	18.93	90.03	69.90
mDeBERTa-TrSQuAD	46.63	85.50	74.50

başarısı da analiz edilmiştir. Tablo en başarılı modelin mDeBERTa-TrSQuAD olduğunu göstermektedir.

Cevaplar için öngörülen olasılıklar incelendiğinde, modellerin sorularda düşük sonuçlar gösterdiği Tablo 4'te gözükmektedir. Bu sonuçları ele almak için, LLM'lerin ürettiği cevapların tahmin olasılıkları için 0.3, 0.4 ve 0.5 eşik değerleri belirlenerek analiz gerçekleştirilmiştir. Bu olasılıkların altındaki tüm tahmin edilen cevaplar "Cevap Yok" olarak etiketlenmiştir. LLM'lerin eşik değerlerine dayalı doğruluk değerleri Tablo 5'te sunulmaktadır. Tablo, LLM'lerin cevabı olmayan soruları cevaplamada daha başarılı olduğunu göstermektedir. ALBERT-TrSQuAD modeli cevabı olmayan sorular için daha iyi bir doğruluğa sahipken, mDeBERTa-TrSQuAD cevaplanabilir sorular için daha iyi performans göstermiştir.

Önceki tablolar LLM'lerin soruları yanıtlamada nispeten düşük bir başarı oranına sahip olduğunu göstermektedir. Bunun nedeni, QA modelleri tarafından sağlanan yanıtlar doğru yanıtı yansıtsa da, yapılarındaki farklılıklar nedeniyle farklı olarak görülebilmesidir (Örneğin; "Kiliselerde korunmus Norman sanatinin en önemli türü nedir?" sorusu için doğru cevap "mozaikler" iken model "mozaikleri" sonucunu geri döndürmesi).

Sonuç olarak, LLM'lerin soruları yanıtlamadaki başarısı, gerçek ve tahmin edilen yanıtlar arasındaki anlamsal benzerliğin model (https://huggingface.co/emrecan/bert-base-turkish-cased-mean-nli-stsb-tr) ile ölçülmesiyle yeniden analiz edilmiştir. Bu model ile beraber Türkçe'deki cümleler arasındaki semantik benzerlik değerlendirilmektedir. Model, NLI ve STS-b veri kümelerinin Türkçe makine çevirisi sürümleri üzerinde eğitilmiştir [29]. Model ile elde edilen analizin bulguları Tablo 6'da sunulmuştur. Bu tablodaki verileri incelendiğinde, cevaplanabilir soruların doğruluğunun arttığı gözlemlenmiştir.

Tüm deneysel çalışmalara dayalı olarak Türkçe QA görevi için en başarılı LLM mDeBERTa-TrSQuAD olmuştur. Analiz olarak ayrıca bu modelin en doğru cevaplanan soru zamirlerine odaklanılmıştır. Bu analizden elde edilen sonuçlar Tablo 7'de verilmiştir. Bu tablo, mDeBERTa-TrSQuAD modelinin "diğer" kategorisi hariç, benzer olasılıkla soru zamirlerini doğru cevapladığını göstermiştir. Soru zamiri "ne (what)" en yüksek doğruluğa sahipken, "hangi (which)" en yüksek cevap oranına sahip olmuştur.

mDeBERTa-TrSQuAD modeli için yanlış ve doğru cevaplanan sorulara yönelik bazı örnekler Tablo 8'de verilmiştir. Model çoğu soru zamirini %75'in üzerinde doğru tahmin etmiştir (Tablo 7).

Tablo 5. LLM'lerin eşik değerlerine göre doğruluk değerleri (Accuracies of LLMs according to threshold values)

	Eşik Değeri (Threshold)	Cevaplanabilir (Answerable) (%)	Cevabı Olmayan (No-Answer) (%)	Genel Doğruluk (General Accuracy) (%)
	0.3	43.65	75.16	66.24
BERT-TrSOuAD	0.4	43.65	78.81	68.86
	0.5	43.65	82.37	71.41
	0.3	14.19	94.23	71.58
ALBERT-TrSOuAD	0.4	14.19	94.23	71.58
	0.5	14.19	94.26	71.61
	0.3	18.93	90.03	69.91
DistilBERT-TrSOuAD	0.4	18.93	90.03	69.91
	0.5	18.93	90.19	70.03
	0.3	46.63	85.50	74.50
mDeBERTa-TrSOuAD	0.4	46.63	85.53	74.53
	0.5	46.63	85.54	74.55

Tablo 6. Anlamsal benzerlik süreci ile LLM'lerin doğruluk değerleri
(Accuracy values of LLMs with semantic similarity process)

LLM'ler (LLMs)	Cevaplanabilir (Answerable) (%)	Cevabı Olmayan (No-Answer) (%)	Genel Doğruluk (General Accuracy) (%)
BERT-TrSQuAD	45.27	87.42	75.49
ALBERT-TrSQuAD	16.71	94.23	72.30
DistilBERT-TrSQuAD	23.36	90.03	71.16
mDeBERTa-TrSQuAD	62.83	85.50	79.09

Tablo 7. MDeBERTa-TrSQuAD için doğru cevaplanan sorularda soru zamirlerinin istatistikleri
(Statistics of question pronouns in correctly answered questions for mDeBERTa-TrSQuAD)

Soru Zamirleri (Pronouns)	Doğru Cevaplanan (Correctly answered)	Toplam Sayı Total Count	Oran-% (Rate)
Ne (what)	3093	3830	80.76
Hangi (which)	1593	1894	84.11
Kim (who)	476	664	71.69
Ne zaman (when)	483	576	83.85
Kaç (how many/much)	278	341	81.52
Nerede (where)	252	336	75.00
Neden (why)	199	239	83.26
Nasıl (how)	181	230	78.70
Diğer - mı, mi vb. (others)	2	181	1.10

Sonraki aşamada, metin oluşturma modelinin Türkçe soruları yanıtlamadaki performansı analiz edilmiştir. Bu görev için mT5 modeli Türkçe SQuAD ile ince ayarlanmıştır ve modelin başarısı test seti kullanılarak değerlendirilmiştir. Metin oluşturma modelleri cümleler halinde soruları cevaplayabildiğinden bu aşamada gerçek ve tahmin edilen cevap benzerlik ölçümü için yine model kullanılmıştır. Ayrıca, model düşük olasılıklı yanıtları (eşik değeri 0,5'ten küçük) "Cevap Yok" olarak işaretleyerek analiz edilmiştir. Daha sonra, gerçek ve tahmin edilen yanıtlar arasındaki anlamsal benzerliği (AB) ölçerek bu görevin başarı üzerindeki etkisi incelenmiştir. Bu analizin sonuçları Tablo 8'de sunulmuştur. Sonuçlar, mT5 modelinin cevapları olan soruları daha başarılı bir şekilde yanıtladığını göstermiştir.

Bu çalışmada QA görevi için çeşitli analizler yürütülmüştür. Tüm analizleri içeren LLM'lerin ve mT5-Tr'nin doğruluk değerleri Şekil 4'te gösterilmektedir. Bu şekil incelendiğinde en başarılı modelin mDeBERTa-TrSQuAD olduğu görülmektedir. Bu çalışma eşik değeri (ED) işleminin cevap tespitini önemli ölçüde etkilediğini ortaya koymuştur. Ayrıca, gerçek ve tahmin edilen cevaplar arasında anlamsal benzerlik işleminin uygulanması modelin başarısını olumlu yönde etkilemiştir.

5. Sonuçlar ve Tartışmalar (Results and Discussions)

Bu çalışma Türkçe dili için QA görevine odaklanmıştır. Bu görevi gerçekleştirmek için ilk aşamada LLM'ler kullanılmıştır. Bu LLM'ler, Türkçe dili için ince ayar süreciyle ön-eğitilmiş Türkçe BERT, DistilBERT, ALBERT ve mDeBERTa LLM'leri kullanılarak Türkçe SQuAD ile eğitilmiş ve test edilmiştir. Analiz sonucunda en başarılı LLM %74.50 doğruluk değeriyle mDeBERTa-TrSQuAD olmuştur. DeBERTa V3, ayrıştırılmış dikkati ve maske kod çözmeyi geliştirmiştir ve Gradient Disentangled Embedding Sharing ile ön-eğitilmiş ELECTRA Style'dır. Bu, BERT'ye kıyasla önemli bir performans artışına yol açmıştır. Öte yandan, DistilBERT-TrSQuAD ve ALBERT-TrSQuAD modelleri en az başarılı olanlardır, çünkü BERT modelinin ölçeklendirilmiş ve parametresi azaltılmış sürümleridir. Böylece bu modeller için eğitim performansında kayıplara neden olmuştur.

Diğer aşamada, tahmin edilen cevaplar ile LLM'leri analiz etmek için modeller ile eşik değeri arasındaki ilişki araştırılmıştır. Eşik değerlerini 0.3, 0.4 ve 0.5 olarak belirlenmiştir. Analiz sonucunda, BERT-TrSQuAD modeli hariç, 0.5 eşik değeri kullanmanın %0.03 ile %0.13 arasında LLM'lerin başarısını artırdığı gözlemlenmiştir. Bu, BERT-TrSQuAD modeli dışında, diğer modellerde "Cevap Yok" için varsayılan eşik değerlerinin 0.5'ten düşük olduğunu ve başarılarını olumsuz etkilediğini göstermiştir. BERT-TrSQuAD modeli için varsayılan eşik değerinin 0.5'ten yüksek olduğu analiz sonucunda ortaya çıkmıştır.

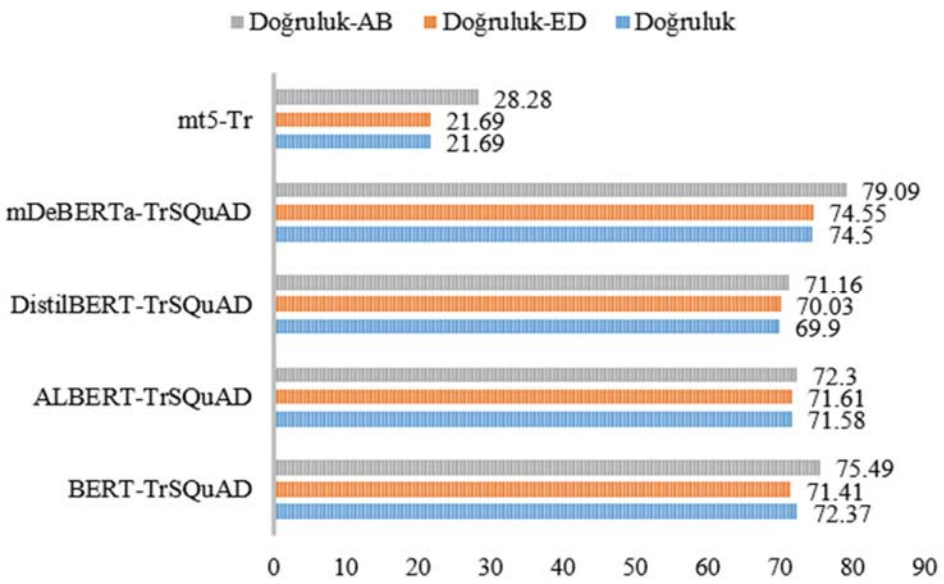
QA için en kritik sorun, modelin döndürdüğü yanıtların gerçek yanıtı içermesine (ek içermesi, cümleye dahil olması vb.) rağmen, modelin yalnızca bir gerçek yanıt nedeniyle bunu yanlış kabul edebilmesidir. Bu nedenle, gerçek ve tahmin edilen yanıtlar arasındaki anlamsal benzerliği hesaplayarak modelleri test etmek esastır. Analiz aşamasında bu yanıtların benzerliği 0.5'ten fazlaysa, bu LLM'ler yanıtı doğru olarak kabul etmiştir. LLM'lerin bu analiz sonucunda, doğruluk değerleri yaklaşık 0.7% ile 4.5% artış göstermiştir. Bu aşama, anlamsal benzerliğin QA sistemleri için ne kadar vazgeçilmez olduğunu göstermiştir.

Tablo 8. mDeBERTa-Tr için doğru ve yanlış cevaplanan soru-cevap çiftleri
(Question-answer pairs with correct and incorrect answers for mDeBERTa-Tr)

Soru Zamirleri (Pronouns)	Doğru Cevaplanan (Correctly answered)	Yanlış Cevaplanan (Wrongly answered)
Ne (what)	Soru: Normanlar ne din vardi Cevap: Katolik	Soru: Richard'in Acre'ye ulasma planlarini ne mahvetti? Cevap: bir firtina
Hangi (which)	Soru: Normandiya hangi ülkede bulunur? Cevap: Fransa	Soru: Kanarya Adalari kiyilarinda hangi kita vardir? Cevap: Afrika
Kim (who)	Soru: Hastings savasinda dük kimdi? Cevap: Fatih William	Soru: Iskandinav lideri kimdi? Cevap: Rollo
Ne zaman (when)	Soru: Norman'lar ne zaman Normandiya'daydilar? Cevap: 10. ve 11. yüzyillar	Soru: Ilk Fransizca Incil ne zaman ortaya çikti? Cevap: 1294 civari
Kaç (how many/much)	Soru: Robert'in ordusunda kaç adam vardi? Cevap: 30.000	Soru: Bir islev problemindeki her giris için kaç çikis beklenir? Cevap: tek
Nerede (where)	Soru: Normanlar ve Bizanslilar baris antlasmasini nerede imzaladi? Cevap: Deabolis	Soru: Huguenots ve Valonlar Ingiltere'de nerede yerlesti? Cevap: Canterbury
Neden (why)	Soru: Hayvanlar neden biyolojik solunum için O2 kullaniyor? Cevap: Cevap Yok	Soru: Berlin Huguenots hizmetlerinde neden Fransizca'dan Almanca'ya geçti? Cevap: Napolyon tarafindan Prusya'nin isgaline karsi protesto
Nasıl (how)	Soru: Hesaplanan bir soruna çözüm elde etmek için gereken süre nasil hesaplanir? Cevap: örnegin boyutunun bir islevi olarak	Soru: Fonksiyon sorunlari genellikle nasil yeniden döndürülebilir? Cevap: karar problemleri
Diğer -mı, mi vb. (others)	Soru: Eureka Stockade'de maden vergileri hakkinda silahli bir protesto yapildi mi? Cevap: Cevap Yok	Soru: Konsey Baskani Avrupa Merkez Bankasi ile ilgili önemli konularda oy verebilir mi? Cevap: oy hakki yok

Tablo 8. mT5 modelinin doğruluk değerleri (%) (Accuracy values of the mT5 model)

LLM'ler (LLMs)	Cevaplanabilir (Answerable) (%)	Cevabı Olmayan (No-Answer) (%)	Genel Doğruluk (General Accuracy) (%)
mT5-Tr	63.00	5.38	21.69
mT5-Tr + AB	86.27	5.38	28.28



Şekil 4. Süreçlerin uygulanmasının ardından tüm modellerin doğruluk değerleri (%)
(Accuracy values of all models after implementation of the processes)

Son aşamada QA görevleri için son zamanlarda popüler olan metin oluşturma modeli kullanılmıştır. mT5 modeli bu görev için veri kümesiyle ince ayar yapılarak test edilmiştir. Bu modeller cevaba yakın çıktılar üretebildiğinden, gerçek cevap ile model çıktısı arasındaki anlamsal benzerlik hesaplanmış ve 0.5'in üzerindeki cevaplar doğru kabul edilmiştir. Ayrıca bu modelde 0.5'lik eşik değerinden düşük model çıktıları "Cevap Yok" olarak etiketlenmiştir. Analiz sonucunda modelin başarısının %21.69 olduğu, ancak cevapları olan sorular için %63.00'lık bir doğruluk değeri verdiği görülmüştür. Anlamsal benzerlik gerçekleştirilmesi sonucunda ise bu modelin başarısı %6.59 artarken, cevapları olan sorular için %23.27'lik bir doğruluk artışı sağlanmıştır. Bu süreç, anlamsal benzerliğin metin oluşturma modelleri için olumlu bir performans artışı sağladığını göstermiştir. Cevabı olmayan sorularda başarının düşük olmasının sebebi, bu modellerin cevap üretmeye yönelik olması ile açıklanabilir.

Türkçe dili için geçmişteki çalışmalar incelendiğinde ise SQuAD veri seti üzerine çalışmalar çok fazla bulunmamaktadır. Ancak soru cevaplama üzerine Türkçe dilinde birçok çalışma uygulanmıştır. [19], [20], [21], [22] ve bu çalışmanın veri kümelerini ve sonuçlarını içeren veriler Tablo 9'da verilmiştir. Tablo gösteriyor ki Türkçe QA görevi için farklı veri kümeleri üzerinde çalışmalar yapılmıştır. Geçmiş çalışmalara kıyasla, bu çalışmadaki en başarılı model olan mDeBERTa-TrSQuAD oldukça iyi bir sonuç göstermiştir. Bu çalışma, "Cevap Yok" cevabına sahip soruları da incelemesi açısından Türkçe için literature önemli katkı sağlamıştır. Ayrıca önemli bir İngilizce kıyaslama veri setinin Türkçe versiyonu değerlendirilerek Türkçe dili açısından etkili bir analiz gerçekleştirilmiştir.

Tablo 9. Türkçe dili için geçmiş QA çalışmaları ile kıyaslama
(Comparison with previous QA studies for Turkish language)

Çalışmalar	Veri Seti	Doğruluk/F1 (%)
[19]	Thquad	81.55
[20]	TQuAD, XQuAD	74.70
[21]	MedTurkQuAD	77.18
[22]	Turkish SQuAD 1.1	63.00
Bu çalışma	Turkish SQuAD 2.0	79.09

6. Sonuçlar (Conclusions)

Bu çalışma, LLM'lerin Türkçe QA görevi için kullanılmasına ek olarak, eşik ve anlamsal benzerlik değerlerinin LLM'lerin performansına etkisinin incelenmesine öncülük etmektedir. Bu görev için Türkçe BERT, DistilBERT, ALBERT ve mDeBERTa LLM'leri ve mT5 metin üretme modeli kullanılmıştır. İlk aşamada bu modeller Türkçe SQuAD eğitim veri kümesi üzerinde ince ayar yapılarak eğitilmiş ve test veri kümesi ile analiz edilmiştir. Analizler sonucunda, en başarılı modelin %74.50 ile mDeBERTa-TrSQuAD olduğu gösterilmiştir. Sonraki aşamada LLM'lerin cevapladığı sorulara verilen cevapların olasılıkları incelendiğinde eşik değerinin modeller üzerindeki etkisi araştırılmıştır. Modellerin analizi sonucunda BERT-TrSQuAD modeli hariç diğer LLM'lerde doğruluk değerinde %0.13'e varan bir artış görülmüştür. Son aşamada ise modeller tarafından tahmin edilen cevapların gerçek cevaplarla anlamsal ilişkisi incelenmiştir. Bu görev için gerçek ve tahmin edilen cevaplar arasındaki değer, anlamsal benzerlik modeli kullanılarak analiz edilmiştir. Bu analiz sonucunda, tüm modellerin doğruluk değerinin 0.7% ile 6.59% arasında arttığı görülmüş ve en yüksek doğruluk değeri %79.09 ile mDeBERTa-TrSQuAD modeli tarafından elde edilmiştir.

Gelecekteki çalışmalarda bu modellerin Türkçe için alan-özgü QA görevleri için karşılaştırılması planlanmaktadır. Ayrıca Türkçe için metin oluşturma, özetleme modelleri oluşturarak literatüre katkıda bulunulması hedeflenmektedir.

Kaynaklar (References)

1. Karanikolas, N., Manga, E., Samaridi, N., Tousidou, E., Vassilakopoulos, M., Large language models versus natural language understanding and generation, In Proceedings of the 27th Pan-Hellenic Conference on Progress in Computing and Informatics, 278-290, 2023.
2. Dwivedi S.K., Singh V., Research and reviews in question answering system, Procedia Technology, 10, 417–424, 2013.
3. Agarwal A., Sachdeva N., Yadav R.K., Udandarao V., Mittal V., Gupta A., Mathur A., Eduqa: Educational domain question answering system using conceptual network mapping, ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing – Proceedings, 2019-May, 8137–8141, 2019.

4. Jin D., Pan E., Oufattole N., Weng W.H., Fang H., Szolovits P., What disease does this patient have? a large-scale open domain question answering dataset from medical exams, Applied Sciences 2021, 11, 6421, 2021.

5. Etezadi R., Shamsfard M., The state of the art in open domain complex question answering: a survey, Applied Intelligence, 53, 4124–4144, 2023.

6. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L., Polosukhin I., Attention is all you need, Advances in Neural Information Processing Systems, 2017-December, 5999–6009, 2017.

7. Devlin J., Chang M.W., Lee K., Toutanova K., Bert: Pre-training of deep bidirectional transformers for language understanding, NAACL HLT 2019-2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies Proceedings of the Conference, 1, 4171–4186, 2018.

8. Improving language understanding with unsupervised learning-OpenAI. https://openai.com/index/language-unsupervised/. Erişim tarihi Temmuz 20, 2024.

9. Angelis L.D., Baglivo F., Arzilli G., Privitera G.P., Ferragina P., Tozzi A.E., Rizzo C., Chatgpt and the rise of large language models: the new ai-driven infodemic threat in public health, Frontiers in Public Health, 11, 2023.

10. Alzubi J.A., Jain R., Singh A., Parwekar P., Gupta M., Cobert: Covid-19 question answering system using bert, Arabian Journal for Science and Engineering, 48, 11003–11013, 2023.

11. Kierszbaum S., Lapasset L., Applying distilled bert for question answering on asrs reports, Proceedings of the 22nd International Conference on New Trends in Civil Aviation, NTCA 2020, 33–38, 2020.

12. Singhal K., Azizi S., Tu T., Mahdavi S.S., Wei J., Chung H.W., Scales N., Tanwani A., Cole-Lewis H., Pfohl S., Payne P., Seneviratne M., Gamble P., Kelly C., Babiker A., Scharli N., Chowdhery A., Mansfield P., DemnerFushman D., Arcas B.A., Webster D., Corrado G.S., Matias Y., Chou K., Gottweis J., Tomasev N., Liu Y., Rajkomar A., Barral J., Semturs C., Karthikesalingam A., Natarajan V., Large language models encode clinical knowledge, Nature, 620, 172–180, 2023.

13. Abdelhay M., Mohammed A., Hefn H.A., Deep learning for arabic healthcare: Medicalbot, Social Network Analysis and Mining, 13, 1–17, 2023.

14. Wang Z., Ng P., Ma X., Nallapati R., Xiang B., Multi-passage bert: A globally normalized bert model for open-domain question answering, EMNLP-IJCNLP 2019-2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, 5878–5882 2019.

15. Qu C., Yang L., Qiu M., Croft W.B., Zhang Y., Iyyer M., Bert with history answer embedding for conversational question answering, SIGIR 2019-Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 1133–1136, 2019.

16. Zhang Z., Wu Y., Zhou J., Duan S., Zhao H., Wang R., Sg-net: Syntax-guided machine reading comprehension, Proceedings of the AAAI Conference on Artificial Intelligence, 34, 9636–9643, 2020.

17. Zhang Z., Wu Y., Zhao H., Li Z., Zhang S., Zhou X., Zhou X., Semantics-aware bert for language understanding, Proceedings of the AAAI Conference on Artificial Intelligence, 2020.

18. Lan Z., Chen M., Goodman S., Gimpel K., Sharma P., Soricut R., ALBERT: A Lite BERT for Self-supervised Learning of Language Representations, arXiv preprint arXiv:1909.11942, 2019.

19. Soygazi F., Çiftçi O., Kök U., Cengiz S., Thquad: Turkish historic question answering dataset for reading comprehension, Proceedings-6th International Conference on Computer Science and Engineering, UBMK 2021, 215–220, 2021.

20. Akyön F., Çavuşoğlu D., Cengiz C., Altinuç S.O., Temizel, A.: Automated question generation and question answering from Turkish texts, Turkish Journal of Electrical Engineering and Computer Sciences, 30, 1931–1940, 2022.

21. İncidelen, M., Aydoğan, M. (2024, Developing Question-Answering Models in Low-Resource Languages: A Case Study on Turkish Medical Texts Using Transformer-Based Approaches, In 2024 8th International Artificial Intelligence and Data Processing Symposium (IDAP), 1-4, 2024.

22. Tutar, K., Yıldız, O.T., Turkish Question-Answer Dataset Evaluated with Deep Learning, In 2024 9th International Conference on Computer Science and Engineering (UBMK), 101-105, 2024.

23. Kahraman S.Y., Durmuşoğlu A., Dereli T., Patent classification with pre-trained Bert model, Journal of the Faculty of Engineering and Architecture of Gazi University, 39 (4), 2484-2496, 2024.

24. Tepecik A., Demir E., Analysis of Turkish audio recording data labeled with three emotions popular machine learning algorithms, Journal of the Faculty of Engineering and Architecture of Gazi University, 39 (2), 709-716, 2024.

25. Budur E., Ozçelik R., Soylu D., Khattab O., Güngör T., Potts C., Building Efficient and Effective OpenQA Systems for Low-Resource Languages, Knowledge-Based Systems, 302, 112243, 2024.

26. Sanh V., Debut L., Chaumond J., Wolf T., Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, arXiv preprint arXiv:1910.01108, 2019.

27. He P., Liu X., Gao J., Chen W., Deberta: Decoding-enhanced bert with disentangled attention, ICLR 2021-9th International Conference on Learning Representations, 2021.

28. Xue L., Constant N., Roberts A., Kale M., Al-Rfou R., Siddhant A., Barua A., Raffel C., mt5: A massively multilingual pre-trained text-to-text transformer, NAACL-HLT 2021-2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 483–498, 2020.

29. Hugging Face. emrecan/bert-base-turkish-cased-mean-nli-stsb-tr. https://huggingface.co/emrecan/bert-base-turkish-cased-mean-nli-stsb-tr . Erişim tarihi Aralık 25, 2024.