Bilimsel Metinlerde Anlamsal Özetleme: Derin Öğrenme Yaklaşımlarının Performans Analizi

Araştırma Makalesi/Research Article

Senanur ÖZCAN[1*], Ahmet ÖZCAN[2]

[1]Yazılım Geliştirme, Kapadokya Üniversitesi, Nevşehir, Türkiye
[2]Yapay Zekâ ve Makine Öğrenmesi, Kapadokya Üniversitesi, Nevşehir, Türkiye

senanur.ozcan@kun.edu.tr, ahmet.ozcan@kapadokya.edu.tr, iletisim@ahmetozcan.com
(Geliş/Received:15.01.2026; Kabul/Accepted:03.07.2026)
DOI: 10.17671/gazibtd.1863483

Özet— Bu çalışmada biyomedikal soyutlayıcı metin özetleme görevinde alan odaklı ve genel amaçlı Transformer tabanlı modeller karşılaştırılmıştır. BART Large CNN, PEGASUS Large, FLAN T5 XL ve BioMistral 7B modelleri, 133120 PubMed makalesinden oluşan hazır bir veri seti üzerinde eğitilmiş ve test edilmiştir. Tüm modeller 1024 token girdi ve 256 token çıktı sınırı ile eğitilmiş, performansları ROUGE 1, ROUGE 2, ROUGE L ve ROUGE Lsum metrikleri kullanılarak değerlendirilmiştir. Sonuçlar, BioMistral 7B modelinin ROUGE 1 değerinde 47.74, ROUGE 2 değerinde 21.08 ve ROUGE L değerinde 27.99 skorları ile en yüksek genel performansı elde ettiğini göstermektedir. Niteliksel değerlendirmeler, BioMistral 7B modelinin biyomedikal terminolojiye daha yüksek uyum sağladığını ve klinik bilgileri daha tutarlı biçimde özetleyebildiğini ortaya koymuştur. Bu bulgular, alan uyarlamalı ön eğitimin biyomedikal soyutlayıcı özetleme görevlerinde model performansını artırdığını göstermektedir.

Anahtar Kelimeler— biyomedikal metin özetleme, soyutlayıcı özetleme, transformer modelleri, biomistral, PEGASUS, BART, FLAN-T5, ROUGE metrikleri

Semantic Summarization of Scientific Texts: Performance Analysis of Deep Learning Approaches

Abstract— This study compares domain oriented and general purpose Transformer based models for biomedical abstractive text summarization. BART Large CNN, PEGASUS Large, FLAN T5 XL and BioMistral 7B models are trained and evaluated on a prepared dataset consisting of 133120 PubMed articles. All models are trained with a maximum input length of 1024 tokens and an output length of 256 tokens, and their performance is evaluated using ROUGE 1, ROUGE 2, ROUGE L and ROUGE Lsum metrics. The results show that the BioMistral 7B model achieves the highest overall performance with ROUGE 1 score of 47.74, ROUGE 2 score of 21.08 and ROUGE L score of 27.99. Qualitative analyses indicate that BioMistral 7B provides better alignment with biomedical terminology and produces more coherent clinical summaries. These findings demonstrate that domain adaptive pretraining improves model performance in biomedical abstractive summarization tasks.

Keywords— biomedical text summarization, abstractive summarization, transformer models, biomistral, PEGASUS, BART, FLAN-T5, ROUGE metrics

1. GİRİŞ (INTRODUCTION)

Son yıllarda yayımlanan bilimsel makale sayısındaki hızlı artış, akademik metinlerin çeşitliliğini ve kapsamını olağanüstü bir şekilde arttırmıştır. Bu artış, araştırmacıların güncel gelişmeleri takip etmesini ve literatürü kapsamlı olarak değerlendirmesini giderek zorlaştırmaktadır. Özellikle biyomedikal araştırmalar gibi hem kavramsal karmaşıklığın hem de terminolojik yoğunluğun yüksek olduğu alanlarda, literatürün hızla genişlemesi bilgiye erişim süreçlerini zorlaştırmaktadır. Örneğin yalnızca PubMed veri tabanında bile her yıl milyonlarca yeni biyomedikal makale yayımlanmakta ve bu yayınların önemli bir kısmı klinik araştırmalar, vaka raporları ve tedavi süreçlerine ilişkin ayrıntılı bilgiler içermektedir. Bu bağlamda, otomatik metin özetleme sistemleri; araştırmacıların uzun metinleri daha verimli taraması, klinik raporları hızlı değerlendirmesi ve bilgiye daha kısa sürede ulaşabilmeleri açısından kritik bir önceme sahiptir.

Metin özetleme yöntemleri genel olarak iki ana yaklaşım altında sınıflandırılmaktadır: çıkarımsal (extractive) ve soyutlayıcı (abstractive) özetleme. Çıkarımsal yöntemler, metinlerdeki en önemli cümleleri seçmeye çalışır bu yüzden bilimsel metinlerde anlam bütünlüğü, mantıksal akış ve bağlamsal tutarlılık gibi kritik unsurları her zaman koruyamamaktadır [1]. Buna karşın soyutlayıcı özetleme, metni derinlemesine analiz ederek içeriği yeni ifadelerle yeniden oluşturur [2]. Transformer mimarisiyle birlikte PEGASUS, BART ve T5 gibi modellerin ortaya koyduğu yüksek başarı, soyutlayıcı özetlemeyi bilimsel metin işleme alanında önemli bir araştırma konusu hâline getirmiştir.

Ancak mevcut soyutlayıcı modellerin büyük çoğunluğu genel amaçlı dil verileri üzerinde eğitildiğinden, biyomedikal alanın kendine özgü terminolojisini ve bilgi yoğunluğunu tam olarak yansıtamamaktadır. Biyomedikal metinlerin analizi için geliştirilen BioBERT [3], BioGPT [4] ve PubMedBERT[5] gibi alan odaklı modeller, sınıflandırma, Adlandırılmış Varlık Tanıma(Named Entity Recognition, NER) ve ilişki çıkarımı (Regular Expression, RE) gibi anlama tabanlı görevlerde genel modellerden daha yüksek performans göstermiştir. Patterns dergisinde yayımlanan güncel çalışmalar, alan uyarlamalı ön eğitimin (domain-adaptive continued pretraining) biyomedikal görevlerde önemli kazanımlar sağladığını, ancak özetleme gibi soyutlayıcı görevlerde hâlen araştırma boşlukları bulunduğunu ifade etmektedir [6].

Bu bağlamda, Mistral mimarisinin biyomedikal literatür üzerinde ek ön eğitimle dönüştürülmüş hâli olan BioMistral, alan uyumlu bilgi temsil kabiliyeti ve soyutlayıcı dil modelleme kapasitesiyle dikkat çekmektedir. BioMistral ve benzeri modeller, biyomedikal metin işleme görevlerinde dikkat çekici başarılar göstermiş olsa da bu modellerin soyutlayıcı özetleme bağlamında PubMed veri seti üzerinde sistematik biçimde değerlendirildiği çalışma sayısının sınırlı olduğu görülmektedir. Bu çalışma, literatürdeki bu eksikliği gidermeyi amaçlayarak, BioMistral modelinin PubMed üzerindeki özetleme performansını PEGASUS, FLAN-T5-XL ve BART-Large-CNN gibi güçlü modeller ile sistematik biçimde karşılaştırılmaktadır. Böylece çalışmanın temel amacı, alan odaklı büyük dil modellerinin genel amaçlı modellere kıyasla biyomedikal özetleme görevlerinde anlamsal bütünlülük, bağlamsal tutarlılık ve kavramsal doğruluk açısından ne ölçüde üstünlük sağlayabileceğini göstermektir.

2. LITERATÜR ANALİZİ (LITERATURE REVIEW)

Bu bölüm, metin özetleme alanındaki temel yöntemleri bütüncül bir çerçevede incelemekte; çıkarımsal ve soyutlayıcı yaklaşımları, Transformer tabanlı modellerin yükselişini ve bilimsel doküman özetlemesini diğer metin türlerinden ayıran yapısal özellikleri ele almaktadır.

2.1. Çıkarımsal Özetleme Yaklaşımları
(Extractive Summarization Approaches)

Çıkarımsal özetleme, metindeki bilgi taşıyıcı cümlelerin belirli ölçütlere göre seçilmesiyle özet oluşturmayı amaçlayan en eski ve en yaygın yaklaşımlardan biridir. Bu yöntemde cümle seçimi benzerlik ölçümleri, kapsama (coverage), tekrarın azaltılması (redundancy reduction) ve içerik önceliklendirmesi gibi kriterlere dayanır. Literatürde özellikle çok-belgeli özetleme çalışmalarında maksimum kapsama ve minimal tekrar ilkelerine dayalı yöntemlerin etkili sonuçlar verdiği gösterilmiştir [7], [8].

Çıkarımsal yaklaşımın uygulamada karşılık bulan örnekleri arasında, istatistiksel özelliklere dayalı cümle ağırlıklandırma yöntemleri yer almaktadır. Yapılan çalışmalar tek-belgeli özetleme için cümleleri istatistiksel puanlama ile sıralayan bir cümle sıralama (sentence-ranking) metodu önermiş ve kısa metinlerde tutarlı özetler üretebildiğini göstermiştir [9]. Yadav ve arkadaşları tarafından yapılan güncel bir incelemede ise çıkarımsal yöntemleri istatistiksel, graf tabanlı, optimizasyon temelli, girdi gömme (embedding) tabanlı ve derin öğrenme tabanlı modeller olmak üzere çeşitli kategoriler altında sınıflandırarak güncel eğilimleri ve karşılaşılan sınırlılıkları kapsamlı biçimde ortaya konmuştur [10]. Bu çalışmalar, çıkarımsal özetlemenin önemli avantajlarına rağmen, cümleleri doğrudan kopyalama yapısı nedeniyle özellikle uzun ve karmaşık bilimsel metinlerde özet üretmede sınırlı kaldığını göstermektedir.

2.2. Soyutlayıcı Özetleme Yaklaşımları
(Abstractive Summarization Approaches)

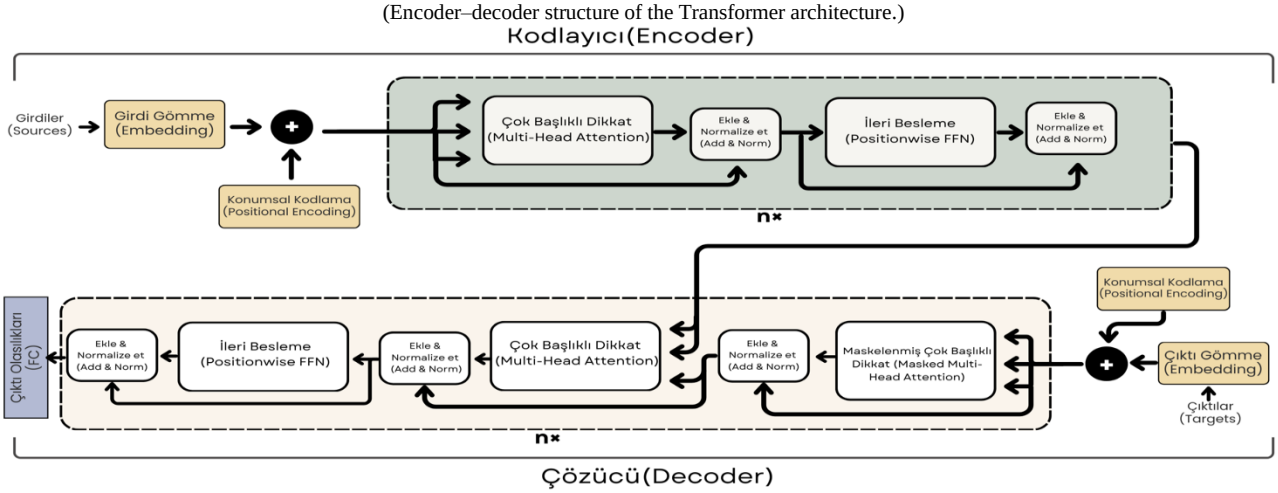
Çıkarımsal yöntemlerin metni doğrudan kopyalama yapıları nedeniyle bağlamsal bütünlüğü tam olarak yansıtamaması, literatürde anlamı yeniden ifade eden soyutlayıcı özetleme yaklaşımlarının gelişmesine zemin hazırlamıştır. Soyutlayıcı yöntemler, metni yalnızca seçilmiş cümlelerden oluşturmak yerine yeni ifadeler ve cümle yapıları üreterek insan benzeri özetler oluşturmayı amaçlamaktadır. İlk soyutlayıcı özetleme çalışmaları, kodlayıcı-çözücü (encoder–decoder) mimarisine dayalı RNN/LSTM tabanlı seq2seq modelleri kullanmış;

attention mekanizmasının eklenmesi bu modellerin önemli ilişkileri yakalama kapasitesini belirgin biçimde artırmıştır [11]. Ardından geliştirilen pointer–generator mimarisi, hem kaynaktan kopyalama hem de yeni içerik üretimi yapabilmesi sayesinde eğitimde görülmeyen kelimeler (out-of-vocabulary, OOV) problemini azaltmış ve özetlerin akıcılığını iyileştirmiştir [12].

tarafındaki çok katmanlı dikkat mekanizmaları sayesinde bilgi kaybı olmadan sağlanır [17]. Transformer'ın ölçeklenebilir tasarımı, PEGASUS BART ve T5 gibi modellerde anlamsal açıdan tutarlı ve yüksek kaliteli özet üretimi sunmasını sağlar, böylece mimari güncel özetleme yöntemlerinin temel yapı taşı hâline gelir [18].

PEGASUS, gap-sentence generation (boşluklu cümle üretimi) adı verilen ön eğitim stratejisi sayesinde metindeki en bilgilendirici cümleleri maskeleyerek özetleme görevine doğrudan uyumlu bir temsil öğrenmektedir [16].

Şekil 1. Transformer mimarisinin kodlayıcı-çözücü yapısı
(Encoder–decoder structure of the Transformer architecture.)



2.2.1 Transformer Tabanlı Özetleme Yaklaşımları (Transformer-based Summarization Approaches)

Soyutlayıcı özetlemede kaydedilen en büyük gelişme Transformer mimarisinin ortaya çıkması ile gerçekleşmiştir [13]. Kodlayıcı-çözücü yapısı ve tamamen self-attention mekanizmasına dayanması sayesinde Transformer, uzun bağımlılıkların modellenmesi büyük ölçekli ön-eğitimle uyumluluğu açısından yinelemeli sinir ağı (Recurrent Neural Network, RNN) ve evrişimsel sinir ağı (Convolutional Neural Network, CNN) tabanlı önceki yaklaşımlara kıyasla önemli bir üstünlük sağlamıştır. Çok başlıklı dikkat (Multi-head attention) katmanları ve konumsal kodlama yapıları, modelin bağlamsal ilişkileri daha etkili öğrenmesini sağlamış; böylece Transformer tabanlı modeller kısa sürede metin özetleme görevlerinde standart hâline gelmiştir [14]. Ayrıca bu mimarinin verileri paralel işleyebilmesi, büyük veri üzerinde çalışan ön eğitimli dil modellerinin ortaya çıkmasını ve yüksek performanslı özetler üretilmesini mümkün kılmıştır [15]. Kodlayıcı ve çözücü bileşenlerinden oluşan Transformer mimarisinin genel yapısı Şekil 1'de sunulmaktadır.

Transformer mimarisi, metni işlemeye girdi aşamasında gömme ve konumsal kodlama bileşenleriyle başlar; ardından çok başlıklı dikkat katmanları üzerinden farklı bağlamsal ilişkilerin paralel olarak öğrenilmesini sağlar. Bu paralel yapı, RNN tabanlı modellerdeki sıralı hesaplama zorunluluğunu ortadan kaldırarak daha verimli çalışmasını sağlar [16]. Özellikle uzun metinlerin bağlamının korunması hem kodlayıcı hem çözücü

Ancak temel model çoklu doküman ve uzun metin özetlemede önemli içerik seçimi ve gereksiz tekrar (redundancy) açısından sınırlılıklar göstermektedir. Bu eksiklikleri gidermek üzere geliştirilen PEGASUS-XL mimarisi maksimum marjinal alaka (Maximal Marginal Relevance, MMR) ile tekrar azaltımı ve Longformer tabanlı uzun girdi kodlaması sayesinde hem içerik seçimi hem de bağlamsal tutarlılık açısından anlamlı iyileştirmeler sağlamış ve güçlü modelleri geride bırakan sonuçlar elde etmiştir [17]. Ayrıca tıbbi alan özelinde yapılan çalışmalar, PEGASUS ailesinin sınırlı veriyle bile tutarlı özetler üretebildiğini göstermekte [19]; farklı veri setlerinde yapılan ince ayar (fine-tuning) çalışmaları ise modelin ROUGE tabanlı metriklerde belirgin performans artışları sağlayabildiğini ortaya koymaktadır [20].

BART modeli maskeleme (masking), silme (deletion) ve metin içi boşluk doldurma (text infilling) gibi çeşitli denoising (bozulma) stratejileriyle eğitildiği için metnin bağlamsal bütünlüğünü yeniden yapılandırmada güçlü performans göstermektedir. Uzun ve bilgi açısından yoğun olan metinlerde yaşanan token sınırı sorununa yönelik olarak yapılan bir çalışmada, metinlerin semantik bütünlüğü korunarak parçalara ayrıldığı ve her bir parçanın ayrı ayrı özetlenip daha sonra birleştirildiği bir yaklaşımın BART'ın PubMed gibi uzun bilimsel dokümanlarda ROUGE-1/2 skorlarını belirgin şekilde artırdığını göstermiştir [21]. Yapılan başka bir çalışmada semantik benzerlik temelli değerlendirmeler, BART'ın ürettiği özetlerin yalnızca n-gram örtüşmesine değil, aynı zamanda metnin özünü ve anlamsal yapısını koruma açısından da yüksek doğruluk sağladığını ortaya koymaktadır [22].

T5 modeli, tüm Doğal Dil İşleme (Natural Language Processing, NLP) görevlerini metinden metne çeviri (text-to-text) paradigması altında birleştiren bir mimaridir [23]. Bu yapı, Transformer mimarisinin attention tabanlı hesaplama prensipleri üzerine inşa edilmiştir [13]. Google tarafından geliştirilen yaklaşık 3 milyar parametreli FLAN-T5-XL modeli, talimat temelli ince ayar stratejisiyle kullanılmakta olup çeşitli metin işleme görevlerinde yüksek genelleme performansı sergilemektedir [24]. Güncel çalışmalar, T5 tabanlı modellerin farklı veri setlerinde uygulanan ince ayar süreçleri sonucunda ROUGE tabanlı ölçütlerde belirgin performans artışı sağladığını göstermekte [25]; ayrıca yarı denetimli öğrenme yaklaşımlarının ve görev ön-ekleriyle (prefix-controlled) yönlendirilen üretim mekanizmalarının modelin bağlamsal tutarlılığını güçlendirdiği raporlanmaktadır [26]. Buna ek olarak, farklı veri setleri üzerinde yapılan karşılaştırmalı analizler, T5 modellerinin hem kısa hem uzun metin özetleme görevlerinde rekabetçi sonuçlar verdiğini ortaya koymaktadır [27].

2.3. Bilimsel Makale Özetleme Çalışmaları
(Scholarly / Scientific Document Summarization)

Bilimsel makalelerin özetlenmesi, genel amaçlı metin özetleme görevlerinden farklı olarak yapısal bütünlük (Giriş, Yöntem, Bulgular ve Tartışma; Introduction, Methods, Results, Discussion, IMRaD), teknik terminoloji, uzun bağlam ve disiplinler arası bilgi yoğunluğu gibi ek zorluklar içermektedir. İlk çalışmalar, bilimsel yayınlardaki yenilik ve katkı bölümlerinin otomatik olarak çıkarılmasına odaklanarak katkı odaklı özetleme (contribution-focused) yaklaşımlarını ortaya çıkartmıştır [28]. Ardından metinsel modeller, bilimsel makalelerin bölüm yapısını dikkate alarak daha tutarlı ve bölüm uyumlu özetler üretmeyi amaçlamış ve IMRaD yapısına dayalı bölüm farkındalığının özet kalitesini belirgin şekilde artırdığı gösterilmiştir [29].

Bilimsel özetleme çalışmalarında kullanılan modeller kadar veri setlerinin niteliği de önemli bir araştırma konusu olarak öne çıkmaktadır. Veri setlerinin kapsam, dağılım ve önyargı yönünden heterojen yapılar barındırdığı; bu durumun modellerin genelleme performansını etkileyebildiği çeşitli analizlerle ortaya konmuştur [30]. Yapılan bir çalışmada, biyomedikal özetleme alanında Transformer tabanlı modellerin son yıllarda öne çıktığı rapor edilmektedir [31]. Bu çalışma, özetleme yöntemlerinin büyük bölümünün hâlâ tek-belgeli ve çoğunlukla PubMed tabanlı biyomedikal literatür üzerinde yürütüldüğünü ve ROUGE'un en yaygın değerlendirme metriği olduğunu rapor etmektedir. Bu konu ile ilgili yapılan bir çalışmada ise bilimsel makalelerin özetlenmesinde yalnızca temel içeriğin değil, çalışmanın literatüre yaptığı özgün katkının da ayrı ve doğru biçimde yakalanması gerektiğini göstermektedir [32]. Özellikle PubMed gibi yüksek bilgi yoğunluklu alanlarda BioMistral ve benzeri Transformer modellerinin bağlamsal anlamı yeniden yapılandırma kapasitesinin neden kritik olduğunu açıklamaktadır.

2.4. Alan Odaklı Dil Modelleri
(Domain-Specific / Biomedical Language Models)

Genel amaçlı dil modellerinin biyomedikal metinlerde teknik terminoloji, karmaşık sözdizimi ve yüksek bilgi yoğunluğu nedeniyle performans kaybı yaşadığı, literatürde tutarlı şekilde gösterilmiştir [33], [34]. Bu nedenle alan odaklı modeller geliştirilmiş ve önemli ilerlemeler kaydedilmiştir. BioBERT [6], PubMed ve PMC üzerinde yapılan ön-eğitimi sayesinde NER ve RE gibi görevlerde genel modelleri belirgin biçimde aşmıştır. BioGPT [4], jeneratif kapasitesi sayesinde klinik bilgi üretimi ve tıbbi doküman sınıflandırmada güçlü sonuçlar vermiştir. SCIBERT gibi bilimsel söylem temelli modeller, akademik makale dilinin yapısal özelliklerini daha tutarlı biçimde yakalayarak bilimsel alanlara özgü görevlerde yüksek doğruluk sağlamıştır [35]. Uzun bağlam işleyebilen ModernBERT ve benzeri yeni modeller ise klinik rapor özetleme gibi uzun giriş gerektiren görevlerde bağlamsal bütünlüğü koruma açısından dikkat çekici iyileşmeler sunmuştur [36]. Son olarak, büyük dil modellerinin biyomedikal veri üzerinde ince ayarlama yapıldığında ciddi performans artışı sağladığı literatürde ortaya konmuş, alan-uyarlamalı(domain-adaptive) eğitimin bu alanda kritik bir öneme sahip olduğu vurgulanmıştır [6]. Tüm bu çalışmalar, alan odaklı dil modellerinin bilimsel ve klinik dokümanlarda hem doğruluk hem bağlamsal tutarlılık açısından vazgeçilmez hâle geldiğini göstermektedir.

Büyük dil modellerinin tıbbi alana uyarlanmasına yönelik çalışmalar son yıllarda artmış olsada, açık kaynak ve geniş kapsamlı biyomedikal modeller hâlâ sınırlıdır. Bu boşluğu hedefleyen BioMistral, Şubat 2024'te tanıtılmış olup Mistral 7B mimarisi üzerinde, çok kaynaklı tıbbi veri kümeleriyle eğitilen ilk büyük ölçekli biyomedikal model ailesi olarak öne çıkmaktadır [37]. Mevcut araştırmalar, Mistral tabanlı modellerin PubMedQA ve benzeri biyomedikal değerlendirme setlerinde etkili sonuçlar verebildiğini göstermektedir; örneğin Léon ve arkadaşları, Mistral 7B'nin LoRA tabanlı uyarlamalarla tıbbi soru-cevap görevlerinde güçlü performans sergilediğini raporlamıştır [38]. Buna karşılık BioMistral gibi alan-özel olarak eğitilmiş modeller, tıbbi terminolojiye hâkimiyetleri ve uzun bağlam içinde kavramsal tutarlılığı sürdürebilme yetenekleri sayesinde, bilimsel metinlerde bağlam bütünlüğünü koruma ve kritik bilgileri doğru biçimde yeniden üretme noktasında genel modellerden daha güvenilir sonuçlar sunmaktadır [39]. Özellikle uzun bilimsel ve klinik metinlerde, büyük dil modellerinin bağlamın orta bölümlerini düşük oranda kullandığı ve bu nedenle özetleme performansının ciddi biçimde etkilenebildiği gösterilmiştir [40]. Öte yandan klinik uzun metin özetleme üzerine yapılan güncel çalışmalar, sıfırdan ayarlama (zero-shot) senaryolarında bile, alan-odaklı modellerin zamansal çıkarım ve kavramsal tutarlılık açısından genel modellerden daha güvenilir çıktılar üretebildiğini ortaya koymaktadır [41]. Bu bulgular değerlendirildiğinde, BioMistral'in biyomedikal özetleme için önemli bir potansiyel sunduğu görülmektedir.

Tablo *1*. Literatürde incelenen yaklaşım ve yöntemlerin karşılaştırılması
(Comparison of approaches and methods reviewed in the literature)

YAKLAŞIM	KAYNAK	KULLANILAN YÖNTEM	TEMEL BULGU / KATKI
Çıkarımsal Özetleme	[7], [8]	Maksimum kapsama, tekrarın azaltılması ilkeleri	Çok-belgeli özetlemede etkili sonuçlar
	[9], [10]	Cümle puanlama yöntemi; mevcut çıkarımsal yöntemler istatistiksel, graf tabanlı, optimizasyon ve derin öğrenme temelli	Kısa metinlerde tutarlı özetler; mevcut yöntemlerin kapsamlı sınıflandırması ve sınırlılıkları
Soyutlayıcı Özetleme	[11]	Attention mekanizmalı encoder–decoder RNN/LSTM mimarisi	Uzun metinlerdeki önemli ilişkileri yakalama oranı artmıştır
	[12]	Pointer–generator mimarisi	OOV probleminin azaltılması, akıcılığın iyileşmesi
Transformer Tabanlı	[13]	Self-attention, encoder–decoder mimarisi	Transformer'ın temel mimarisi
	[16], [17], [19], [20]	PEGASUS ailesi: Gap-Sentence Generation ön eğitimi, MMR + Longformer tabanlı uzun girdi kodlama sınırlı veriyle ve farklı veri setleriyle ince ayar	Özetlemeye özgü temsil öğrenme, içerik seçimi/bağlamsal tutarlılıkta iyileşme ve ROUGE artışı
	[21], [22]	BART tabanlı çalışmalar: parçalara ayırma + ayrı özetleyip birleştirme, masking/deletion/text infilling	PubMed'de ROUGE-1/2 artışı; anlamsal yapıyı koruma ve n-gram ötesi doğruluk
	[23], [24], [25],[26],[27]	T5/FLAN-T5 ailesi: metinden metne çeviri, talimat temelli ince ayar , ince ayar ve yarı denetimli öğrenme yaklaşımları	Görev birleştirme, yüksek genelleme performansı, ROUGE artışı ve bağlamsal tutarlılık
Bilimsel Makale Özetleme	[28], [32]	Katkı odaklı özetleme ve özgün katkı tespiti	Yenilik/katkı bölümlerinin ayrı ve doğru biçimde yakalanması
	[29]	Discourse-aware attention, IMRaD yapı farkındalığı	Bölüm uyumlu özet kalitesinde artış
	[30], [31]	Veri seti analiz çalışması ve sistematik derleme	Veri setlerindeki önyargının model genellemesini etkilediği; Transformer tabanlı modellerin biyomedikal özetlemede son yıllarda öne çıkması
Alan Odaklı Modeller	[3], [4], [35], [36]	Alan-özel önceden eğitilmiş modeller: BioBERT, BioGPT, SciBERT, ModernBERT (PubMed/PMC ön eğitimi, jeneratif ön eğitim, bilimsel söylem/uzun bağlam encoder)	NER ve RE görevlerinde üstünlük, klinik bilgi üretimi, akademik dilde yüksek doğruluk ve bağlamsal bütünlük
	[6]	Alan-uyarlamalı sürekli ön eğitim	Biyomedikal görevlerde performans artışı
	[37], [38]	Mistral tabanlı biyomedikal modeller: çok kaynaklı tıbbi veriyle ek ön eğitim (BioMistral), LoRA ile Mistral-7B ince ayar	İlk büyük ölçekli açık kaynak biyomedikal LLM; PubMedQA'da güçlü performans
	[39], [40], [41]	Karşılaştırmalı değerlendirme, bağlam kullanım analizi ve sıfırdan ayarlama zamansal çıkarım (temporal reasoning) çalışmaları	Alan-özel modellerin terminoloji avantajı; uzun metinlerde orta bölüm bilgisinin az kullanılması; sıfırdan ayarlamada üstünlük

3. YÖNTEM (METHOD)

Bu bölümde, çalışmada uygulanan veri işleme adımları, model mimarileri, ince ayar süreci ve değerlendirme yöntemleri ayrıntılı olarak açıklanmaktadır. Çalışmanın genel işlem akışı Şekil 2'de sunulmuştur.

Şekil 2. Özetleme deneylerinde izlenen süreç
(Workflow followed in the summarization experiments)

3.1. Veri Seti (Dataset)

Bu çalışmada deneysel değerlendirmeler için, bilimsel metin özetleme alanında yaygın olarak kullanılan PubMed Article Summarization veri seti kullanılmıştır [42]. Veri seti toplam 133.120 biyomedikal makaleden oluşmakta olup; 119.832 makale eğitim (train), 6.632 makale doğrulama (validation) ve 6.656 makale test bölümü olarak ayrılmıştır. PubMed veri setindeki referans özetler, makalelerin arka planını, amacını, kullanılan yöntemleri, elde edilen bulguları ve ulaşılan sonuçları sistematik biçimde sunan yapılandırılmış bir formatta hazırlanmıştır. Bu yapı, modellerin yalnızca yüzeysel bilgi sıkıştırması değil, makalenin mantıksal akışını öğrenmesini de mümkün kılmaktadır.

Veri kümesi, biyomedikal alanlarda yayımlanan hakemli akademik makalelerin tam metin gövdelerini ve bu makalelere ait yapılandırılmış referans özetleri içermektedir. Ortalama bir kayıt, 2.000–4.000 kelimelik bir makale gövdesi ve 150–250 kelimelik bir referans özetten oluşmaktadır. Veri seti; onkoloji, kardiyoloji, immünoloji, nörobilim, farmakoloji ve moleküler biyoloji gibi farklı klinik ve biyomedikal alt alanlardan makaleler içerdiği için modellerin çeşitli konu alanlarında genelleme performansını değerlendirmeye uygundur.

Tüm deneylerde aynı eğitim, doğrulama ve test bölünmesi kullanılmıştır. Tokenizasyon analizleri, makale gövdelerinin ortalama 3.500–5.000 token, referans özetlerin ise ortalama 180–250 token uzunluğunda olduğunu göstermektedir. Bazı makale gövdeleri 10.000 token'i aşarak büyük dil modellerinin maksimum giriş sınırlarını zorlamaktadır. Bu nedenle uzun makalelerde truncation (kırpma) stratejisi uygulanmış ve her makalenin yalnızca ilk 1.024 token'i modele giriş olarak alınmıştır. Böylece tüm modeller eşit uzunluktaki girdilerle eğitilmiş ve karşılaştırmalar adil koşullar altında gerçekleştirilmiştir.

3.1.1. Veri Temizleme (Data Cleaning)

Ham PubMed dosyaları incelendiğinde, özellikle CSV yapısını bozan biçimsel hatalar tespit edilmiştir. Bu hatalar; satır kırılmaları, eksik veya fazla sütunlu kayıtlar, hatalı kaçış dizileri, kontrol dışı özel karakterler ve tırnaklama–virgül uyumsuzlukları gibi nedenlerle bazı fonksiyonların düzgün çalışmasını engellemiştir. Bu nedenle verinin bütünlüğünü koruyarak bu teknik sorunları gidermek için özel bir veri temizleme süreci uygulanmıştır. Bu kapsamda, standart pandas fonksiyonları yerine Python'un daha esnek csv kütüphanesi kullanılarak iki temel standartlaştırma adımı gerçekleştirilmiştir:

Alan Limitinin Yükseltilmesi: PubMed makalelerinin uzun gövde metinlerinin sorunsuz okunabilmesi için Python'un satır başına okuma limiti sys.maxsize değeri kullanılarak maksimum düzeye çıkarılmıştır. Böylece varsayılan 128 KB sınırının yol açtığı satır kırılmaları ve okuma hataları tamamen ortadan kaldırılmıştır.

Yeniden Biçimlendirme: Tüm kayıtlar, sütunların zorunlu olarak tırnak karakterleriyle çevrelendiği tutarlı bir CSV formatına dönüştürülmüştür. Bu işlem, gelecekte ortaya çıkabilecek tüm virgül–kaçış ve hücre kayması sorunlarını yapısal olarak ortadan kaldırmıştır.

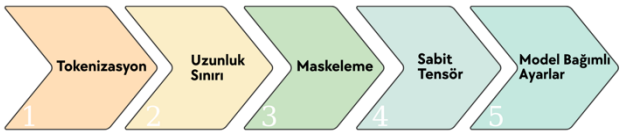
Bu standartlaştırma süreci sırasında veri kaybı minimum tutulmuştur. Train setinde yalnızca bir adet satır, sütun yapısının başlık ile onarılamayacak derecede tutarsız olması nedeniyle çıkarılmıştır. validation.csv ve test.csv dosyalarında ise hiçbir satır silinmemiştir.

Sonuç olarak bu veri temizleme adımları, veri setinin içerik bütünlüğünü koruyarak sadece teknik bozulmaları gidermiş ve modellerin eğitimi için güvenilir, tutarlı ve standartlaştırılmış bir veri kaynağı oluşturmuştur.

3.1.2. Ön İşleme Adımları (Preprocessing Steps)

Veri temizlendikten sonra, tüm modellerin aynı girdi formatını alabilmesi için ortak bir ön işleme boru hattı oluşturulmuştur. Bu adımlar, büyük dil modellerinin veriyi tutarlı ve standart bir biçimde işlemesini sağlamak için gereklidir. Uygulanan ön işleme süreci Şekil 3'te gösterilmiştir.

Şekil 3. Ön işleme süreci (Preprocessing Process)



Tokenizasyon : Her makale gövdesi ve özet, ilgili mimariye özgü BARTTokenizer, PegasusTokenizer, T5Tokenizer, MistralTokenizer olarak ayrıştırılmıştır. Bu sayede her model kendi kelime dağarcığı ve alt-birim (subword) yapısına uygun şekilde veri almıştır.

Maksimum Uzunluk Sınırı (Truncation & Padding): Farklı uzunluklardaki biyomedikal metinlerde eğitimin istikrarlı olması için sabit uzunluklar kullanılmıştır bu uzunluklar Girdi olarak (body text): 1024 token, çıktı olarak (abstract): 256 token kullanılmıştır. Bu sınırı aşan örnekler truncation (maksimum uzunluğa göre kesme), kısa olanlar ise padding (token ekleme) yöntemiyle işlenmiştir.

Maskeleme (Loss Hesabı İçin −100 Maskesi): Hedef özetlerdeki doldurma token'ları öğretici bilgi içermediği için kayıp fonksiyonuna dahil edilmemiştir. Bu amaçla padding token'ları −100 ile maskelenmiştir. Bu yöntem HuggingFace seq2seq modelleri için standart bir uygulamadır.

Sabit Tensör Formatı: Tüm modeller eğitim boyunca sabit boyutlu tensörler üzerinden çalıştırılmıştır. Bu durum hem GPU verimliliği arttırmış hem de tüm modellerin tamamen eşit şartlarda eğitilmesini sağlamıştır.

Model Bağımlı Ayarlar: Mimarilerin özel gereksinimleri göz önünde bulundurulmuştur:

FLAN-T5: Özel mask token yapısı korunmuştur.
PEGASUS: GAP pretraining ile uyumlu hedef özet tokenizasyonu korunmuştur.
BART: Kodlayıcı-çözücü denoising yapısına uygun input-target standardizasyonu sağlanmıştır.
BioMistral: Geniş (Byte-Pair Encoding ,BPE) kelime dağarcığına uygun unicode temizliği yapılmıştır.
Bu ortaklaştırılmış ön işleme süreci sayesinde tüm modeller, birbirinden bağımsız yapıda olmasına rağmen aynı veri akışı ile eğitilmiş ve karşılaştırmalar bir bütün halinde gerçekleştirilmiştir.

3.2. Model Mimarileri (Model Architectures)

Biyomedikal metin özetleme görevinde farklı Transformer mimarilerinin etkisini incelemek amacıyla dört model değerlendirilmiştir. Modeller, hem kodlayıcı-çözücü hem de yalnızca çözücü tabanlı (decoder-only) yapıları temsil ederek mimari çeşitliliğin özetleme başarısına etkisini gözlemlemeye olanak sağlamaktadır. Kodlayıcı-çözücü modellerinde giriş metni kodlayıcı tarafından bağlamsal bir temsile dönüştürülürken, çözücü bu temsili kullanarak özet üretir. Yalnızca çözücü tabanlı modeller ise metni soldan sağa olarak üreterek farklı bir üretim dinamiği sunmaktadır.

3.2.1. Model Teknik Özellikleri (Technical Specifications)

Kullanılan modeller tokenizer altyapıları ve kelime dağarcıkları bakımından farklılık göstermektedir. BART modeli Roberta tabanlı BPE tokenizer kullanırken, PEGASUS SentencePiece'in Unigram varyantıyla çalışmaktadır. FLAN-T5-XL modeli T5 mimarisine özgü SentencePiece tabanlı tokenizer kullanmakta, BioMistral-7B ise LLaMA/Mistral ailesiyle uyumlu geniş kapsamlı bir BPE kelime dağarcığına sahiptir. Bu teknik farklılıklar, modellerin alt-birim düzeyindeki bilgi yakalama kapasitesini ve kelime parçalama stratejilerini etkilemektedir.

3.2.2. Ön Eğitim Stratejileri (Pre-training Strategies)

Her modelin ön-eğitim süreci, özetleme görevine aktarılabilir bilgi miktarını ve alan uyumluluğunu belirleyen temel etkendir. BART, denoising autoencoder yaklaşımıyla bozulmuş metni onarma üzerine eğitilmiş olup, metinsel yapıyı güçlü şekilde modelleyen ancak biyomedikal alan için doğrudan optimize edilmemiş bir yapıya sahiptir. PEGASUS, Gap-Sentence Generation (GAP) mekanizmasıyla özetleme görevine en yakın pretraining stratejisini sunar; metnin en "özet değerindeki" cümlelerini maskelerek modelin bu cümleleri yeniden üretmesini öğretir. FLAN-T5, span corruption pretraining'ini instruction tuning ile birleştirerek çok görevlilik (multi-task adaptation) avantajı sağlar ve yeni görevleri hızlı şekilde öğrenebilir. BioMistral-7B ise Mistral modelinin biyomedikal literatür üzerinde ek pretraining uygulanmış sürümüdür; bilimsel terimleri, klinik ifadeleri ve alan-odaklı (domain-spesifik) yapıları doğrudan modelin parametrelerine işlemiştir.

3.3. İnce Ayar Süreci (Fine-Tuning Process)

Bu çalışmada kullanılan modeller, mimari kapasiteleri ve donanım gereksinimleri doğrultusunda farklı ince ayar stratejileriyle optimize edilmiştir. PEGASUS-Large ve BART-Large modelleri tam ince ayar ile güncellenirken, FLAN-T5-XL modeli parametre-verimli LoRA yaklaşımı, BioMistral-7B ise yüksek bellek gereksinimleri nedeniyle QLoRA yöntemiyle eğitilmiştir. Böylece her model, donanım sınırlamalarıyla uyumlu biçimde optimize edilmiş ve adil bir karşılaştırma sağlanmıştır.

3.3.1. Eğitim Ortamı ve Hiperparametreler (Training Environment and Hyperparameters)

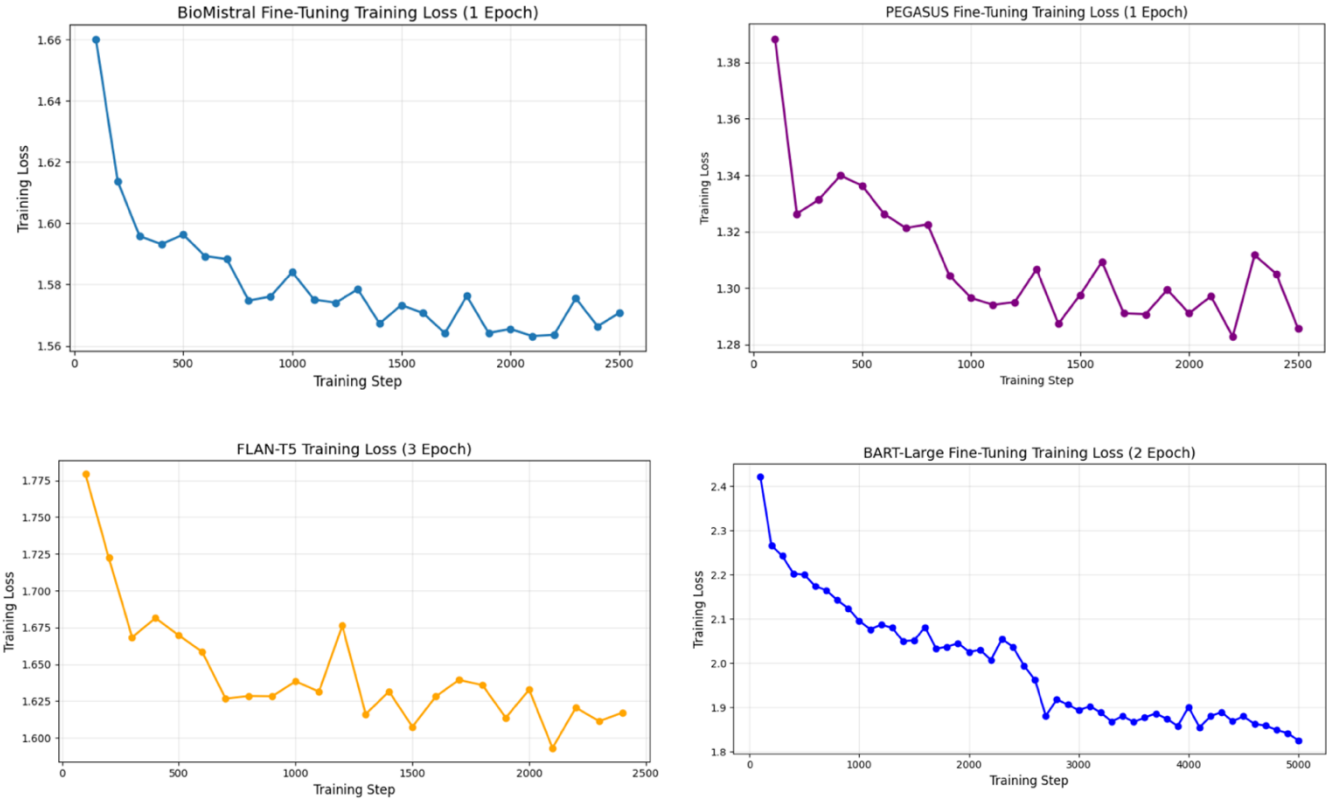
Tüm eğitimler NVIDIA A100 80GB GPU üzerinde, ortak yazılım bileşenleri kullanılarak gerçekleştirilmiştir. Modeller arasında tutarlı bir yapı oluşturmak amacıyla temel hiperparametreler sabit tutulmuştur. Öğrenme oranı 2e-5, optimizasyon algoritması olarak AdamW, maksimum giriş ve çıktı uzunlukları sırasıyla 1024 ve 256 token olarak belirlenmiştir. Büyük modellerin bellek

gereksinimlerini karşılamak için batch size değerleri gradient accumulation tekniğiyle artırılmış; BioMistral-7B ve FLAN-T5-XL için efektif batch size 16, daha küçük modeller için ise 32 olacak şekilde yapılandırılmıştır. Bu ortaklaştırılmış ayarlar, farklı modellerin eşit koşullarda değerlendirilmesini sağlamıştır.

3.3.2. Eğitim Dinamikleri (Training Dynamics)

Tüm modellerde aşırı öğrenme (overfitting) eğilimi gözlenmemiş olup, eğitim kaybı değerlerinin kararlı bir seviyeye ulaştığı noktada süreç sonlandırılmıştır. BioMistral-7B ve PEGASUS-Large modellerinin hızlı stabilizasyon göstermesi, alan-uyarlamalı ön eğitim yaklaşımlarının biyomedikal özetleme görevine güçlü bir başlangıç sağladığını göstermektedir. FLAN-T5-XL modeli ise talimat temelli ince ayar (instruction-tuning)tabanlı yapısı sayesinde kısa sürede uyum sağlayarak dengeli bir yakınsama eğrisi sergilemiştir. Buna karşılık BART-Large modeli, genel amaçlı metinler üzerinde eğitildiğinden biyomedikal içerikle uyumlanmak için daha uzun bir yakınsama sürecine ihtiyaç duymuştur.

Şekil 4. Dört modelin eğitim kaybı eğrileri: (a) BioMistral-7B (b) PEGASUS-Large (c) FLAN-T5-XL (d) BART-Large.
(Training loss curves of the four models: (a) BioMistral-7B (b) PEGASUS-Large (c) FLAN-T5-XL (d) BART-Large)



3.3.3. Eğitim Davranışı ve Epoch Optimizasyonu (Training Behavior and Epoch Optimization)

Bu çalışmada kullanılan dört modelin ince ayar sürecinde gözlemlenen eğitim kaybı eğrileri şekil 4'te gösterilmiştir. (a) BioMistral-7B modeli biyomedikal pretraining'e sahip olduğu için PubMed metinlerine doğal uyum sergilemiş, hızlı bir yakınsama göstermiştir. (b) PEGASUS-Large özetleme görevine özgü GAP pretraining avantajı ile erken stabilizasyona ulaşmıştır. (c) FLAN-T5-XL üç epoch sonunda kararlı bir kayıp düzeyi sergilemiştir. (d) BART-Large modeli ise diğer modellere kıyasla pretraining'i haber metinlerine dayalı olduğu için daha uzun bir eğitim sürecinde yakınsamıştır.

Modeller mimari olarak farklı oldukları için sabit epoch sayısı kullanmak uygun görülmemiştir. Bunun yerine, her modelin eğitim boyunca gösterdiği training loss davranışı incelenmiş ve kendi yakınsama noktalarında durdurulmuş; böylece her modelin en iyi hâli karşılaştırmaya dahil edilmiştir.

3.4. Değerlendirme (Evaluation)

Modellerin çıktı kalitesini ölçmek ve farklı mimariler arasındaki performans farklarını nesnel olarak karşılaştırmak için, metin özetleme alanında yaygın olarak kullanılan otomatik değerlendirme metrikleri uygulanmıştır. Değerlendirme süreci, standartlaştırılmış bir test kümesi üzerinde yürütülmüş ve tüm modeller aynı koşullar altında karşılaştırılmıştır.

3.4.1. Kullanılan Metrikler (Evaluation Metrics)

Bu çalışmada model performansını ölçmek için ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metrikleri kullanılmıştır. Tüm ROUGE metrikleri, metin

özetleme literatürünün standardına uygun olarak F1-Skoru (F-measure) bazında hesaplanmıştır. ROUGE, aday özet ile referans özet arasındaki n-gram ve dizi tabanlı örtüşmeleri hesaplayan sektör standardı bir ölçüm yaklaşımıdır [43].

ROUGE-1: Aday özet ile referans özet arasında unigram (tek kelime) örtüşmesini ölçer. Modelin temel kavramları yakalama başarısını yansıtır.

ROUGE-2: Bigram (ikili kelime dizisi) örtüşmesini hesaplar. Dil akıcılığı ve yerel bağlam tutarlılığı hakkında bilgi verir.

ROUGE-L: En Uzun Ortak Alt Dizi (Longest Common Subsequence, LCS) yöntemine dayanır. Üretilen özetin yapısal bütünlüğünü ve sentaktik akıcılığını değerlendirir. Soyutlayıcı özetleme literatüründe en kritik göstergelerden biridir.

ROUGE-Lsum: Cümle seviyesinde LCS ölçümü yaparak çok cümleli özetlerin yapısal tutarlılığını değerlendirir. Özellikle PubMed gibi uzun ve çok cümleli özetlerin bulunduğu görevlerde kullanımı yaygındır.

ROUGE metrikleri, modellerin sözcük düzeyinde ve yapısal düzeyde ne ölçüde benzer özetler ürettiğini karşılaştırmak için standart bir çerçeve sunmaktadır.

3.4.2. Değerlendirme Prosedürü (Evaluation Procedure)

Değerlendirme aşaması, modellerin aynı veri bölmesi üzerinde eşit koşullarda test edilmesi prensibine dayanmaktadır. Tüm modeller, önceden ayrılmış PubMed test seti üzerinde çalıştırılmış ve her bir model için: Girdi metni modele kodlayıcı-çözücü veya yalnızca çözücü tabanlı yapısına uygun biçimde verilmiştir. Model tarafından üretilen özetler toplanmıştır. ROUGE metrikleri HuggingFace Evaluate kütüphanesi kullanılarak otomatik şekilde hesaplanmıştır. Tüm modeller için aynı tokenizasyon ayarları korunmuştur; böylece karşılaştırma adil hâle getirilmiştir.

Değerlendirme süreci tamamen otomatik ve n-gram tabanlı olduğundan, modellerin performans farklarının nesnel şekilde ortaya çıkması sağlanmıştır. Bu prosedür sayesinde özet kalitesi, dil yapısı ve içerik örtüşmesi açısından modeller arasında güvenilir bir karşılaştırma yapılmıştır.

3.5. Sonuçların Hesaplanması ve Karşılaştırma (Result Computation & Comparison)

Bu bölümde, eğitilmiş modellerden elde edilen performans sonuçlarının nasıl çıkarıldığı, modellerin hız ve hesaplama yükü açısından nasıl değerlendirildiği ve deneylerde kullanılan yazılım bileşenleri açıklanmaktadır.

3.5.1. Üretim Süresi ve Hesaplama Yükü (Inference Speed & Computational Cost)

Modellerin yalnızca çıktı kalitesi değil, aynı zamanda hesaplama verimliliği ve pratik kullanım maliyetleri de değerlendirilmiştir. Bu amaçla her model için ince ayar ve değerlendirme (inference) süreleri ayrı olarak ölçülmüş ve Tablo 1'de sunulmuştur. Sonuçlar, modellerin parametre boyutu, mimari yapısı ve kullanılan ince ayar yöntemine göre belirgin performans farklılıkları ortaya koymaktadır.

Tablo 2. Modellerin Eğitim Süreleri
(Training Times of the Models)

MODEL	EĞİTİM SÜRESİ	DEĞERLENDİRME SÜRESİ
BioMistral-7	2 Saat 46 Dakika	3 Saat 51 Dakika
PEGASUS	22 Dakika	10 Dakika
BART-Large-CNN	30 Dakika	9 Dakika
FLAN-T5-XL	6 Saat	1.5 Saat

BioMistral-7B, 7 milyar parametre boyutuna sahip model olup, eğitim ve değerlendirme aşamalarında en yüksek hesaplama süresini gerektirmiştir. QLoRA optimizasyonu bellek kullanımında tasarruf sağlamasına rağmen, modelin geniş BPE kelime dağarcığı ve yüksek hesaplama yoğunluğu nedeniyle eğitim süresi 2 saat 46 dakika, değerlendirme süresi ise 3 saat 51 dakika olarak ölçülmüştür.

FLAN-T5-XL, 3 milyar parametreli encoder–decoder mimarisine sahip olup, LoRA tabanlı ince ayar uygulanmasına rağmen eğitim ve değerlendirme aşamalarında görece yüksek hesaplama süresi gerektirmiştir. Model için eğitim süresi yaklaşık 6 saat, değerlendirme süresi ise 1.5 saat olarak ölçülmüş olup, bu değerler diğer modellere kıyasla daha yüksek bir hesaplama maliyetine işaret etmektedir.

PEGASUS-Large ve BART-Large-CNN, daha kompakt parametre boyutları ve optimize edilmiş encoder–decoder mimarileri sayesinde eğitim ve değerlendirme aşamalarında daha kısa sürelerde tamamlanmıştır. PEGASUS-Large için eğitim süresi 22 dakika, değerlendirme süresi 10 dakika olarak ölçülürken; BART-Large-CNN için eğitim süresi 30 dakika, değerlendirme süresi ise yaklaşık 9 dakika olarak kaydedilmiştir. Bu sonuçlar, her iki modelin eğitim ve değerlendirme süreçlerinin diğer modellere kıyasla daha kısa sürdüğünü göstermektedir.

Bu bulgular birlikte değerlendirildiğinde, model kapasitesi arttıkça hesaplama maliyetinin doğrusal olmayan bir biçimde yükseldiği görülmektedir. Küçük ve orta ölçekli Transformer tabanlı modeller görece kısa sürede eğitilebilirken, büyük dil modellerinin hem eğitim hem de değerlendirme aşamalarında önemli ölçüde zaman ve donanım kaynağı gerektirdiği anlaşılmaktadır. Bu durum, özetleme sistemlerinin pratik kullanım senaryolarında doğruluk kazanımları ile hesaplama maliyeti arasındaki

dengenin kritik bir değerlendirme ölçütü hâline geldiğini göstermektedir.

Tablo 3. Modellerin Çıkarım Süresi ve Üretilen Özet Uzunluğu Karşılaştırması
(Inference Time and Summary Length Comparison)

Model	Ortalama Çıkarım Süresi (sn)	Ortalama Kelime Sayısı
BART-Large-CNN	3.67	90
PEGASUS-Large	2.79	110
FLAN-T5-XL	8.88	55
BioMistral-7B	7.96	89

Tablo 2'de sunulan sonuçlar, modeller arasında belirgin hesaplama verimliliği farklılıkları bulunduğunu göstermektedir. PEGASUS-Large ve BART-Large-CNN modelleri, görece daha düşük parametre boyutları sayesinde en kısa çıkarım sürelerini elde etmiş olup, gecikmeye duyarlı uygulamalar için avantajlı bir profil sergilemiştir. Özellikle PEGASUS-Large, en hızlı çıkarımı gerçekleştirirken aynı zamanda en uzun özetleri üretmiş ve hız–çıktı uzunluğu dengesi açısından dikkat çekici bir performans göstermiştir.
FLAN-T5-XL modeli, diğer modellere kıyasla daha uzun çıkarım süresine sahip olmasına rağmen daha kısa özetler üretmiştir. Bu durum, modelin özetleme sürecinde daha agresif bir sıkıştırma stratejisi benimsediğini ve çıktı uzunluğu üzerinde daha katı bir kontrol uyguladığını göstermektedir.
BioMistral-7B modeli ise yüksek parametre sayısına bağlı olarak çıkarım süresi açısından daha maliyetli bir yapı sergilese de, ürettiği özetlerin uzunluğu ve içerik bütünlüğü göz önünde bulundurulduğunda dengeli bir hesaplama profili ortaya koymaktadır. Bu bulgu, alan odaklı büyük dil modellerinin daha yüksek hesaplama maliyetine rağmen daha tutarlı ve anlamsal açıdan zengin özetler üretebildiğini desteklemektedir.
Genel olarak değerlendirildiğinde, elde edilen sonuçlar özetleme modellerinin yalnızca doğruluk metrikleriyle değil, aynı zamanda hesaplama maliyeti, çıkarım süresi ve çıktı uzunluğu gibi pratik kullanım kriterleriyle birlikte ele alınması gerektiğini ortaya koymaktadır.

3.5.4. Kullanılan Yazılım Kütüphaneleri (Software Libraries Used)

Bu çalışmada tüm deneysel süreçler PyTorch tabanlı bir derin öğrenme altyapısı üzerinde yürütülmüştür. Model yükleme, ince ayar (fine-tuning) ve özet üretimi işlemleri için HuggingFace Transformers kütüphanesi; PubMed veri setinin okunması ve işlenmesi için HuggingFace Datasets; ROUGE metriklerinin hesaplanması için ise HuggingFace Evaluate kullanılmıştır. Sayısal işlemler NumPy ile, bozuk CSV satırlarının temizlenmesi ve standartlaştırılması Python'un yerleşik csv modülü ile gerçekleştirilmiştir. FLAN-T5-XL ve BioMistral-7B modelleri için parametre-verimli ince ayar amacıyla PEFT (Parameter-Efficient Fine-Tuning) kütüphanesi, BioMistral-7B'de düşük-bitli ağırlıkların yönetimi için BitsAndBytesConfig ve denetimli ince ayar sürecinin yönetimi için TRL (Transformer Reinforcement Learning) kullanılmıştır. Bu kütüphanelerin birlikte kullanılması, eğitim ve değerlendirme aşamalarının standartlaştırılmış, tekrarlanabilir ve yeniden üretilebilir bir deneysel çerçeve içinde yürütülmesini sağlamıştır.

4. BULGULAR (RESULTS)

Bu bölümde, ince ayar yapılan dört farklı modelin performans sonuçları ROUGE metrikleri üzerinden değerlendirilmiştir. Elde edilen bulgular, BioMistral modelinin özellikle ROUGE-1, ROUGE-2 ve ROUGE-L skorlarında diğer modellere kıyasla daha yüksek performans gösterdiğini ortaya koymaktadır.

4.1. Modeller (Models)

Bu çalışmada dört farklı mimari değerlendirilmektedir: BART-Large, PEGASUS-Large, FLAN-T5-XL ve BioMistral-7B. BART, PEGASUS ve T5 modelleri kodlayıcı-çözücü yapısına sahip seq2seq mimarileridir; buna karşılık BioMistral, LLaMA türevi yalnızca çözücü tabanlı bir büyük dil modelidir. Modeller mimari olarak farklı olsa da tümü özetleme görevine uygun olarakm çıkış üretmektedir. Parametre boyutları incelendiğinde BioMistral-7B en büyük modeldir; FLAN-T5-XL yaklaşık 3B parametreye sahipken, PEGASUS ve BART modelleri sırasıyla 568M ve 406M parametre ile daha kompakt yapılardır. Tokenizasyon süreçleri de model ailelerine göre değişmektedir: BART Roberta tabanlı BPE tokenizer kullanırken, PEGASUS ve T5 SentencePiece tabanlı unigram tokenizasyonu tercih etmektedir; BioMistral ise LLaMA2 ile uyumlu geniş bir BPE vocabulary'si kullanmaktadır.

Tablo 4. PubMed veri seti üzerinde değerlendirilen özetleme modellerinin ROUGE skorları
(ROUGE scores of summarization models evaluated on the PubMed dataset)

MODEL	İNCE AYAR STRATEJİSİ	R-1	R-2	R-L	R-LSUM
BioMistral-7B	QLoRA	**47.74**	**21.08**	**27.99**	30.18
PEGASUS-Large	Full Fine-Tuning	45.84	20.03	27.99	27.98
BART-Large-CNN	Full Fine-Tuning	44.37	17.88	25.91	**38.28**
FLAN-T5-XL	LoRA	42.34	17.44	25.96	25.91
PEGASUS-Large [16]	Pre-trained	45.09	19.56	27.42	-
Ptr-Gen-Seq2Seq [29]	Supervised training	35.86	10.22	29.69	-
Seq2seq Attentional Model [44]	Seq2Seq + Attention	27.62	15.70	27.11	-
SUMPUBMED [45]	Supervised training	30.61	7.66	27.80	-
flan-t5-xlsum [20]	Full Fine-Tuning	35.41	14.45	28.21	28.24

4.2. Model Karşılaştırması (Model Comparison)

Deneysel sonuçlar, farklı mimarilerin metin özetleme görevindeki başarı düzeylerini açıkça ortaya koymaktadır Model karşılaştırmasına ilişkin nihai ROUGE-1, ROUGE-2, ROUGE-L ve ROUGE-Lsum skorları Tablo 3'te sunulmuştur.Dört model arasında en yüksek genel performans, BioMistral-7B modeli tarafından elde edilmiştir. Model, çoğu ROUGE metriklerinde lider konumda yer almıştır.

BART-Large-CNN modeli, geleneksel kodlayıcı-çözücü mimarilerinin tipik bir örneği olarak çıkarımsal özetleme eğilimi göstermiş ve ROUGE-Lsum skorunda diğer tüm modelleri geride bırakmıştır. Bu durum, modelin cümle yapısı ve kelime diziliminde yüksek yüzeysel benzerlikler yakalasa da, anlam düzeyinde tutarlılığın daha sınırlı kaldığını göstermektedir.

Sonuç olarak, BioMistral-7B, hem akıcılık (ROUGE-L) hem de kelime düzeyi doğruluğu (ROUGE-1) açısından en yüksek başarıyı elde etmiş; bu da modelin biyomedikal terminolojiye özgü semantik ilişkileri yakalamada üstün performans sergilediğini ortaya koymuştur.

BART-Large metni büyük ölçüde kopyalayarak yapısal bir dönüşüm sunmuş ancak özgün bir soyutlama gerçekleştirememiştir.

PEGASUS, güçlü soyutlama kabiliyetine rağmen klinik açıdan hatalı halüsinasyonlar üretmiştir.

FLAN-T5, biçimsel olarak tutarlı olmakla birlikte, çalışmanın klinik anlamını BioMistral kadar derinlemesine ifade edememiştir.

Buna karşın BioMistral, özetin bir bölümünde yorumlayıcı bir üslup kullanmış olsa da, bu yaklaşım modelin çalışmayı kavramsal düzeyde anlamlandırabildiğini göstermektedir. Model, ctDNA değişimlerinin prognostik değerini doğru bağlama yerleştirerek, metastatik meme kanseri hastalarında erken progresyonun biyolojik temelini kavramış ve bunu klinik olarak anlamlı bir çerçevede yeniden ifade etmiştir. Bu yönüyle BioMistral, yüzeysel n-gram eşleşmesinin ötesine geçerek içerik düzeyinde semantik temsil üretme kapasitesini açıkça ortaya koymuştur.

Genel olarak değerlendirildiğinde BioMistral-7B, yalnızca ROUGE skorlarında değil, kavramsal doğruluk, klinik mesajın korunması, terminolojik hassasiyet ve anlam yoğunlaştırma açılarından da en dengeli ve başarılı modeli temsil etmektedir. Bu niteliksel örnek, BioMistral'ın biyomedikal özetleme görevlerinde neden daha güvenilir ve etkili bir seçenek olduğunu somut biçimde göstermektedir.

4.2.1. Niteliksel Örnek Analizi
(Qualitative Example Analysis)

Tablo 4'te farklı mimarilere sahip modellerin aynı PubMed makalesi üzerinde ürettiği özetler karşılaştırmalı olarak sunarak, nicel ROUGE metriklerinin yanında niteliksel farklılıklarında ortaya koymayı amaçlamaktadır. Bu analiz, modellerin yalnızca yüzeysel n-gram örtüşmeleri açısından değil; kavramsal soyutlama kapasitesi, terminolojik doğruluk ve bağlamsal tutarlılık bakımından da değerlendirilmesine imkân sağlamaktadır.

Bu niteliksel analizde kullanılan örnek, PubMed veri setinden seçilen ve metastatik meme kanserinde dolaşımdaki tümör DNA'sının (ctDNA) prognostik değerini inceleyen bir biyomedikal çalışmaya dayanmaktadır [46]

Çıktılar incelendiğinde, BART-Large-CNN modelinin büyük ölçüde kaynak metindeki cümle yapılarını ve ifadeleri koruyarak çıkarımsal bir özetleme eğilimi sergilediği görülmektedir. Bu durum, modelin ROUGE-Lsum skorunda yüksek bir değer elde etmesini açıklamakla birlikte, özgün soyutlama ve kavramsal yeniden ifade etme kapasitesinin sınırlı kaldığını göstermektedir.

PEGASUS-Large modeli ise daha soyutlayıcı bir yaklaşım benimsemiş; ancak bazı klinik ayrıntılarda bağlamdan kopuk veya metin dışı ifadeler üreterek halüsinasyon eğilimi sergilemiştir.

FLAN-T5-XL modeli biçimsel açıdan tutarlı ve dilsel olarak düzgün özetler üretmiş olsa da, içeriği aşırı derecede sıkıştırması nedeniyle çalışmanın klinik bağlamını ve elde edilen bulguların önemini sınırlı düzeyde yansıtabilmiştir.

BioMistral-7B modeli, makalenin temel klinik amacını, kullanılan yöntemi ve biyolojik çıkarımları daha bütüncül bir çerçevede ele almış; kavramsal ilişkileri yorumlayarak yeniden ifade edebilmiştir. Modelin özetinde gözlemlenen yorumlayıcı üslup, yüzeysel kopyalamadan ziyade anlamsal düzeyde bir soyutlama gerçekleştirdiğini göstermektedir.

Tablo 5. BART, PEGASUS, FLAN-T5 ve BioMistral modellerinin ürettiği özet çıktılarının karşılaştırılması
(Comparison of summary outputs generated by BART, PEGASUS, FLAN-T5, and BioMistral models)

Model	Üretilen Özet (Model Çıktısı)
BART-Large-CNN	objective : to evaluate circulating tumor DNA (ctDNA) change as an early prognostic biomarker of subsequent radiologic progression and progression-free survival in patients with metastatic breast cancer (mbc).methods : blood samples from patients with mbc were analyzed for levels of ctDNA , carcinoembryonic antigen , and cancer antigen 15-3 at baseline and during treatment .a clinical laboratory Improvement Amendments-certified panel of 73 genes was used to quantify tumor-specific point mutations in circulating ctDNA.results:genomic progression was detected in 76 baseline samples (90.5%) and 71 during-treatment
PEGASUS	background : early diagnosis of metastatic breast cancer (mbc) is essential for initiating appropriate treatment and avoiding disease progression . we evaluated the utility of genomic profiling in the early detection of metastatic mbc.methods:genomic profiling was performed on blood and tumor samples from patients with mbc at baseline . the results were compared with those of clinical and radiological features.results:sixty-three patients were included in the study . on average , patients had a tumor size of 2 cm at presentation , a lymph node diameter of 4 cm , and a skin color of tan to brown . all but one patient had metastatic disease at some point during the course of their disease . genomic profiling detected a germline
FLAN-T5-XL	PURPOSE: Accurate monitoring of therapeutic response remains an important unmet need for patients with metastatic breast cancer (MBC). Analysis of tumor genomics obtained via circulating tumor DNA (ctDNA) can provide a comprehensive overview of tumor evolution. Here, we evaluated ct DNA change as an early prognostic biomarker of subsequent radiologic progression and survival in MBC.PATIENTS AND METHODS: Paired blood samples from patients with MBC were analyzed for levels of tumor DNA, carcinoembryonic antigen, and cancer antigen 15-3 at baseline and during treatment. A Clinical Laboratory Improvement Amendments-certified sequencing panel of 73 genes was
BioMistral-7B	This is a very interesting article that highlights the importance of circulating tumor DNA (ctDNA) as a potential biomarker of early disease progression in patients with metastatic breast cancer (MBC). The authors describe a cohort of patients with MBC who had paired blood samples collected at baseline and during treatment. The authors used a clinical laboratory improvement amendments-certified sequencing panel of 73 genes to quantify tumor-specific point mutations in ctDNA. The authors found that patients with genomic progression (defined as a rise in ctDNA from baseline to during

5. TARTIŞMA (DISCUSSION)

Bu çalışmada elde edilen sonuçlar, biyomedikal metin özetleme görevlerinde alan-uyarlamalı ön-eğitimin model başarısı üzerinde önemli bir rol oynadığını ortaya koymaktadır. Nicel sonuç metrikleri ve niteliksel örnek analizleri birlikte değerlendirildiğinde, BioMistral-7B modelinin genel amaçlı Transformer tabanlı modellere kıyasla daha bağlamsal olarak tutarlı, kavramsal ilişkileri koruyan ve klinik içeriği anlamlı biçimde yeniden yapılandırabilen özetler ürettiği gözlemlenmiştir. Bu durum, biyomedikal alana özgü ve bilgi yapılarının doğrudan model parametrelerine entegre edilmesinin, soyutlayıcı özetleme gibi anlamsal derinlik gerektiren görevlerde avantaj sunduğunu göstermektedir.

ROUGE metriklerine dayalı değerlendirmeler, bazı modellerin yüksek skorlar elde etmesine rağmen niteliksel açıdan sınırlı performans sergileyebildiğini ortaya koymaktadır. Özellikle BART-Large-CNN modelinin ROUGE-Lsum skorunda öne çıkması, modelin kaynak metindeki cümle yapıları ve sözcük dizilimleriyle yüksek yüzeysel örtüşme yakalayabildiğini göstermektedir. Ancak niteliksel analizler, bu başarının çoğunlukla çıkarımsal üretim eğilimi ve sınırlı soyutlama kapasitesiyle ilişkili olduğunu; klinik bilginin yeniden ifade edilmesi ve kavramsal ilişkilerin korunması açısından yetersiz kaldığını ortaya koymuştur. Bu bulgu, ROUGE'un tek başına anlamsal kaliteyi tam olarak yansıtamadığını ve özellikle biyomedikal özetleme gibi bilgi yoğun alanlarda niteliksel değerlendirmelerle desteklenmesi gerektiğini bir kez daha vurgulamaktadır.

PEGASUS-Large modeli, özetleme görevine özgü GAP ön-eğitim stratejisi sayesinde belirgin bir soyutlayıcı üretim yeteneği sergilemiş; ancak bazı örneklerde klinik bağlamdan kopuk veya metin dışı ifadeler üretme eğilimi göstermiştir. Bu durum, modelin güçlü soyutlama kapasitesine rağmen biyomedikal terminolojiye ve alan-özel bilgiye yeterince hâkim olmamasının bir sonucu olarak değerlendirilebilir.

Benzer şekilde FLAN-T5-XL modeli, dilsel akıcılık ve biçimsel tutarlılık açısından dengeli çıktılar üretmiş; ancak içeriği aşırı derecede sıkıştırması nedeniyle çalışmanın klinik önemini yeterince yansıtamamıştır.

Buna karşılık BioMistral-7B modeli, biyomedikal literatür üzerinde gerçekleştirilen ek ön-eğitim sayesinde klinik terminolojiye yüksek düzeyde hâkimiyet, kavramsal ilişkileri koruma ve bağlamsal bütünlüğü sürdürme açısından belirgin bir üstünlük sergilemiştir. Niteliksel örnek analizinde modelin biyolojik çıkarımları doğru biçimde yorumlayarak yeniden ifade edebilmesi, yüzeysel n-gram eşleşmesinin ötesinde gerçek bir anlamsal soyutlama gerçekleştirdiğini göstermektedir. Bu bulgu, BioMistral'ın yüksek ROUGE skorlarının yalnızca sözcük düzeyindeki örtüşmeden değil, içerik düzeyindeki doğru temsil yeteneğinden kaynaklandığını ortaya koymaktadır.

Bu çalışmanın bazı sınırlılıkları da bulunmaktadır. Tüm modeller, uzun klinik makaleler için sabit bir maksimum giriş uzunluğu ile eğitilmiş olup bu durum, özellikle uzun bağlamlı metinlerde bilginin tamamının modele aktarılmasını sınırlandırmıştır. Ayrıca değerlendirme süreci büyük ölçüde otomatik ROUGE metriklerine dayandığından, insan uzman değerlendirmeleri çalışmaya dâhil edilmemiştir. Bu sınırlılıklar, elde edilen sonuçların yorumlanmasında göz önünde bulundurulmalıdır.

6. SONUÇ (CONCLUSION)

Bu çalışma, PubMed veri seti üzerinde gerçekleştirilen kapsamlı deneyler aracılığıyla, alan odaklı büyük dil modellerinin biyomedikal soyutlayıcı özetleme görevlerinde genel amaçlı modellere kıyasla daha güvenilir ve anlamsal açıdan daha iyi çıktılar üretebildiğini göstermiştir. Elde edilen bulgular, BioMistral-7B modelinin PEGASUS-Large, FLAN-T5-XL ve BART-Large-CNN gibi güçlü karşılaştırma modellerine kıyasla hem nicel performans metriklerinde hem de içerik düzeyinde daha tutarlı özetler ürettiğini ortaya koymaktadır.

Nicel değerlendirme sonuçları, BioMistral-7B'nin ROUGE-1, ROUGE-2 ve ROUGE-L metriklerinde en yüksek performansı elde ettiğini göstermiş; niteliksel analizler ise bu başarının biyomedikal bağlamın doğru yorumlanmasına, klinik bilginin anlamlı biçimde yeniden yapılandırılmasına ve anlamsal bütünlüğün korunmasına dayandığını doğrulamıştır. Bu durum, alan-uyarlamalı ön-eğitimin biyomedikal özetleme gibi yüksek bilgi yoğunluklu görevlerde avantaj sunduğunu açıkça ortaya koymaktadır.

Çalışmadan elde edilen sonuçlar, biyomedikal literatür gibi uzun bağlamlı ve teknik metinlerde, genel amaçlı dil modellerinin sınırlılıklarını vurgularken; alan odaklı büyük dil modellerinin klinik bilgi yönetimi ve bilimsel doküman özetleme uygulamaları için daha uygun bir çözüm sunduğunu göstermektedir. Ayrıca bu çalışma, otomatik değerlendirme metriklerinin niteliksel analizlerle birlikte ele alınmasının gerekliliğini vurgulayarak, biyomedikal özetleme değerlendirmelerine daha bütüncül bir bakış açısı kazandırmaktadır.

Gelecek çalışmalarda, BioMistral modelinin daha uzun bağlamları işleyebilen mimarilerle değerlendirilmesi, çok belgeli özetleme senaryolarında test edilmesi ve insan uzman değerlendirmeleriyle desteklenmesi planlanmaktadır. Ayrıca Türkçe biyomedikal veri setleriyle yapılacak uyarlamalar ve çok dilli özetleme yeteneklerinin incelenmesi, özellikle yerel sağlık metinlerinin otomatik işlenmesi açısından önemli bir araştırma yönü sunmaktadır.

KAYNAKLAR (REFERENCES)

[1] M. Azam, S. Khalid, S. Almutairi, H. A. Khattak, A. Namoun, A. Ali, and H. S. M. Bilal, "Current trends and advances in extractive text summarization: A comprehensive review," *IEEE Access*, vol. 13, pp. 28150–28166, 2025, doi: 10.1109/ACCESS.2025.3538886.

[2] M. Zhang, G. Zhou, W. Yu, N. Huang, and W. Liu, "A comprehensive survey of abstractive text summarization based on deep learning," *Computational Intelligence and Neuroscience*, vol. 2022, Article ID 7132226, pp. 1–18, Aug. 2022, doi: 10.1155/2022/7132226.

[3] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020, doi: 10.1093/bioinformatics/btz682.

[4] R. Luo, L. Sun, Y. Xia, T. Qin, S. Zhang, H. Poon, and T.-Y. Liu, "BioGPT: generative pre-trained transformer for biomedical text generation and mining," *Briefings in Bioinformatics*, vol. 23, no. 6, pp. 1–11, 2022, doi: 10.1093/bib/bbac409.

[5] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, H. Poon, "Domain-specific language model pretraining for biomedical natural language processing**", Proceedings of the 2021 Conference on Health, Inference, and Learning (CHIL)**, Virtual Conference, 55–65, 2021.

[6] R. Tinn, H. Cheng, Y. Gu, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Fine-tuning large neural language models for biomedical natural language processing," *Patterns*, vol. 4, no. 4, Article 100729, Apr. 2023, doi: 10.1016/j.patter.2023.100729.

[7] P. Verma and H. Om, "MCRMR: Maximum coverage and relevancy with minimal redundancy based multi-document summarization," *Expert Systems with Applications*, vol. 120, pp. 43–56, 2019, doi: 10.1016/j.eswa.2018.11.022.

[8] A. John, P. S. Premjith, and M. Wilscy, "Extractive multi-document summarization using population-based multicriteria optimization," *Expert Systems with Applications*, vol. 89, pp. 1–15, 2017, doi: 10.1016/j.eswa.2017.05.075.

[9] J. N. Madhuri, R. G. Kumar, "Extractive text summarization using sentence ranking", **International Conference on Data Science and Communication (IconDSC)**, Bangalore, Hindistan, pp. 112–117, 01–02 Mart 2019.

[10] A. K. Yadav, Ranvijay, R. S. Yadav, A. K. Maurya, "State-of-the-art approach to extractive text summarization: a comprehensive review", *Multimedia Tools and Applications*, vol. 82, no. 1, pp. 29135–29197, 2023.

[11] A. M. Rush, S. Chopra, and J. Weston, "A neural attention model for abstractive sentence summarization," **in Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 379–389, 2015.

[12] A. See, P. J. Liu, and C. D. Manning, "Get to the point: Summarization with pointer-generator networks," **in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)**, pp. 1073–1083, 2017.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, "Attention is all you need", **Advances in Neural Information Processing Systems (NeurIPS)**, Long Beach, CA, ABD, 5998–6008, 2017.

[14] F. A. Acheampong, H. Nunoo-Mensah, W. Chen, "Transformer models for text-based emotion detection: a review of BERT-based approaches", *Artificial Intelligence Review*, vol. 54, no. 8, pp. 6177–6208, 2021.

[15] W. Ansar, S. Goswami, and A. Chakrabarti, "A survey on transformers in NLP with focus on efficiency", 2024.

[16] J. Zhang, Y. Zhao, M. Saleh, and P. J. Liu, "PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization," **in Proceedings of the 37th International Conference on Machine Learning (ICML)**, PMLR, vol. 119, pp. 11328–11339, 2020.

[17] R. Alsultan, A. Sagheer, H. Hamdoun, L. Alshamlan, and L. Alfadhli, "PEGASUS-XL with saliency-guided scoring and long-input encoding for multi-document abstractive summarization," *Scientific Reports*, vol. 15, no. 26529, 2025, doi: 10.1038/s41598-025-11062-2.

[18] E. Daraghmi, L. Atwe, and A. Jaber, "A comparative study of PEGASUS, BART, and T5 for text summarization across diverse datasets," *Future Internet*, vol. 17, no. 389, pp. 1–33, Aug. 2025.

[19] C. Benzoni, M. Langhals, M. Boeker, L. Modersohn, M. E. Maros, "Deep learning and abstractive summarisation for radiological reports: an empirical study for adapting the PEGASUS models' family with scarce data", https://arxiv.org/abs/2509.15419, Erişim tarihi: 15.02.2025.

[20] S. A. Rafi, N. Rahman, K. N. Islam, H. Ahmad, Optimizing Abstractive Summarization With Fine-Tuned PEGASUS, Teknik Rapor, Department of Computer Science and Engineering, Brac University, Bangladeş, 2023.

[21] K. N. Lam, T. G. Doan, K. T. Pham, J. Kalita, "Abstractive Text Summarization Using the BRIO Training Paradigm", **Findings of the Association for Computational Linguistics (ACL)**, Toronto, Kanada, 2023, doi: 10.18653/v1/2023.findings-acl.7.

[22] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation", https://arxiv.org/abs/1910.13461, Erişim tarihi: 12.10.2025.

[23] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020.

[24] H. W. Chung, L. Hou, S. Longpre, B. Zoph, A. Tay, A. Fedus, "Scaling instruction-finetuned language models", https://arxiv.org/abs/2210.11416, Erişim tarihi: 11.10.2025.

[25] A. G. Etemad, A. I. Abidi, M. Chhabra, "Fine-tuned T5 for abstractive summarization", *International Journal of Performability Engineering*, vol. 17, no. 10, pp. 900–906, 2021.

[26] M. Wang, P. Xie, Y. Du, and X. Hu, "T5-Based Model for Abstractive Summarization: A Semi-Supervised Learning Approach with Consistency Loss Functions," *Applied Sciences*, vol. 13, no. 12, p. 7111, 2023, doi: 10.3390/app13127111.

[27] T. Tawmo, M. Borah, P. Dadure, P. Pakray, Comparative analysis of T5 model for abstractive text summarization on different datasets, Teknik Rapor, National Institute of Technology Silchar, Assam, Hindistan, 2022.

[28] M. Yasunaga, J. Kasai, R. Zhang, A. R. Fabbri, I. Li, D. Friedman, and D. R. Radev, "ScisummNet: A Large Annotated Corpus and Content-Impact Models for Scientific Paper Summarization with Citation Networks," **in Proceedings of the AAAI Conference on Artificial Intelligence (AAAI-19)**, vol. 33, pp. 7386–7393, 2019.

[29] A. Cohan, F. Dernoncourt, D. S. Kim, T. Bui, S. Kim, W. Chang, and N. Goharian, "A Discourse-Aware Attention Model for Abstractive Summarization of Long Documents," **in Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)**, New Orleans, LA, USA, pp. 615–626, 2018.

[30] P. Tejaswin, D. Naik, P. Liu, "How well do you know your summarization datasets?", https://arxiv.org/abs/2106.11388, Erişim tarihi: 14.02.2025.

[31] A. Chaves, C. Kesiku, and B. García-Zapirain, "Automatic text summarization of biomedical text data: A systematic review," *Information*, vol. 13, no. 393, pp. 1–21, Aug. 2022, doi: 10.3390/info13080393.

[32] H. Hayashi, W. Kryściński, B. McCann, N. Rajani, C. Xiong, "What's New? Summarizing Contributions in Scientific Literature", **Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)**, pp. 1019–1031, Mayıs 2023.

[33] T. Kerner, "Domain-specific pretraining of language models: a comparative study in the medical field", https://arxiv.org, Erişim tarihi: Şubat 2025.

[34] D. Van Veen, C. Van Uden, L. Blankemeier, J.-B. Delbrouck, A. Aali, C. Bluethgen, A. Pareek, M. Polacin, E. P. Reis, A. Seehofnerova, N. Rohatgi, P. Hosamani, W. Collins, N. Ahuja, C. Langlotz, J. Hom, S. Gatidis, J. Pauly, and A. S. Chaudhari, "Clinical Text Summarization: Adapting Large Language Models Can Outperform Human Experts," Oct. 2023, doi: 10.21203/rs.3.rs-3483777/v1.

[35] I. Beltagy, K. Lo and A. Cohan, "SCIBERT: A Pretrained Language Model for Scientific Text", https://arxiv.org/abs/1903.10676, Erişim tarihi: Kasım 2025.

[36] S. A. Lee, A. Wu, and J. N. Chiang, "Clinical ModernBERT: An Efficient and Long-Context Encoder for Biomedical Text", https://arxiv.org/abs/2504.03964, Erişim tarihi: Kasım 2025.

[37] Y. Labrak, A. Bazoge, E. Morin, P.-A. Gourraud, M. Rouvier and R. Dufour, "BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains", https://arxiv.org/abs/2402.10373, Erişim tarihi: Kasım 2025.

[38] J. L. Léon, V. B. Nogueira and P. Quaresma, "Are Small Language Models Enough for Biomedical QA Tasks? Fine-Tuning Mistral-7B Using LoRA and PubMedQA," **in Proceedings of the 17th International Conference on Machine Learning and Computing (ICMLC '25), Lecture Notes in Networks and Systems**, vol. 1475, pp. 420–432, Aug. 2025, doi: 10.1007/978-3-031-94892-3_31.

[39] F. J. Dorfner, A. Dada, F. Busch, T. Han, D. Truhn, J. Kleesiek, M. Sushil, J. Lammert, M. R. Makowski, L. C. Adams, and K. K. Bressem, "Biomedical large language models seem not to be superior to generalist models on unseen medical data", https://arxiv.org/abs/2408.13833, Erişim tarihi: Ekim 2025.

[40] M. Ravaut, A. Sun, N. F. Chen, S. Joty, "On context utilization in summarization with large language models", **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)**, pp. 2764–2781, 2024.

[41] M. Kruse, S. Hu, N. Derby, Y. Wu, S. Stonbraker, B. Yao, D. Wang, E. Goldberg and Y. Gao, "Zero-Shot Large Language Models for Long Clinical Text Summarization with Temporal Reasoning," **in Machine Learning and Knowledge Discovery in Databases. Research Track (ECML PKDD 2024), Lecture Notes in Computer Science**, vol. 15120, Springer, Cham, pp. 553–569, 2024, doi: 10.1007/978-3-031-94892-3_31.

[42] Cohan, A., Dernoncourt, F., Kim, D. S., Bui, T., Kim, S., Chang, W., & Goharian, N. (2018). A Discourse-Aware Attention Model for Abstractive Summarization of Long Documents. Proceedings of NAACL-HLT 2018, 615–621.

[43] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries", **Workshop on Text Summarization Branches Out**, Barcelona, İspanya, pp. 74–81, 2004.

[44] A. Ben Abacha and D. Demner-Fushman, "On the summarization of consumer health questions," **in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)**, Florence, Italy, pp. 2228–2234, 2019.

[45] V. Gupta, P. Bharti, P. Nokhiz, and H. Karnick, "SUMPUBMED: Summarization Dataset of PubMed Scientific Articles," **in Proceedings of the ACL Student Research Workshop (ACL SRW)**, pp. 292–303, 2021.

[46] JCO Precision Oncology, "Evaluation of circulating tumor DNA as an early prognostic biomarker in metastatic breast cancer," *JCO Precision Oncology*, vol. 4, pp. 1246–1262, Nov. 2020, doi: 10.1200/PO.20.00117. PMID: 35050782.