T.C.
YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

BÜYÜK DİL MODELLERİ İLE BİLDİRİM METNİ ÜRETİMİ

Hakan TAŞKÖPRÜ

YÜKSEK LİSANS TEZİ
Bilgisayar Mühendisliği Anabilim Dalı
Bilgisayar Mühendisliği Programı

Danışman
Prof. Dr. Banu Diri

Temmuz, 2025

T.C.
YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

BÜYÜK DİL MODELLERİ İLE BİLDİRİM METNİ ÜRETİMİ

Hakan TAŞKÖPRÜ tarafından hazırlanan tez çalışması 23.07.2025 tarihinde aşağıdaki jüri tarafından Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı Bilgisayar Mühendisliği Programı **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Prof. Dr. Banu Diri
Yıldız Teknik Üniversitesi
Danışman

Jüri Üyeleri

Prof. Dr. Banu Diri, Danışman
Yıldız Teknik Üniversitesi _____

Dr. Öğr. Üyesi Göksel Biricik, Üye
Yıldız Teknik Üniversitesi _____

Doç. Dr. Aysun Güran, Üye
Doğuş Üniversitesi _____

Danışmanım Prof. Dr. Banu Diri sorumluluğunda tarafımca hazırlanan Büyük Dil Modelleri ile Bildirim Metni Üretimi başlıklı çalışmada veri toplama ve veri kullanımında gerekli yasal izinleri aldığımı, diğer kaynaklardan aldığım bilgileri ana metin ve referanslarda eksiksiz gösterdiğimi, araştırma verilerine ve sonuçlarına ilişkin çarpıtma ve/veya sahtecilik yapmadığımı, çalışmam süresince bilimsel araştırma ve etik ilkelerine uygun davrandığımı beyan ederim. Beyanımın aksinin ispatı halinde her türlü yasal sonucu kabul ederim.

Hakan TAŞKÖPRÜ

İmza

Daha iyi bir geleceğe

TEŞEKKÜR

Büyük Dil Modelleri (LLM) ve Doğal Dil İşleme (NLP) alanındaki bu çalışmanın geliştirilmesinde, değerli katkıları, görüşleri, yol göstericiliği ve desteğiyle her zaman yanımda olan danışmanım Prof. Dr. Banu Diri'ye en içten teşekkürlerimi sunarım. Çalışmalarım süresince sağladıkları bilgi birikimi ve destekleri için Müge Yalçınkaya'ya ve Ali Alperen Sönmez'e teşekkür ederim. Ayrıca süreç boyunca yanımda olan aileme ve dostlarıma da şükranlarımı sunarım.

Hakan TAŞKÖPRÜ

İÇİNDEKİLER

KISALTMA LİSTESİ vii

ŞEKİL LİSTESİ viii

TABLO LİSTESİ ix

ÖZET x

ABSTRACT xi

1 GİRİŞ 1

1.1 Literatür Özeti . . . 2

1.1.1 Bildirimler . . . 2

1.1.2 Büyük Dil Modelleri . . . 2

1.1.3 İstem Mühendisliği . . . 3

1.1.4 Değerlendirme Yöntemleri . . . 4

1.2 Tezin Amacı . . . 8

1.3 Hipotez . . . 9

2 VERİ KÜMESİ 11

2.0.1 Kaynak ve Toplama Hikâyesi . . . 12

3 YÖNTEM 16

3.1 Elle Yöntem ile Bildirim Üretimi . . . 18

3.2 Otomatik Yöntem ile Bildirim Üretimi . . . 19

3.3 Optimize Yöntem ile Bildirim Üretimi . . . 20

3.4 Bildirimlerin Değerlendirilmesi . . . 21

3.4.1 Klasik Otomatik Değerlendirme Metrikleri . . . 21

3.4.2 Makine Öğrenmesi Tabanlı Yardımcı Metrikler . . . 25

3.4.3 Görev Odaklı LLM Tabanlı Metrikler . . . 27

3.5 Üretici Dil Modeli: Google Gemini 1.5 Pro . . . 28

3.6 Üretici Dil Modeli: Turkish-Gemma-9b-v0.1 . . . 30

3.7 Değerlendirici Dil Modeli: GPT-4o . . . 31

4 DENEYLER VE SONUÇLAR **34**
4.1 Deney Tasarımı 34
4.2 Sonuçlar 34

5 SONUÇ **40**

KAYNAKÇA **42**

A VURGU TANIMLARI **46**

B ÜSLUP TANIMLARI **47**

TEZDEN ÜRETİLMİŞ YAYINLAR **48**

KISALTMA LİSTESİ

LLM	Büyük Dil Modeli/Large Language Model

ŞEKİL LİSTESİ

Şekil 2.1 Kampanya tabanlı varyasyon üretim akışı. 14
Şekil 3.1 Elle, Otomatik ve Optimize Yöntemlerin Karşılaştırmalı Süreç Akışı 17
Şekil 3.2 Klasik metrikler için referans seçim süreci 22
Şekil 3.3 Anlamsal emoji tutarlılığı metriği için değerlendirme süreci . . . 26

TABLO LİSTESİ

Tablo 2.1 Varyasyon Örnekleri 13
Tablo 2.2 Veri Kümesi Şeması – Kampanya vs. Segment Alanları 14
Tablo 4.1 Gemma Türkçe Sonuçları — Yöntem bazında metrik karşılaştırması. En iyi değerler kalın gösterilmiştir. 35
Tablo 4.2 Gemini 1.5 Pro Sonuçları — Yöntem bazında metrik karşılaştırması. En iyi değerler kalın gösterilmiştir. 35
Tablo 4.3 `B_ALL_CHURN_400_500` kampanyası için yöntem ve LLM bazında bildirim örnekleri 38
Tablo 4.4 `X_HML_ACQ_400_475` kampanyası için yöntem ve LLM bazında bildirim örnekleri 38
Tablo 4.5 `ALL_PD_MILKA_1E1_1510` kampanyası için yöntem ve LLM bazında bildirim örnekleri 39

ÖZET

Büyük Dil Modelleri ile Bildirim Metni Üretimi

Hakan TAŞKÖPRÜ

Bilgisayar Mühendisliği Anabilim Dalı
Yüksek Lisans Tezi

Danışman: Prof. Dr. Banu Diri

Bildirimler, kullanıcıların uygulamalarla etkileşiminde önemli bir rol oynamaktadır. Bu tezde, bildirim metinlerinin üretiminde Büyük Dil Modelleri'nin (LLM) performansı değerlendirilmiş; Elle Yöntem, Otomatik Yöntem ve Optimize Yöntem olmak üzere üç yaklaşım karşılaştırılmıştır. Üretimde Google Gemini 1.5 Pro ve Turkish-Gemma-9b-v0.1, değerlendirmede ise GPT-4o modeli kullanılmıştır.

Kısa ve yaratıcı yapıları nedeniyle, geleneksel metriklerin (BLEU, ROUGE, METEOR) sınırlı kaldığı tespit edilmiş; bunun yerine anlamsal uygunluk, dilbilgisel doğruluk ve kullanıcı algısını ölçen LLM tabanlı yeni metrikler geliştirilmiştir. Sonuçlar, Optimize Yöntem'in dil bilgisi ve genel kalite açısından en başarılı yaklaşım olduğunu göstermiştir. Elle Yöntem ise çeşitlilik ve emoji kullanımı açısından öne çıkmıştır.

Bu çalışma, LLM'lerin kısa, eyleme dönük metin üretimindeki gücünü ortaya koymakta; optimize edilmiş istem mühendisliğinin, insan emeğini azaltarak yüksek kaliteli içerik üretiminde oynadığı kritik rolü vurgulamaktadır.

Anahtar Kelimeler: Büyük Dil Modelleri, bildirim üretimi, istem mühendisliği, LLM tabanlı değerlendirme metrikleri, kısa metin üretimi.

ABSTRACT

Notification Text Generation with Large Language Models

Hakan TAŞKÖPRÜ

Department of Computer Engineering
Master of Science Thesis

Supervisor: Prof. Dr. Banu Diri

Notifications play a crucial role in user interaction with applications. This thesis evaluates the performance of Large Language Models (LLMs) in generating notification texts and compares three approaches: Manual, Automatic, and Optimized. Google Gemini 1.5 Pro and Turkish-Gemma-9b-v0.1 were used for generation, while GPT-4o was employed for evaluation.

Due to the short and creative nature of notifications, traditional metrics (BLEU, ROUGE, METEOR) proved insufficient. Therefore, new LLM-based metrics were developed to assess semantic relevance, grammatical accuracy, and user perception. Results show that the Optimized Method outperforms others in grammar and overall quality, while the Manual Method excels in diversity and emoji usage.

This study demonstrates the potential of LLMs in generating concise, action-oriented texts and emphasizes the critical role of optimized prompt engineering in reducing human effort while ensuring high-quality output.

Keywords: Large Language Models, notification generation, prompt engineering, LLM-based evaluation metrics, short text generation.

YILDIZ TECHNICAL UNIVERSITY
GRADUATE SCHOOL OF SCIENCE AND ENGINEERING

1
GİRİŞ

Kullanıcılara doğru zamanda, uygun içerikle bildirim göndermek, dijital platformlarda kullanıcı etkileşimini artırma açısından kritik bir rol oynamaktadır [1]. Bildirimler, kullanıcıları bilgilendirmekten çok, onları belirli bir eyleme yönlendirme amacı taşıyan kısa metinlerdir [1, 2].

Son yıllarda, Büyük Dil Modelleri (LLM'ler), doğal dil işleme alanında devrim niteliğinde gelişmeler sunarak metin üretimi gibi görevlerde başarılı sonuçlar elde etmişlerdir [3]. Ancak, LLM tabanlı metin üretiminin özellikle bildirim üretimi gibi kısa ve hedef odaklı içerikler üzerindeki potansiyelinden literatürde yararlanılmadığı görülmektedir. Mevcut çalışmalar çoğunlukla uzun metin üretimi, soru-cevap sistemleri veya içerik özetleme gibi daha genel görevleri hedeflemektedir [4, 5].

Üretilen metinlerin kalitesini değerlendirmede insan merkezli yaklaşımlar daha başarılı olmakla birlikte, bu yöntemlerin yüksek zaman ve maliyet gereksinimi nedeniyle ölçeklenebilirliği sınırlıdır [6]. Bu durum, otomatik ve nesnel olan makine merkezli metriklerin kullanımını zorunlu kılmaktadır. Ancak mevcut literatürde yaygın olarak kullanılan ROUGE, BLEU gibi metrikler, bildirim üretimi gibi özgün ve kısa metin üretimi görevlerinin değerlendirilmesinde yetersiz kalmaktadır [7].

Bu tez çalışması, bildirim üretimi gibi ihmal edilmiş bir içerik türünde LLM'lerin nasıl etkili kullanılabileceğini araştırmayı hedeflemektedir. Özellikle, Elle İstem Oluşturma (Manual Prompt Generation, Elle Yöntem), Otomatik İstem Oluşturma (Automated Prompt Generation, Otomatik Yöntem) ve Optimizasyonlu Otomatik İstem Oluşturma (Optimized Automated Prompt Generation, Optimize Yöntem) olmak üzere üç farklı yöntem karşılaştırılmıştır.

Yöntemlerin karşılaştırılmasında geleneksel metriklerin yanı sıra, bu tez kapsamında geliştirilen LLM tabanlı değerlendirme metrikleri de kullanılmıştır. Bu

yeni metrikler, üretilen bildirimin dilbilgisi doğruluğu, başlık-içerik uyumu gibi niteliklerini ölçmeyi hedeflemektedir. Deneysel sonuçlar, Optimize Yöntem'in, diğer iki yönteme kıyasla daha yüksek performans sergilediğini göstermektedir.

Tezin devamında, literatürdeki benzer çalışmalar Literatür Özeti bölümünde incelenmekte, önerilen yöntemler Yöntem bölümünde detaylandırılmakta, deney tasarımı ve bulgular Deneyler ve Sonuçlar bölümünde sunulmakta ve nihai çıkarımlar ile gelecekte yapılabilecek çalışmalar Sonuçlar bölümünde tartışılmaktadır.

1.1 Literatür Özeti

1.1.1 Bildirimler

Bildirimler, kullanıcıları bilgilendirmek ve uygulamalarla etkileşimi artırmak için yaygın olarak kullanılan bir yöntemdir. Mobil ve web uygulamalarında, kullanıcıların uygulamaya dönüş oranlarını ve etkileşim düzeylerini yükseltmek amacıyla bildirimler sıklıkla kullanılmaktadır [8]. Ancak bildirim sayısının her geçen gün hızla artması, kullanıcı deneyimi ve etkileşimi açısından önemli bir sorun haline gelmiştir. Kullanıcılar gün içerisinde çok fazla bildirim aldıklarında, bunları göz ardı etme veya tamamen kapatma eğilimi göstermektedir [8]. Özellikle gereksiz veya bağlamsız bildirimler, kullanıcıların bildirimlere olan ilgisini azaltarak etkileşim oranlarını düşürebilmektedir [9].

Bununla birlikte, bir bildirimin kullanıcı için önemli olması, onun etkileşime girmesi için tek başına yeterli değildir [8]. Bildirim içeriği ve sunumu, kullanıcının bildirime tıklama kararında kritik bir rol oynamaktadır [1, 8, 9]. Ayrıca e-ticaret uygulamalarında kullanıcının ilgisini çekecek ürünün bildirimde geçmesi etkileşimi artırmaktadır [10].

Mevcut literatürde, bildirimler üzerine yapılan çalışmalar genellikle ne zaman[9, 11, 12] ve hangi bildirimin[13] gönderileceği, hangi kullanıcılara gönderileceği[14], bildirim stratejileri[15] ve kullanıcı segmentasyonu[16] gibi konulara odaklanmaktadır. Ancak, bildirim içeriği üretiminin özellikle LLM tabanlı bir yaklaşımla ele alınmasına yönelik kapsamlı bir çalışma bulunmamaktadır.

1.1.2 Büyük Dil Modelleri

Büyük Dil Modelleri (LLM'ler), genellikle Transformer mimarisine dayanan, büyük ölçekli ve önceden eğitilmiş sinir ağı tabanlı modellerdir. Bu modeller, geniş ve çeşitli metin verileri üzerinde eğitilerek doğal dil işleme (NLP) alanında çok

sayıda görevde kullanılabilir hale gelmiştir [17].

LLM'ler; metin üretimi, makine çevirisi, metin özetleme, talimat takibi, bilgi çıkarımı, çok dilli görevler, diyalog sistemleri ve bilgiye dayalı akıl yürütme gibi alanlarda yüksek performans göstermektedir [3, 4, 17]. Ayrıca bu modeller, yalnızca teknik uygulamalarla sınırlı kalmayıp, giderek artan bir şekilde eğitim, finans, tıp ve bilimsel araştırma gibi uzmanlık gerektiren alanlarda da benimsenmektedir [4].

LLM tabanlı içerik üretimi, son yıllarda pazarlama ve iletişim alanlarında da yaygınlık kazanmıştır. Özellikle arama motoru optimizasyonu (SEO) içerikleri[18], e-posta kampanyaları[19], reklam metinleri[20] ve sosyal medya gönderileri[21] gibi pazarlama odaklı metinlerin üretiminde bu modellerin kullanımı artmaktadır.

Ancak literatürde, bildirim gibi kısa, kullanıcıyı harekete geçirmeyi amaçlayan ve sınırlı bağlam içeren metinlerin, LLM'ler aracılığıyla üretilmesine yönelik kapsamlı bir çalışmaya rastlanmamaktadır. Mevcut araştırmalar genellikle daha uzun metin üretimi veya genel içerik oluşturma görevlerine odaklanmakta; bildirim metinlerinin sahip olduğu kendine özgü kısıtlar—örneğin sınırlı uzunluk, dikkat çekici olma zorunluluğu, kişiselleştirme gerekliliği ve kullanıcıyı eyleme geçirme hedefi—çoğu zaman göz ardı edilmektedir [4, 5]. Bu bağlamda, bildirim özelinde etkili istemlerin nasıl oluşturulması gerektiği, hem teorik hem de pratik açıdan yeterince incelenmemiş önemli bir boşluk alanı oluşturmaktadır.

1.1.3 İstem Mühendisliği

Üretken Yapay Zeka modellerinin çıktıları, bu modellere verilen istemler (prompt) aracılığıyla yönlendirilir [22]. İstem mühendisliği, bu modellerden istenilen çıktıları elde edebilmek için en uygun istemlerin tasarlanması sürecidir [23]. Bu alan, modellerin farklı görevlerde etkili bir şekilde kullanılabilmesini sağlamak adına son yıllarda büyük önem kazanmıştır.

Literatürde istem mühendisliği teknikleri 6 temel başlık altında toplanabilir: Bağlam İçi Öğrenme (In-Context Learning), Zero-Shot, Thought Generation, Decomposition, Ensembling ve Self-Criticism [22]. Bu çalışmada ise yalnızca Zero-Shot ve Few-Shot/Bağlam İçi Öğrenme (In-Context Learning) tekniklerine odaklanılmıştır.

Bağlam İçi Öğrenme, dil modellerinin yalnızca birkaç örnekle yeni görevleri öğrenmesini sağlayan bir paradigmadır [24]. Bu yaklaşım, herhangi bir ek eğitim gerektirmeden öğrenmeyi mümkün kılarak özellikle üretken yapay

zeka modelleriyle çalışan sistemlerde büyük esneklik sağlar [24]. Büyük Dil Modelleri, bu özellik sayesinde sınırlı sayıda örnekle bile birçok görevde etkileyici performanslar sergileyebilmektedir [25]. Zero-Shot öğrenme, dil modelinin herhangi bir görev örneği sunulmadan, yalnızca görev tanımı içeren bir istemle yönlendirilmesidir. Bu yöntemde model, daha önce benzer bir görevle karşılaşmamış olsa dahi, genel dil bilgisi ve önceden öğrenilmiş dünya bilgisiyle yanıt üretmeye çalışır [26]. Buna karşılık, Few-Shot öğrenme yaklaşımında, model görevle ilgili birkaç örnek içeren bir istem aracılığıyla yönlendirilir. Bu örnekler, modelin görevi daha iyi anlamasını ve bağlamdan yola çıkarak doğru çıktılar üretmesini sağlar. Few-Shot öğrenme, Bağlam İçi Öğrenme (In-Context Learning) paradigmasının bir uygulaması olarak kabul edilir ve genellikle Zero-Shot yöntemine kıyasla daha başarılı sonuçlar vermektedir [25].

Bağlam İçi Öğrenme'nin başarısı; örneklerin doğru seçimi ve sıralanması gibi faktörlerden önemli ölçüde etkilenmektedir [27]. Ayrıca, örneklerin gösterim formatı ve etiket alanı gibi yapısal özelliklerin performans üzerinde, doğru etiketlerden daha belirleyici olduğu gösterilmiştir [28]. Bağlam İçi Öğrenme'nin güçlendirilmesine yönelik teknikler de geliştirilmiştir; örneğin MetaICL yöntemi, modellerin ön eğitim (pretraining) aşamasında bu yeteneğini artırarak görevleri daha etkili öğrenmelerini sağlamaktadır [29].

Bununla birlikte, istemlerin elle oluşturulması zaman alıcı[30] ve emek gerektiren[31] bir süreçtir. Üstelik, dil modellerine ve görev türlerine göre optimum istemlerin değişiklik göstermesi, elle istem mühendisliğini daha da karmaşık hale getirmektedir [32].

Bu zorlukları aşmak amacıyla, literatürde otomatik istem üretimi ve optimizasyonu alanında pek çok çalışma gerçekleştirilmiştir [31–37]. Otomatikleştirilmiş istem mühendisliği, yalnızca insan emeğini azaltmakla kalmayıp, aynı zamanda Büyük Dil Modellerinin çeşitli görevlerdeki performansını artırma potansiyeli de taşımaktadır [38]. Ayrıca, istemlerin çıktılarda kişiselleştirme sağlayacak şekilde otomatik yeniden yazımı üzerine de çalışmalar mevcuttur [39].

1.1.4 Değerlendirme Yöntemleri

İnsan merkezli metrikler, genellikle en güvenilir değerlendirme yöntemleri olarak kabul edilmekte ve model çıktılarının insanlar tarafından doğrudan değerlendirilmesine dayanır. Bu yöntemler, içerik kalitesi ve kullanıcı deneyimi gibi bağlamsal unsurları da göz önünde bulundurduğu için daha doğru sonuçlar vermektedir [6]. Ancak, insan değerlendirmeleri yüksek maliyetli, zaman alıcı ve

ölçeklenebilirlik açısından sınırlayıcıdır [6, 40]. Ayrıca, değerlendirmeler kişiden kişiye değişebildiği için tutarlılık (consistency) problemi de bulunmaktadır [6]. Ayrıca uzmanlık gerektirebilmektedir[6].

Eğitimsiz otomatik metrikler, herhangi bir model eğitimi gerektirmeden, üretilen metin ile insan yazımı referans metinler arasındaki benzerliği ölçen değerlendirme araçlarıdır. Bu metrikler; makine çevirisi, görsel betimleme ve soru oluşturma gibi doğal dil üretimi (NLG) görevlerinde yaygın olarak kullanılmaktadır [6]. Genellikle n-gram örtüşmesi, metinler arası mesafe, içerik benzerliği, çeşitlilik ve dilbilgisel özelliklere dayalı ölçümler olmak üzere beş ana grupta sınıflandırılmaktadır. Eğitimsiz metriklerin en önemli avantajı, hesaplama açısından hızlı ve verimli olmalarıdır; bu sayede model geliştirme süreçlerinde günlük ilerlemelerin izlenmesini mümkün kılar. Bununla birlikte, bu metriklerin insan yargılarıyla olan korelasyonu genellikle sınırlı kalmakta; yani, metriklerin verdiği skorlar, insanlar tarafından yapılan kalite değerlendirmeleriyle yüksek düzeyde örtüşmeyebilmektedir [6]. Bu durum, özellikle bağlam bilgisi gerektiren ya da yaratıcı üretim barındıran metinlerde daha da belirgin hale gelmektedir. Örneğin, bir şiirsel anlatım, hiciv veya kültürel referans içeren ifade biçimi, n-gram örtüşmesine dayanan bir metrik tarafından düşük puanlanabilir; oysa insan değerlendiriciler bu tür içerikleri daha yüksek kaliteli olarak değerlendirebilir. Bu nedenle, bu metriklerin kullanımı pratik olsa da, sınırlılıkları dikkate alınmalı ve mümkünse insan değerlendirmeleri veya daha gelişmiş öğrenmeli metriklerle desteklenmelidir [6].

Bu sebeplerle, daha az maliyetli ve ölçeklenebilir olan otomatik metrikler tercih edilebilmektedir. İlk ortaya çıkan otomatik metriklerden biri olan BLEU (Bilingual Evaluation Understudy), başlangıçta makine çeviri çıktılarının değerlendirilmesi için geliştirilmiştir [41]. Ancak BLEUyalnızca kelime düzeyinde benzerlikleri ölçmektedir ve özellikle yaratıcı, kısa veya bağlamdan bağımsız metinlerde yetersiz kalmaktadır [42]. Hatta sadece BLEU'ya dayalı sistemlerin, makine çevirisi gibi kendi alanlarında bile model performansını sınırlandırabildiği gösterilmiştir [43, 44].

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metriği ise özetleme gibi görevlerde yaygın şekilde kullanılır ve çıktılar ile referans metinler arasındaki örtüşen kelimeleri ölçer [45]. Ancak ROUGE'un da bazı sınırlılıkları vardır: anlamsız olsa dahi aynı kelimeleri içeren [45] ve uzun özetleri yüksek puanlama eğilimindedir [45].

Bunun yanında, literatürde METEOR, NIST, ROUGE-N ve ROUGE-L gibi

metrikler de yer almaktadır; ancak bunların çoğu, kelime veya n-gram düzeyinde benzerliklere dayandığı için bağlamsal anlamı yeterince yansıtamamaktadır [6].

ROUGE-N, ROUGE metriğinin bir parçasıdır ve üretilen metinler ile insan tarafından yazılmış (referans) metinler arasındaki n-gram örtüşmesini ölçen sekiz farklı n-gram sayma yöntemini içerir [46].

ROUGE-L, yine ROUGE metriğinin bir parçası olarak, en uzun ortak alt dizi (Longest Common Sub-sequence - LCS) kullanarak iki metin arasındaki en uzun eşleşen kelime dizisini ölçer. ROUGE-L, eşleşen kelimelerin metin içinde ardışık olması gerekmeyen en uzun ortak diziyi bulmaya odaklanır[46].

METEOR (Metric for Evaluation of Translation with Explicit ORdering), BLEU'nun bazı kısıtlarını gidermek amacıyla geliştirilmiş[47] ve makine çevirisi, görüntü altyazılama, soru oluşturma ve özetleme gibi görevlerde model değerlendirmesi için yaygın olarak kullanılan bir metriktir. İnsan değerlendirmeleriyle yüksek korelasyon göstermesiyle dikkat çeker[6]. Sadece kesin kelime eşleşmelerini değil, aynı zamanda kök ve eşanlamlı kelime eşleşmelerini de dikkate alarak referans çevirilerdeki varyasyon sorununu azaltır. Değerlendirmeyi, kelime kesinliği ve geri çağırmanın harmonik ortalamasına göre yapar ve geri çağırmayı daha yüksek ağırlıklandırır. METEOR, BLEU'nun aksine, corpus düzeyinde değil, cümle düzeyinde karşılaştırma yapmak üzere tasarlanmıştır[6].

NIST, ABD Ulusal Standartlar ve Teknoloji Enstitüsü tarafından önerilmiş bir değerlendirme yöntemidir. Metin kalitesini ölçmek için BLEU'ya benzer şekilde n-gram eşleşmelerine dayanır. Ancak BLEU'nun her n-gram'ı eşit ağırlıkla değerlendirmesinin aksine, NIST nadir görülen n-gram'lara daha fazla ağırlık verir. Bunun nedeni, seyrek görülen n-gram'ların birlikte ortaya çıkmasının, yaygın n-gram'lara kıyasla daha fazla bilgi taşıdığı varsayımıdır [6].

Makine öğrenmesi tabanlı metrikler bu noktada daha güçlü bir alternatif sunmaktadır. Özellikle BERTScore gibi önceden eğitilmiş dil modelleri, metinler arasındaki semantik benzerliği daha iyi yansıtarak insan değerlendirmeleriyle yüksek korelasyon göstermiştir [45]. BLEURT gibi diğer derin öğrenme tabanlı metrikler de benzer şekilde bağlamsal benzerlikleri yakalamaya çalışır [6].

BLEURT, doğal dil üretimi (NLG) sistemlerinin değerlendirilmesi amacıyla geliştirilmiş, BERT tabanlı ve makine öğrenimiyle eğitilen bir otomatik değerlendirme metriğidir. Modelin eğitimi iki aşamalı bir süreçten oluşur. İlk aşamada, bir BERT modeli alınır ve BLEU ya da ROUGE gibi geleneksel metrikler

kullanılarak oluşturulan sentetik cümle çiftleri üzerinde ön ince ayar yapılır. Bu sentetik veriler, modelin üretim kalitesindeki değişimlere karşı daha dayanıklı hale gelmesini sağlamak amacıyla kullanılır. Ardından, ikinci aşamada, gerçek sistem çıktıları ve insan yazımı referans cümleler üzerinden, insan puanlamaları ve otomatik metrik skorları etiket olarak kullanılarak model daha ayrıntılı biçimde eğitilir. Bu çift aşamalı eğitim yaklaşımı, BLEURT'un özellikle zorlu üretim görevlerinde farklı sistemler arasındaki kalite farklarını daha isabetli şekilde ayırt edebilmesini sağlar. Ayrıca yapılan deneyler, eğitim verisinin dengesiz ya da sistem dışı (out-of-domain) olduğu durumlarda BLEURT'un ön eğitim olmadan ciddi bir performans kaybı yaşadığını ortaya koymuştur. Ancak bu ön eğitim aşaması sayesinde, model kalite dalgalanmalarına karşı daha dirençli hale gelmekte ve insan değerlendirmeleriyle yüksek düzeyde korelasyon göstermektedir. Bu özellikleriyle BLEURT, yalnızca yüzeysel benzerliklere değil, daha derin anlamsal uyuma da duyarlı bir metrik olarak öne çıkmaktadır [6].

BERTScore, metinler arasındaki anlamsal benzerliği değerlendirmek amacıyla geliştirilmiş bir metriktir ve BERT modelinin önceden eğitilmiş bağlamsal temsillerini (contextual embeddings) kullanır. Bu yöntemde, aday ve referans cümlelerdeki kelimeler, vektör temsilleri arasındaki kosinüs benzerliği aracılığıyla eşleştirilir. Böylece değerlendirme, yalnızca yüzeysel kelime eşleşmeleriyle sınırlı kalmayıp, kelimelerin bağlam içindeki anlamlarını da dikkate alır. Yapılan çalışmalar, BERTScore'un hem cümle düzeyinde hem de sistem düzeyinde insan yargılarıyla yüksek oranda örtüştüğünü göstermektedir. Ayrıca bu metrik, kesinlik (precision), geri çağırma (recall) ve F1 skoru gibi ölçümleri de hesaplayarak, özellikle doğal dil üretimi (NLG) görevlerinde kapsamlı bir performans değerlendirmesi sunar [6].

Son dönemde ise literatürde, dil modellerinin kendisinin değerlendirme aracı olarak kullanıldığı LLM-as-a-Judge (LLM'lerin değerlendirme aracı olarak kullanılması) yaklaşımı öne çıkmaktadır. GPT tabanlı modellerin, makine çevirisi, özetleme, diyalog üretimi ve soru-cevap gibi doğal dil üretimi (NLG) görevlerinde değerlendirme amaçlı olarak başarıyla kullanılabileceği gösterilmiştir [48–50]. LLM'lerin değerlendirme görevlerinde talimatlara uyum düzeyini ve insan yargılarıyla olan hizalanmasını inceleyen meta-analiz niteliğinde araştırmalar da bulunmaktadır [51].

Bununla birlikte, bu modellerin kendi ürettikleri metinleri değerlendirirken taraflı (biased) davranabildikleri ortaya konmuştur [52]. Bu nedenle, değerlendirme sırasında kullanılan LLM'nin, üretici modelden farklı olması tercih edilmektedir. Örneğin, bu çalışmada üretim için Google Gemini Pro, değerlendirme için ise

GPT-4o kullanılmıştır.

Ayrıca, bildirim gibi kısa, yaratıcı ve aksiyona yönlendiren metinler, klasik metriklerin sınırlarını daha da belirgin hale getirmektedir [7]. Bu tarz metinlerde anlamsal uygunluk, psikolojik tetikleyiciler ve bağlamsal ilgililik gibi ölçütler öne çıkmakta ve geleneksel metrikler bu karmaşık yapıyı ölçmede yetersiz kalabilmektedir [42, 45]. Bu nedenle, bildirim değerlendirmesi özelinde, klasik metriklerin sınırlılıkları dikkate alınarak yeni nesil, bağlamsal ve yaratıcı metinlere duyarlı yöntemlere ihtiyaç duyulmaktadır. Ayrıca, bildirim gibi kısa, yaratıcı ve aksiyona yönlendiren metinlerin değerlendirilmesi, geleneksel otomatik metriklerin sınırlılıklarını daha da görünür kılmaktadır. BLEU gibi yüzeysel benzerliğe dayalı metrikler, bu tür metinlerde belirleyici olan yaratıcılık, çeşitlilik ve pragmatik etkiyi göz ardı etme eğilimindedir. Özellikle modelin hep aynı kalıpta ama referansa çok yakın metinler üretmesi durumunda yüksek skorlar alınabilmesi, modelin üretim yetkinliğini eksik ya da yanıltıcı şekilde yansıtabilir. Bu da metriklerin çeşitlilik açısından yetersizliğini göstermektedir.

1.2 Tezin Amacı

Bu tez çalışmasının temel amacı, kullanıcılarla dijital platformlar arasında gerçekleşen etkileşimlerin en önemli bileşenlerinden biri olan bildirimlerin, Büyük Dil Modelleri (LLM'ler) kullanılarak nasıl daha etkili üretilebileceğini araştırmaktır. Bildirimler, kullanıcılara yalnızca bilgi vermekle kalmayıp, onları belirli bir eyleme yönlendirme amacı güden, kısa, yoğun ve hedef odaklı metinlerdir. Bu bağlamda, bildirim üretimi görevi; klasik metin üretimi görevlerinden, taşıdığı eylem çağrısı (call-to-action), sınırlı uzunluk ve dikkat çekici olma gerekliliği gibi nedenlerle ayrışmaktadır.

Doğal Dil İşleme (NLP) alanında son yıllarda yaşanan gelişmelerle birlikte, LLM'ler farklı metin üretimi görevlerinde yüksek performans göstermeye başlamıştır. Ancak mevcut literatürde, bildirim üretimi gibi yaratıcı ve kısa metin türlerine odaklanan sistematik çalışmalar oldukça sınırlıdır. Bu tez çalışması, LLM'lerin yalnızca uzun metin üretimi gibi geleneksel görevlerde değil, kısa metin üretiminde de etkili kullanılabileceğini göstermeyi hedeflemektedir.

Bu doğrultuda tezde, istem (prompt) mühendisliği çerçevesinde üç farklı yaklaşım karşılaştırılmıştır:

- **Elle Yöntem:** İstemlerin tamamen insan eliyle, görev ve hedef odaklı hazırlanması.

- **Otomatik Yöntem:** İstemlerin bir sistem tarafından otomatik olarak üretilmesi.
- **Optimize Yöntem:** Otomatik üretilen istemlerin çeşitli optimizasyon teknikleri ile iyileştirilmesi.

Bu yöntemlerin karşılaştırılması yalnızca çıktı metinlerinin dilsel kalitesi açısından değil, aynı zamanda bağlamsal uygunluk, kullanıcıya hitap edebilirlik ve eylem çağrısı niteliği gibi bildirim türüne özgü ölçütler bağlamında da gerçekleştirilmiştir. Geleneksel metin değerlendirme metrikleri (BLEU, ROUGE, METEOR gibi) bu tür metinlerde sınırlı kalabildiğinden, bu tezde ayrıca LLM tabanlı yeni nesil değerlendirme metrikleri de geliştirilmiş ve kullanılmıştır. Bu metrikler, klasik ölçütlerin yakalayamadığı anlamsal uygunluk, dilbilgisel akıcılık ve kullanıcı algısı gibi yönleri daha etkin biçimde ölçmeyi hedeflemektedir.

Tezin bir diğer amacı ise, üretken yapay zeka modellerinden alınan çıktılara uygulanabilecek otomatik değerlendirme tekniklerini bildirim üretimi gibi özgül içerik türlerine uyarlamaktır. Böylece hem metin üretimi sürecinde insan emeğinin azaltılması hem de elde edilen çıktının kalitesinin nesnel biçimde değerlendirilebilmesi sağlanmaktadır.

Sonuç olarak bu çalışma, hem üretim süreci (farklı istem stratejileri) hem de değerlendirme süreci (geleneksel ve LLM tabanlı metrikler) açısından kapsamlı bir yaklaşım sunarak, LLM'lerin kısa, hedef odaklı ve yaratıcı metinler üretme konusundaki potansiyelini ortaya koymayı amaçlamaktadır.

1.3 Hipotez

Bu tez çalışması kapsamında test edilmek üzere aşağıdaki hipotezler öne sürülmüştür:

1. Optimize Yöntem, bildirimin üretim kalitesi açısından Elle Yöntem ve Otomatik Yöntem'e göre daha yüksek başarı sağlar. Özellikle, dilbilgisel doğruluk, içerik–başlık uyumu ve eyleme yönlendiricilik gibi kriterlerde istatistiksel olarak anlamlı bir üstünlük gösterir.
2. LLM tabanlı değerlendirme metrikleri, bildirim gibi kısa ve yaratıcı metinlerin değerlendirilmesinde, ROUGE ve BLEU gibi klasik metriklere göre daha yüksek doğruluk ve insan değerlendirmeleriyle daha güçlü korelasyon sağlar.

3. Few-shot bağlam içi öğrenme stratejisi, özellikle bildirim üretimi gibi bağlam gerektiren görevlerde, Zero-shot yaklaşıma kıyasla daha yüksek kalitede metinler üretir. Verilen örneklerin sayısı ve çeşitliliği, üretilen içeriğin kullanıcıya uygunluğunu doğrudan etkiler.

4. Bildirimlerin kısa ve yaratıcı doğası gereği, yalnızca dilbilgisel doğruluk değil, dikkat çekicilik, kişiselleştirme düzeyi ve psikolojik tetikleyicilere dayalı çekicilik gibi nitelikler de performansın değerlendirilmesinde belirleyici rol oynar.

5. Değerlendirme amacıyla kullanılan dil modeli (örneğin GPT-4o), üretim amacıyla kullanılan modelden (örneğin Gemini Pro) farklı olduğunda, LLM-as-a-Judge yaklaşımı daha tutarlı ve tarafsız sonuçlar üretir. Aynı modelin üretici ve değerlendirici olarak kullanılması durumunda yanlılık (bias) oluşma riski artar.

6. Otomatik istem üretimi ve optimizasyonu, elle oluşturulan istemlere göre hem daha az insan emeği gerektirir hem de belirli görev türlerinde daha tutarlı ve yüksek kaliteli çıktılar üretir. Özellikle metinlerdeki bağlamsal bütünlüğün ve görev başarımının artırılmasında optimize istemlerin katkısı anlamlıdır.

2
VERİ KÜMESİ

Bu çalışmada kullanılan veri kümesi, bir şirketin pazarlama ekibi tarafından daha önceden hazırlanmış 17 farklı kampanyaya dayanmaktadır. Her kampanya, hedef kullanıcı davranışlarının analizi sonucunda alt kitlelere ayrılmış; bu alt kitlelere uygun vurgu ve üslup değerleri atanarak kişiselleştirilmiş kampanya varyasyonları tanımlanmıştır. Her kampanya için 5 farklı varyasyon oluşturulmuş, toplamda 85 varyasyon elde edilmiştir.

Varyasyonlar elle yazılmamış, bunun yerine kampanya ve kitle özelliklerine göre tanımlarla kodla, otomatik bir şekilde bir şekilde üretilmiştir. Bu varyasyonlar herhangi bir etiket içermemekte ve bu varyasyonlara karşılık gelen, elle üretilmiş bir bildirim bulunmamaktadır.

Vurgu değerlerine karşılık gelen tanım cümleleri, büyük dil modellerine (LLM'lere) verilecek istemlerde kullanılmış ve hem otomatik hem de optimize yöntemlerde girdileri oluşturmuştur. Her vurgu tanımı için örnek bir bildirim metni de sağlanmıştır.

Veri kümesine herhangi bir ön işleme (preprocessing) uygulanmamış, tanımlar doğrudan model girdisi olarak kullanılmıştır.

Tanım cümleleri, çeşitli vurgu değerlerine karşılık olarak belirlenmiş ve çalışmada otomatik ve optimize yaklaşımlarda istem girdisi olarak kullanılmıştır. Herhangi bir ön işleme (preprocessing) adımı uygulanmamıştır.

Vurgu ve üslup değerleri, pazarlama ekibi tarafından belirlenen tanımlarla desteklenmiştir. Bu tanımlar, LLM istem girdileri için kullanılmış ve her vurgu değeri için örnek bir bildirim cümlesi de sağlanmıştır. Veri kümesine herhangi bir ön işleme (preprocessing) uygulanmamış; tanımlar doğrudan model girdisine eklenmiştir. Bu vurgu ve üslup tanımları ek bölümünde yer almaktadır.

Ayrıca bir kural olarak ürün bulunan varyasyonlar, ürün odaklı vurgular yaparken,

ürün bulunmayanlar doğrudan iş hedeflerine odaklanmaktadır. Her kampanyanın hedef kitlesi farklı olduğu için tüm kampanyalarda aynı alt kitleler yer almayabilir.

Her bir varyasyon şu alfanümerik alanlardan oluşmaktadır:

- **Kampanya Kodu:** Kampanyanın benzersiz kimliği.
- **Vurgu:** Hedef kitlenin özelliğini vurgulayan bir iletişim stratejisi (Örn: Kayıp Müşteri Kümesi, İndirim Odaklı Müşteri Kümesi).
- **İndirim:** Kampanyaya ait indirim bilgisi (Örn: %40 indirim, Ücretsiz Teslimat).
- **İndirimli Ürün:** Kampanyada vurgulanan spesifik bir ürün ya da kategori (boş da olabilir).
- **Üslup:** İletişim diline dair tercih (Örn: Güvence Veren, Heyecan Verici).
- **İş Hedefi:** Kampanyanın iş amacı (Örn: Yeni Kullanıcı Kazanımı, Müşteri Geri Kazanımı, Alışkanlık Oluşturma).

Vurgu ve üslup değerleri hedef kitleye özel seçilmektedir.

Varyasyon veri kümesindeki farklı indirim türleri arasında 100 TL indirim, %40 indirim, "3 al 2 öde" promosyonu, ücretsiz teslimat ve ikinci üründe %50 indirim gibi seçenekler bulunmaktadır. Vurgu (emphasize) değerleri arasında Churn, Convenience Seekers, Loyal and Weekly Shoppers, Promo-Driven Cluster ve Generic gibi gruplar yer almaktadır. Üslup (tone of voice) değerleri ise Reassuring, Exciting, Friendly ve Urgent gibi iletişim tonlarını kapsamaktadır. İş hedefleri arasında Acquisition, Churn ve Habit Building yer almaktadır.

Tablo 2.1'de örnek iki varyasyon gösterilmektedir.

Metriklerin otomatik çıkarılabilmesi için girdi olarak kullanılacak bir varyasyon veriseti gerekli olup veriseti kampanyalara göre -bir kampanyanın tüm varyasyonları aynı parçada kalacak şekilde- 3 parçaya ayrılmıştır(%60 Eğitim, %20 Doğrulama ve %20 Test).

2.0.1 Kaynak ve Toplama Hikâyesi

Veri kümesi, Şirket (adı saklı tutulacaktır) pazarlama ekibinin 2020 Q3–2025 Q1 döneminde yürüttüğü 17 ayrı kampanyadan türetilen toplam 85 varyasyondan oluşmaktadır. Her kampanya, kullanıcı davranış verilerine dayalı segment-bazlı

Tablo 2.1 Varyasyon Örnekleri

Alan	Değer
Kampanya Kodu	YUZDE50_600TL_V37
Vurgu	'Churn' - Geri dönmenin faydalarını, kişiselleştirilmiş teklifler ve teşviklerle vurgulamak. Örn: "Sizi Özledik!", "Yeni Özellikler Sizi Bekliyor!"
İndirim	%50
İndirimli Ürün	Atıştırmalık
Üslup	Güvence Veren
İş Hedefi	Müşteri Geri Kazanımı

Alan	Değer
Kampanya Kodu	MARKET_FREE_DEL_1410
Vurgu	'Convenience Seekers' - Hızlı ve pratik hizmeti vurgulamak. Örn: "Kolaylık Ayağınıza Geldi!", "Hızlı Teslimat, Zahmetsiz Alışveriş!"
İndirim	3 al 2 öde
İndirimli Ürün	Doygun Artizan Zeytinli Ekmek (400 g)
Üslup	Güvence Veren
İş Hedefi	Alışkanlık Kazanımı

analiz ile alt kitlelere ayrılmış, bu alt kitlelere özel Vurgu ve Üslup değerleri atanarak kişiselleştirilmiş varyasyonlar oluşturulmuştur. Böylece aynı kampanya meta verileri (örn. indirim, ürün) sabit kalırken, her alt segment için ayrı mesaj varyasyonu üretilmiştir.

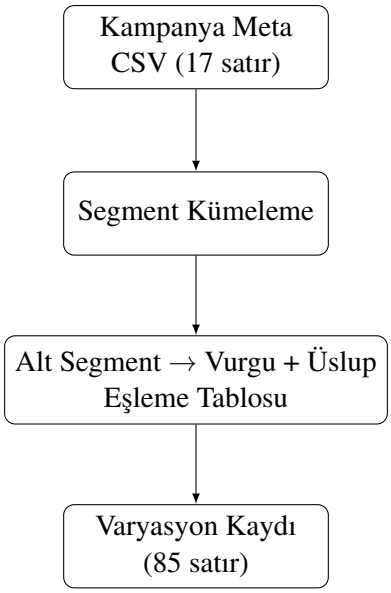
Bu varyasyonlar, Tablo 2.2'de özetlenen altı alanı içeren kayıtlardan ibarettir; her kayıt henüz tam bir bildirim metni barındırmaz, yalnızca metin üretiminde kullanılacak özellikleri içerir. Üslup ve Vurgu alanları, kampanya hedef kitlesinin alt segmentlerine (örn. Churn, Promo-Driven, Loyal) özgü ton ve vurgu bilgisini içerirken; Kampanya Kodu, İndirim, İndirimli Ürün ve İş Hedefi alanları kampanyanın kendisine ait üst düzey meta verilerdir.

Örnek Senaryo: Varsayalım MAR_FREE_DEL_1410 kodlu bir kampanyanın hedef kitlesi 300 bin kullanıcıdan oluşuyor. Segmentasyon sonucu bu kitle, Churn, Promo-Driven, Loyal ve Convenience Seekers olmak üzere dört alt segmente bölünüyor. Kampanya meta verileri (İndirim = "%40", İş Hedefi = Churn) herkes için aynı kalırken, dört farklı Vurgu + Üslup kombinasyonu sayesinde dört özgün varyasyon elde ediliyor.

Tablo 2.2 Veri Kümesi Şeması – Kampanya vs. Segment Alanları

Kampanya Alanları		
Alan	**Açıklama**	**Örnek Değer**
Kampanya Kodu	Benzersiz kimlik	YUZDE50_600TL_V37
İndirim	İndirim biçimi	%50
İndirimli Ürün	Ürün/Kategori (ops.)	Atıştırmalık
İş Hedefi	Üst düzey hedef	Müşteri Geri Kazanımı
Segment Alanları		
Alan	**Açıklama**	**Örnek Değer**
Vurgu	Segmenti motive eden tema	Churn – "Sizi Özledik!"
Üslup	İletişim tonu	Güvence Veren

Şekil 2.1'te özetlenen kampanya tabanlı varyasyon üretim sürecinde, kampanya meta verileri (17 satır) bir Python betiğiyle `.csv` şablonuna aktarılmış, her kampanya satırı $1-N$ alt segmente ayrılmış ve her alt segment, önceden tanımlı *Vurgu* ile *Üslup* etiketleriyle otomatik olarak eşleştirilmiştir. Böylece Tablo 2.2'deki altı alan eksiksiz doldurularak toplam 85 varyasyon kaydı elde edilmiş; süreçte herhangi bir ön-temizlik ya da manuel filtreleme adımı uygulanmamıştır.



Şekil 2.1 Kampanya tabanlı varyasyon üretim akışı.

Tüm işlem, kişisel kullanıcı verisi içermeyen, anonim ve toplulaştırılmış ticari meta veriler üzerinde yürütülmüştür. Veri setinin tez kapsamında kullanılmasına yönelik sözlü onay, 8 Ocak 2024 tarihinde Şirket Data Direktörü tarafından verilmiştir.

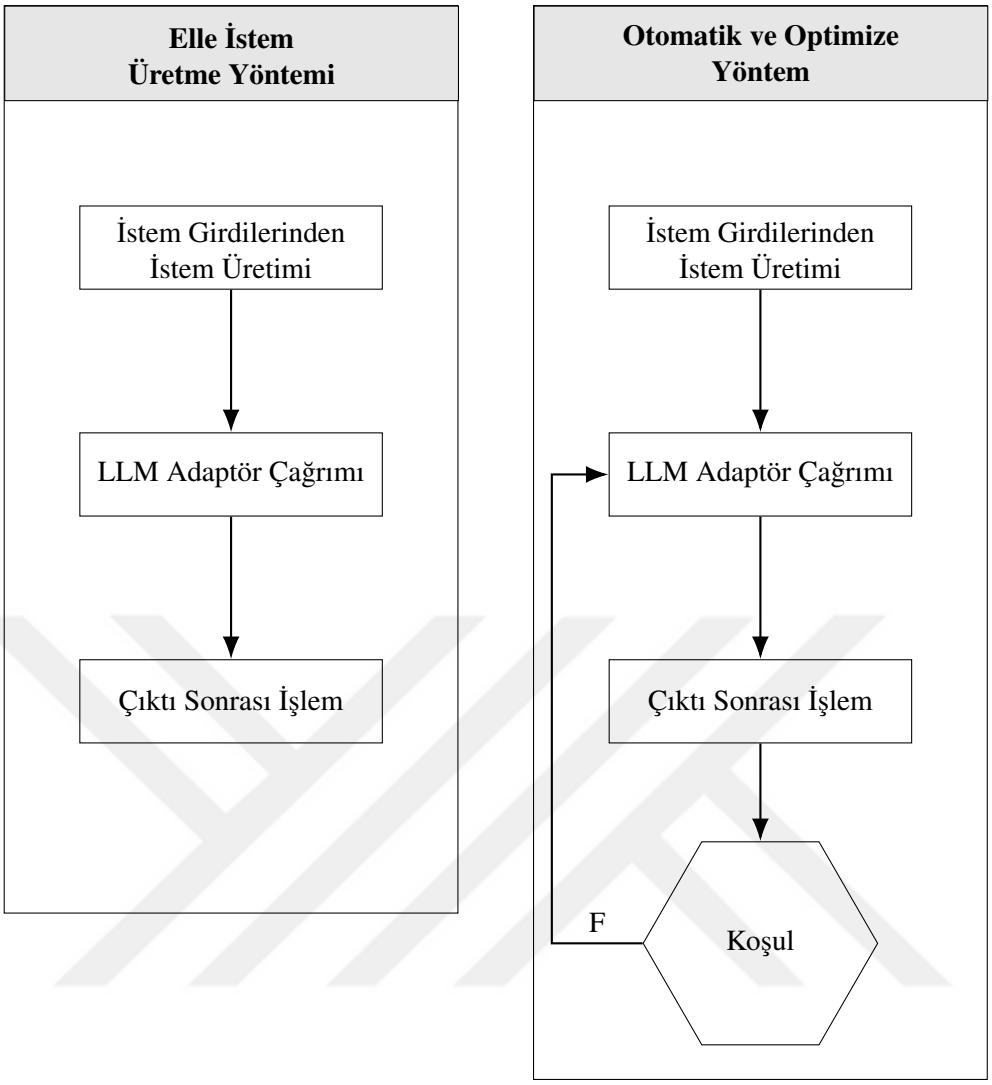
3
YÖNTEM

Bu tez kapsamında, kullanıcının ilgisini çekecek ve dikkat dağıtmadan aksiyona yönlendirecek etkili bildirim başlığı ve metinlerinin üretimi amacıyla Büyük Dil Modelleri (LLM) tabanlı metin üretim yöntemleri incelenmiştir. Bildirimlerin, kullanıcı etkileşimini artırabilecek şekilde hem dikkat çekici hem de doğal bir yapıya sahip olması hedeflenmiş; bu doğrultuda metinlerin gerektiğinde emoji gibi dikkat çekici ögeler içermesi, ancak bu ögelerin kullanımında aşırılığa kaçılmaması gözetilmiştir.

Çalışma sürecinde yapılan ön denemelerde, Türkçe metin üretiminde yüksek başarı gösterdiği gözlemlenen Google Gemini Pro ve Turkish Gemma 9b v0.1 dil modelleri tercih edilmiştir. Modelin yönlendirilmesi amacıyla kullanılan istemlerin (prompt) oluşturulmasında üç farklı yaklaşım benimsenmiştir:

- **Elle Yöntem:** İstem mühendisliği teknikleri kullanılarak, uzmanlarca elle hazırlanan şablonlar aracılığıyla istemlerin oluşturulması.
- **Otomatik Yöntem:** İstemlerin, kod tabanlı araçlar veya algoritmalar aracılığıyla ve herhangi bir metrik hedefi gözetilmeksizin otomatik olarak oluşturulması.
- **Optimize Yöntem:** Otomatik olarak üretilen istemlerin, belirli metrik hedefler ve kısıtlar doğrultusunda optimize edilerek geliştirilmesi.

Şekil 3.1, bu çalışmada karşılaştırmalı olarak değerlendirilen üç farklı bildirim üretim yönteminin süreç adımlarını özetlemektedir. Sol sütunda yer alan Elle Yöntem, uzmanlar tarafından az örnekli istemleme (few-shot prompting) yaklaşımıyla elle oluşturulan istemlerin, öncelikle dil modeli adaptörüne aktarılması ve ardından model çıktılarının işlenmesiyle tamamlanmaktadır. Süreç, sırasıyla üç adımdan oluşur: (i) bağlamsal özellikleri dikkate alarak uzman tarafından elle istem oluşturulması, (ii) bu istemin LLM adaptörüne iletilerek



Şekil 3.1 Elle, Otomatik ve Optimize Yöntemlerin Karşılaştırmalı Süreç Akışı

model yanıtının elde edilmesi, (iii) model çıktısına küçük düzenlemeler veya son işlemler uygulanması (örneğin noktalama düzeltmeleri veya emoji yerleştirme). Bu yöntemde işlem hattı doğrusaldır; çıktı kalitesine dair bir değerlendirme yapılmaz, dolayısıyla geri besleme veya optimizasyon döngüsü içermez.

Sağ sütunda yer alan yapı ise hem Otomatik Yöntem hem de Optimize Yöntem için ortak işlem adımlarını göstermektedir. Bu süreçte istem, insan müdahalesi olmadan otomatik olarak oluşturulur ve doğrudan LLM adaptörüne iletilerek model çıktısı elde edilir. Elde edilen çıktı, tıpkı Elle Yöntem'de olduğu gibi çıktı sonrası işlem (post-process) adımına tabi tutulur. Ancak burada sürece ek olarak bir kontrol aşaması da dâhildir. Bu aşama, model çıktısının istenilen kalite düzeyini karşılayıp karşılamadığını değerlendirir. Eğer çıktı yetersiz bulunursa (şekilde F ile temsil edilmektedir), süreç yeniden başa alınır ve yeni istem varyasyonları denenir.

Bu geri besleme döngüsü yalnızca Optimize Yöntem kapsamında aktiftir. Optimize Yöntem'de, istemler MIPROv2 gibi tekniklerle sistematik biçimde iyileştirilir; görev uyumluluğu ve bağlamsal doğruluğu artırılır. Öte yandan, Otomatik Yöntem'de bu optimizasyon süreci devre dışıdır; istem yalnızca bir kez üretilir ve kontrol aşaması bulunsa da çıktı başarısızsa süreç tekrarlanmaz.

Genel olarak, her üç yöntemde de model çağrısı ve çıktı sonrası işlem adımları yer almakta; ancak yalnızca Optimize Yöntem'de çıktı kalitesine bağlı olarak istem üretimi tekrarlanmakta ve iyileştirilmektedir. Bu yapı, yöntemler arasındaki istem oluşturma biçimi, geri besleme mekanizması ve çıktı işleme farklılıklarını açık biçimde ortaya koymaktadır.

3.1 Elle Yöntem ile Bildirim Üretimi

Bildirim üretimi, özelleşmiş bir metin üretimi problemidir ve bu alanda son yıllarda Büyük Dil Modelleri (LLM) tabanlı çözümler başarılı sonuçlar vermektedir.

LLM'ler ile metin üretiminde yaygın olarak kullanılan ilk yaklaşım, örnek verilmeden istemlerle yanıt üreten Örneksiz İstemleme (Zero-shot Prompting) yöntemidir. Ancak, örneksiz istemleme ve diğer tüm yöntemlerde etkili sonuçlar elde edebilmek için istemlerin dikkatlice tasarlanması ve optimize edilmesi kritik öneme sahiptir; bu süreç istem mühendisliği (prompt engineering) olarak adlandırılmaktadır. Modelin daha iyi performans göstermesi için örnek verilmesi gereken durumlarda ise bağlam içi öğrenme (in-context learning) tercih edilmektedir.

Bu tezde kullanılan Elle Yöntem, LLM tabanlı metin üretimi süreçlerinde yaygın olarak uygulanan istem mühendisliği tekniklerinden biri olan Az Örnekli İstemleme (Few-shot Prompting) yaklaşımına dayanmaktadır. Bu yöntemde, istemler uzmanlar tarafından elle tasarlanmış ve bağlam içi öğrenme yoluyla modelin bildirim formatına daha uygun ve isabetli çıktılar üretmesi sağlanmıştır.

Bu çalışmada, yaygın ve temel istemleme yöntemlerinden biri olan örneksiz istemleme (Zero-shot Prompting) yerine, modelin yönlendirilmesi amacıyla örnek bildirim metinlerinden yararlanılmış ve bu nedenle Az Örnekli İstemleme (Few-shot Prompting) tercih edilmiştir. Özellikle bildirim formatı gibi belirli yapısal kalıpların öğrenilmesinde örneklerin sağladığı yönlendiricilik, model performansını anlamlı şekilde artırmaktadır.

İstem mühendisliği (prompt engineering) sürecinde çeşitli denemeler

gerçekleştirilmiş ve elde edilen çıktılar doğrultusunda bir şablon (template) geliştirilmiştir. Her bir şablon; kampanyanın içeriğini, öne çıkarılması gereken bilgileri, hedef kullanıcı kitlesini ve metinde yer alması istenen unsurları (örneğin ilgi çekici ifadeler, uygun emojiler vb.) dikkate alacak biçimde yapılandırılmıştır. Ayrıca, şablon tasarımında metinlerin dil bilgisel olarak akıcı, doğal ve etkileyici olması gibi kriterler de önemli rol oynamıştır.

Veri kümesindeki varyasyon verileri, hazırlanan bu şablonlara uygun biçimde düzenlenmiş ve modele girdi olarak sunulmuştur. Model, bu girdilerle şablonu doldurarak nihai istemleri oluşturmaktadır.

3.2 Otomatik Yöntem ile Bildirim Üretimi

LLM (Large Language Model) tabanlı uygulamaların hızlı gelişimiyle birlikte, her gün daha iyi ve daha güçlü modellerin yayımlandığı bir döneme girilmiştir. Her LLM için en uygun istem (prompt) biçimi farklılık gösterebildiğinden, belirli bir uygulama için en iyi sonucu veren istemin her model değişiminde yeniden belirlenmesi gerekebilmektedir. Bu durum, elle istem mühendisliği sürecini hem zaman alıcı hem de maliyetli hale getirmektedir.

Bu problemi çözmek amacıyla, otomatik istem üretim yaklaşımları geliştirilmiştir. Bu çalışmada, otomatik istem oluşturmak için DSPy adlı framework kullanılmıştır. DSPy, dil modeli programları yazmaya ve istemleri otomatik olarak üretmeye olanak tanıyan bir Python çerçevesidir. Uygulamada kullanılan Predict modülü, kullanıcıdan gelen girdilere göre doğrudan çıktı üreten bir bileşen olarak öne çıkmaktadır.

Bu kapsamda uygulanan Otomatik Yöntem yaklaşımında, Örneksiz İstemleme (Zero-shot Prompting) kullanılmıştır. Predict modülü, örnek olmadan sadece görev tanımı üzerinden çıktı üretmektedir. Bu süreçte, herhangi bir istem optimizasyonu (örneğin, Chain-of-Thought veya Tree-of-Thought gibi tekniklerle istemlerin yapılandırılması) uygulanmamıştır. Bu nedenle elde edilen istemler, optimize edilmemiştir ve sonuç olarak elle hazırlanmış istemlere göre daha düşük performans göstermiştir.

Otomatik Yöntem'in temel amacı, insan müdahalesine ihtiyaç duyulmadan, modelin kendisinin belirli görevler için istem oluşturmasını ve çıktı üretmesini sağlamaktır. Bu yönüyle, yüksek tekrarlanabilirlik ve düşük maliyet gibi avantajlar sunsa da, performans açısından Elle Yöntem'in gerisinde kalmıştır.

3.3 Optimize Yöntem ile Bildirim Üretimi

Bu çalışmada, dil modeli (DM) programlarında istemlerin otomatik olarak optimize edilmesi amacıyla Örneksiz (Zero-shot) MIPROv2 yöntemi kullanılmıştır. Çok aşamalı DM programları için geliştirilen MIPROv2, her bir modüle ait talimatları iteratif olarak iyileştirirken, Bayesyen optimizasyon ilkelerinden yararlanarak istem konfigürasyonlarını optimize eder. Bu yöntem, istemlerin bağlama duyarlı ve göreve özgü olmasını sağlayan sistematik bir optimizasyon süreci sunmaktadır.

Optimizasyon süreci şu adımlardan oluşmaktadır:

- Görev tanımı ve metriklerin belirlenmesi
- GroundedProposer bileşeni ile bağlamsal istem varyasyonlarının oluşturulması
- Optuna kütüphanesi ile Tree-structured Parzen Estimator (TPE) algoritması yardımıyla en uygun istemlerin seçilmesi
- Mini-batch ve doğrulama seti değerlendirmeleri ile sonuçların ölçülmesi

Optimizasyon süreci, görev ve performans metriklerinin tanımlanmasıyla başlamaktadır. Bu metrikler, istemlerin kalitesini nesnel olarak ölçebilmek ve karşılaştırmalı analiz yapabilmek açısından önemlidir.

Her bir program modülü için aday talimatlar, GroundedProposer bileşeni tarafından üretilir. Bu bileşen, eğitim veri setinin özellikleri, program kodunun özeti ve rastgele seçilen istem ipuçları gibi bağlamsal bilgileri dikkate alarak üretim yapar.

Optimizasyon aşamasında, MIPROv2 Optuna çerçevesini kullanarak Bayesyen optimizasyon uygular. En uygun istem kombinasyonları, Tree-structured Parzen Estimator (TPE) algoritması kullanılarak belirlenir. Arama alanı, CategoricalDistribution kullanılarak tanımlanır ve programın çeşitli talimat kombinasyonlarına karşı performansı değerlendirilir.

Optimizasyon sürecine başlamadan önce MIPROv2, gerçekleştirilecek DM çağrılarının sayısını ve buna bağlı maliyetleri tahmin ederek kullanıcıdan onay talep eder. Sürecin tamamlanmasının ardından en yüksek performansı gösteren talimatları içeren optimize edilmiş DM programı çıktı olarak sunulur.

Bu yöntem, çalışmada Zero-shot senaryo altında uygulanmış ve herhangi bir ön bilgi olmaksızın istemlerin optimize edilmesi sağlanmıştır. Elde edilen sonuçlar,

optimize yöntemin hem elle oluşturulan hem de otomatik yöntemlerle oluşturulan istemlere kıyasla daha yüksek performans sergilediğini göstermektedir.

Optimizasyon süreci, görevle daha uyumlu ve bağlama duyarlı istemlerin elde edilmesini sağlamış; bu da dil modeli programlarının çıktı kalitesine doğrudan olumlu yansımıştır.

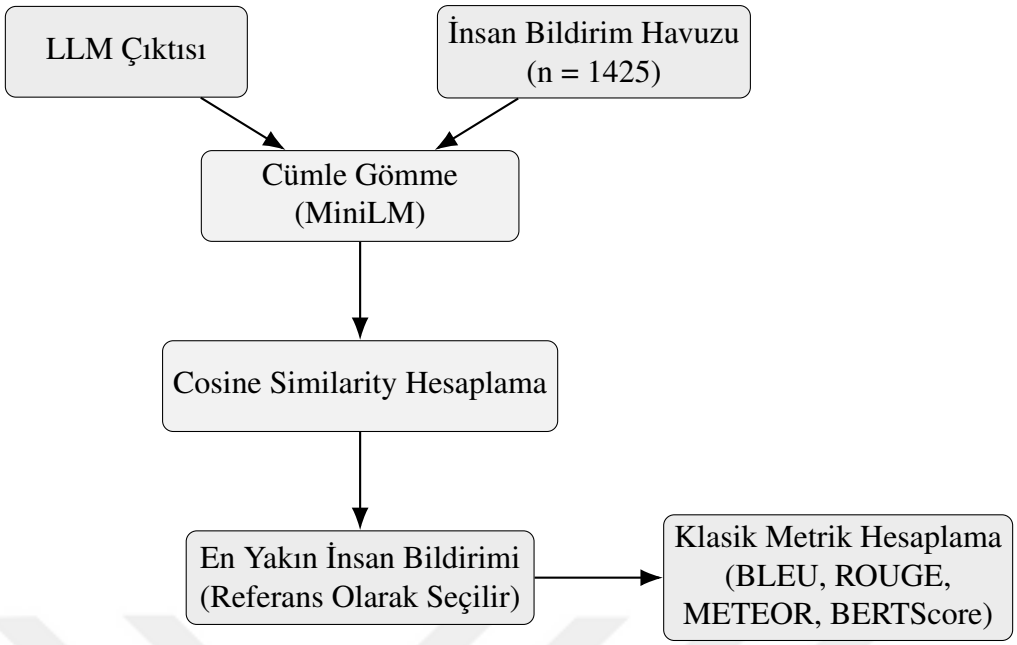
Not: Bu çalışmada kullanılan MIPROv2 yöntemi tarafımızca geliştirilmemiş olup ilgili teknik detaylar DSPy belgelerinde sunulmaktadır [53].

3.4 Bildirimlerin Değerlendirilmesi

Çalışma başlangıcında üretilen bildirimler insan değerlendirmesi (human evaluation) ile belli bir aşamaya getirilmiştir. Daha sonrasında daha hızlı iterasyon ve karşılaştırabilme amacıyla otomatik metriklerin kullanımına ihtiyaç duyulmuştur.

3.4.1 Klasik Otomatik Değerlendirme Metrikleri

Makine çevirisi ve metin özetleme gibi doğal dil işleme alanlarında yaygın olarak kullanılan BLEU, ROUGE, METEOR ve BERTScore gibi otomatik metrikler, bu çalışmada referans değerlendirme sağlamak amacıyla kullanılmıştır. Bildirim üretimi için birebir eşleştirilmiş bir golden output kümesi bulunmadığından, her LLM çıktısı için insan eliyle oluşturulmuş 1425 bildirimlik havuzdan cosine similarity değeri en yüksek metin referans olarak seçilmiştir(Şekil 3.2). Bu işlemde çok dilli cümle gömme (embedding) modeli `paraphrase-multilingual-MiniLM-L12-v2` kullanılmıştır. Böylece klasik metrikler, en yakın referans üzerinden tutarlı biçimde hesaplanabilmiştir. Bu metriklerin teorik temelleri ise şeklin devamında özetlenmiştir.



Şekil 3.2 Klasik metrikler için referans seçim süreci

3.4.1.1 BLEU (Bilingual Evaluation Understudy)

BLEU metriği, model çıktısı ile bir veya daha fazla referans metin arasındaki n-gram eşleşmelerini değerlendirerek üretim kalitesini ölçer. Eşitlik (3.1)'de gösterildiği üzere, BLEU skoru, kısaltma cezası (brevity penalty) ile logaritmik ortalamanın çarpımı olarak hesaplanır:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{3.1}$$

Bu formülde:

- p_n, n-gram düzeyinde doğru eşleşme oranını,
- w_n, her n-gram düzeyi için ağırlığı ($w_n = \frac{1}{N}$),
- BP, kısaltma cezasını ifade eder.

Kısaltma cezası, model çıktısının uzunluğunun referans uzunluğuna oranı dikkate alınarak Eşitlik (3.2)'de tanımlandığı şekilde hesaplanır:

$$\text{BP} = \begin{cases} 1 & \text{eğer } c > r \\ e^{(1-\frac{r}{c})} & \text{eğer } c \leq r \end{cases} \tag{3.2}$$

Burada c, model çıktısının uzunluğu, r ise referans metnin uzunluğudur. BLEU metriği kısa ve yaratıcı metinlerde düşük skor verme eğiliminde olabilir.

3.4.1.2 ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

ROUGE metrikleri, referans metin ile model çıktısı arasındaki kelime dizisi benzerliğini ölçer. Bu çalışmada ROUGE-N ve ROUGE-L metrikleri kullanılmıştır.

ROUGE-N metriği, n-gram düzeyinde örtüşme oranına dayanır ve Eşitlik (3.3)'de tanımlandığı şekilde hesaplanır:

$$\text{ROUGE-N} = \frac{\sum_{\text{gram} \in \text{Ref}} \min(\text{Count}_{\text{gen}}(\text{gram}), \text{Count}_{\text{ref}}(\text{gram}))}{\sum_{\text{gram} \in \text{Ref}} \text{Count}_{\text{ref}}(\text{gram})} \tag{3.3}$$

ROUGE-L metriği ise, model çıktısı ve referans metin arasındaki en uzun ortak alt diziyi (LCS) esas alır. Eşitlik (3.4), (3.5) ve (3.6)'de ilgili formüller verilmiştir:

$$\text{ROUGE-L}_{\text{recall}} = \frac{LCS(X, Y)}{\text{len}(Y)} \tag{3.4}$$

$$\text{ROUGE-L}_{\text{precision}} = \frac{LCS(X, Y)}{\text{len}(X)} \tag{3.5}$$

$$\text{ROUGE-L}_{F1} = \frac{(1 + \beta^2) \cdot \text{precision} \cdot \text{recall}}{\text{recall} + \beta^2 \cdot \text{precision}} \tag{3.6}$$

Burada X, model çıktısını; Y, referans metni; β ise genellikle 1 olarak alınan dengeleme parametresini ifade eder.

3.4.1.3 METEOR (Metric for Evaluation of Translation with Explicit ORdering)

METEOR metriği, eşleşmeleri yalnızca yüzeysel düzeyde değil; aynı zamanda kök, eşanlamlılık ve kelime sırası gibi dilsel faktörleri dikkate alarak değerlendirir. Eşitlik (3.7)'de tanımlandığı üzere, METEOR skoru, harmonik ortalama ile ceza katsayısının çarpımı olarak hesaplanır:

$$\text{METEOR} = F_{\text{mean}} \cdot (1 - \text{Penalty}) \tag{3.7}$$

Burada F_{mean}, precision ve recall'un harmonik ortalamasıdır. Ceza katsayısı (Penalty) ise kelime sırasındaki bozulmaları yansıtarak daha duyarlı bir değerlendirme sunar ve Eşitlik (3.8)'deki gibi tanımlanır:

$$\text{Penalty} = \gamma \cdot \left(\frac{ch}{m}\right)^{\theta} \tag{3.8}$$

Bu ifadede:

- ch, eşleşme bloklarının sayısını,
- m, eşleşen kelime sayısını,
- γ ve θ, önceden belirlenmiş sabitleri ifade eder.

3.4.1.4 BERTScore

BERTScore metriği, model çıktısı ve referans metindeki her kelimeyi, önceden eğitilmiş bir dil modeli (örneğin BERT) aracılığıyla vektör uzayında temsil eder. Bu temsiller arasındaki semantik benzerlik, cosine benzerliği kullanılarak hesaplanır. Eşitlik (3.9) ve (3.10)'te precision ve recall hesaplamaları verilmiştir:

$$\text{Precision} = \frac{1}{|X|} \sum_{x \in X} \max_{y \in Y} \cos(\mathbf{e}_x, \mathbf{e}_y) \tag{3.9}$$

$$\text{Recall} = \frac{1}{|Y|} \sum_{y \in Y} \max_{x \in X} \cos(\mathbf{e}_y, \mathbf{e}_x) \tag{3.10}$$

Burada:

- X, model çıktısındaki kelime kümesini,
- Y, referans metindeki kelime kümesini,
- $\mathbf{e}_x$ ve $\mathbf{e}_y$, sırasıyla X ve Y içindeki kelimelere ait gömme (embedding) vektörlerini ifade eder.

F1 skoru, bu precision ve recall değerlerinin harmonik ortalaması alınarak elde edilir.

3.4.1.5 Klasik Metriklerin Kullanım Gerekçesi ve Sınırlılıkları

Açıklanan bu metrikler, üretimlerin yapısal ve kısmen anlamsal benzerliğini değerlendirmek için kullanılmıştır. Ancak bildirim metinlerinin kısa, yaratıcı ve bağlamsal doğası gereği, bu metriklerin sınırlı kaldığı bazı durumlar gözlemlenmiştir. Özellikle:

- N-gram tabanlı metrikler (BLEU, ROUGE-N) varyasyon içeren ifadeleri cezalandırabilmektedir.
- METEOR bağlamı tam olarak yansıtamayabilir.
- BERTScore semantik yakınlığı ölçebilse de, hedef kitlenin algısı veya kampanya bilgisi gibi bağlamsal kriterlere duyarsızdır.

Bu nedenlerle, klasik metrikler referans değerlendirme sağlamak amacıyla dahil edilmiş ancak bağlamsal başarıyı daha doğru ölçebilmek için LLM tabanlı değerlendirme yaklaşımları da geliştirilmiştir.

3.4.2 Makine Öğrenmesi Tabanlı Yardımcı Metrikler

Bu tezde kullanılan bazı değerlendirme metrikleri, üretken dil modeli (LLM) içermeyen ancak önceden eğitilmiş gömme modelleri gibi makine öğrenmesi bileşenlerine dayanan yöntemler kullanılarak geliştirilmiştir. Bu tür metrikler, üretimlerin semantik yapısını anlamaya yönelik tamamlayıcı içgörüler sağlamaktadır.

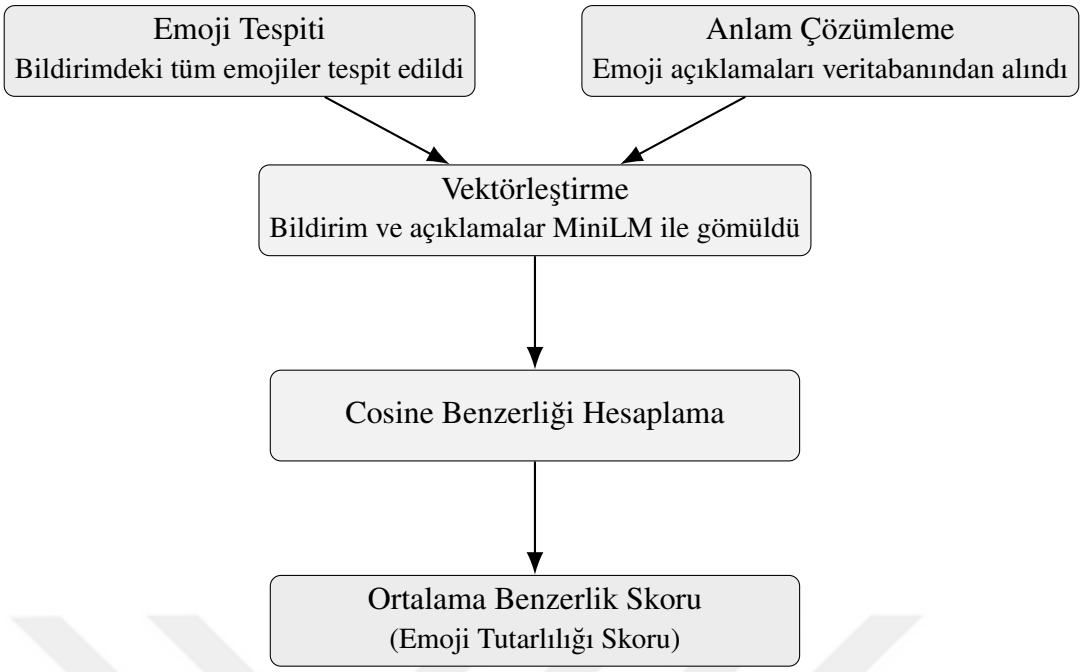
Klasik ve LLM tabanlı metriklerin sınırlılıklarını tamamlayıcı nitelikte tasarlanan bu metrikler, örneğin bildirimin içeriği ile kullanılan emojilerin veya hedeflenen iletişim tonu ile bildirimin üslubunun ne derece örtüştüğünü ölçmek gibi görevlerde kullanılmıştır. Bu bağlamda kullanılan metrikler, LLM tabanlı "emoji uygunluğu" gibi değerlendirmelerden farklı olarak, embedding modelleri ve makine öğrenmesi sınıflandırıcıları gibi araçlara dayanmaktadır.

3.4.2.1 Anlamsal Emoji Tutarlılığı

Bu metrik, bildirimin içinde kullanılan emojilerin, metnin genel anlamı ve tonu ile ne derece örtüştüğünü değerlendirmek amacıyla tasarlanmıştır.

Değerlendirme süreci aşağıdaki adımlarla gerçekleştirilmiştir:

1. Bildirim metninde yer alan tüm emojiler tespit edilmiştir.



Şekil 3.3 Anlamsal emoji tutarlılığı metriği için değerlendirme süreci

2. Her bir emoji için, ilgili Unicode karakterinin Türkçe açıklaması güvenilir bir kaynak olan `emoji` veritabanı kullanılarak elde edilmiştir.

3. Hem bildirim metni hem de emoji açıklamaları, çok dilli cümle gömme (sentence embedding) modeli olan `paraphrase-multilingual-MiniLM-L12-v2` kullanılarak vektör uzayına yansıtılmıştır.

4. Bildirim vektörü ile her bir emoji açıklamasının vektörü arasındaki cosine benzerliği hesaplanmış, ve bu benzerliklerin ortalaması alınarak nihai skor elde edilmiştir.

Bu yaklaşım sayesinde, yalnızca sembolik değil aynı zamanda anlamsal olarak da metne uygun emoji kullanımı teşvik edilmiştir. Örneğin, metnin duygusal tonuyla uyuşmayan veya içerikle alakasız emojiler daha düşük skorla değerlendirilmiştir. Hiç emoji içermeyen bildirime ise doğrudan 0 puan verilmiştir. Bu durum, modelin yalnızca doğru emojiyi seçmesini değil, uygun bağlamda emoji kullanımı kararını da doğru şekilde vermesini gerektirmektedir.

Gömme işlemlerinde kullanılan `paraphrase-multilingual-MiniLM-L12-v2` modeli, çok dilli metinlerde yüksek performans göstermesi ve Türkçe cümleler için de başarılı sonuçlar vermesi nedeniyle tercih edilmiştir. Bu model, düşük boyutlu ve hızlı olmasıyla üretim senaryolarına da uygundur. Ayrıca, Hugging Face

tarafından sağlanan açık kaynaklı sentence-transformers kütüphanesi ile uyumlu olması, hızlı entegrasyon ve çoğaltılabilirlik açısından avantaj sağlamaktadır.

3.4.3 Görev Odaklı LLM Tabanlı Metrikler

Çalışma başlangıcında üretilen bildirimler insan değerlendirmesi (human evaluation) ile belli bir aşamaya getirilmiştir. Daha sonrasında daha hızlı iterasyon ve karşılaştırabilme için otomatik metriklere ihtiyaç duyulmuştur.

Makine çevirisi gibi metin üretimi alanlarında yaygın kullanılan BLEU, ROUGE gibi metriklerin, bildirim üretimi gibi yaratıcı metinlerde yetersiz olduğu bilinmektedir. Bu nedenle şirketin (adı saklı tutulacaktır) pazarlama ve ürün ekipleriyle yapılan görüşmeler sonucunda aşağıdaki değerlendirme kriterlerinin uygun olduğuna karar verilmiştir:

- **Emoji Sayısı:** Uygun sayıda emoji kullanımını kontrol eder (1: uygun, 0: değil).
- **Emoji Uygunluğu:** Emojilerin, mesajın üslubu ve içeriği ile uyumunu ölçer (1-5 arası).
- **Dil Bilgisi Uygunluğu:** Bildirimlerin dil bilgisel doğruluğunu değerlendirir (1-5 arası). LLM kullanılarak metin düzeltilir, ardından orijinal metin ile düzeltilmiş metin arasındaki Levenshtein mesafesi ölçülerek skor hesaplanır.
- **Başlık ve İçerik Uyumu:** Bildirim başlığı ile içeriğin uyumunu ölçer (1-5 arası).
- **Doğallık:** Bildirimin akıcı ve doğal olup olmadığını değerlendirir (1-5 arası).
- **Kampanya Bilgilerinin Doğruluğu:** Kampanya bilgileri aktarımının doğruluğunu ölçer (1-5 arası).

Bu metriklerin her biri, model çıktısının gerçek kullanım bağlamındaki işlevselliğini değerlendirmeyi amaçlamaktadır. Bildirimlerin kalitesini etkilediği düşünülen bu metrikler, makine tabanlı metriklerden oluşmaktadır. Her bir metrik için doğrudan makine öğrenmesi modeli geliştirmek yüksek miktarda veri gerektirdiğinden ve zaman alıcı olduğundan, bunun yerine "LLM-as-a-judge" yaklaşımı benimsenmiştir. Bununla birlikte, yakın tarihli bir çalışmada [52], LLM'lerin değerlendirme yaparken kendi ürettikleri çıktılara sistematik olarak daha yüksek puan verme eğiliminde oldukları gösterilmiştir. Bu durum, özellikle

modelin hem üretici hem de değerlendirici olarak kullanıldığı senaryolarda taraflılığa yol açabileceğinden, değerlendirme sürecinde farklı bir modelin kullanılması önem arz etmektedir. Bu doğrultuda, değerlendirme amacıyla — üretimde kullanılan modelden farklı olarak — GPT-4o modeli tercih edilmiştir. Bununla birlikte, LLM'lerin doğrudan sayısal skor üretiminde istikrarsız sonuçlar verebilmesi nedeniyle, LLM yeteneklerinin güçlü olduğu alanlardan faydalanarak kullanıldığı hibrit değerlendirme stratejileri geliştirilmiştir. Örneğin, Dil Bilgisi Uygunluğu metriğinde, LLM'den metni düzeltmesi istenmiş, ardından orijinal ve düzeltilmiş metin arasındaki Levenshtein mesafesi skor olarak türetilmiştir. Benzer hibrit yaklaşımlar diğer metriklerde de uygulanmıştır.

3.5 Üretici Dil Modeli: Google Gemini 1.5 Pro

Google tarafından geliştirilen çok modlu büyük dil modeli olan Gemini 1.5 Pro (`gemini-1.5-pro-001`) bu tez kapsamında bildirim metinlerinin üretimi amacıyla kullanılmıştır. Model, Gemini 1.5 serisinin en güçlü üyelerinden biri olup hem metin hem de diğer modalitelerde üretim yapabilme kapasitesine sahiptir [54]. Bu çalışmada yalnızca metin üretimi yeteneklerinden faydalanılmıştır.

Gemini 1.5 Pro, mimari olarak klasik Transformer yapısını temel almakla birlikte, Google'ın uzun süredir üzerinde çalıştığı Sparse Mixture-of-Experts (MoE) tasarımını kullanır [54]. Bu yapıda, her katmanda birçok bağımsız uzman (expert) ağı bulunmakta, ancak her giriş örneği için yalnızca belirli bir alt küme (genellikle 2 uzman) etkinleştirilmektedir. Bu durum, toplam parametre sayısını artırırken koşullu hesaplama (conditional computation) sayesinde her bir hesaplamada yalnızca aktif parametrelerin kullanılmasıyla hesaplama verimliliği sağlanır. Bu yapı, modelin daha geniş kapasiteye sahip olmasını mümkün kılarken, donanımsal maliyeti kontrol altında tutar.

Modelin eğitimi, Google veri merkezlerinde bulunan yüksek performanslı TPUv4 donanımları üzerinde gerçekleştirilmiştir [54]. Her bir eğitim kümesi 4096 TPU çekirdeğinden oluşan devasa pod'lar üzerinde çalıştırılmış; bu altyapı sayesinde paralel işlem ve parametre paylaşımı sağlanmıştır. Eğitimin yazılımsal bileşenleri arasında, yüksek performanslı lineer cebir için kullanılan XLA (Accelerated Linear Algebra), çok ölçekli model paralelleştirmesi için GSPMD (General and Scalable Parallelism for ML) ve vektörleştirilmiş hesaplama için JAX yer almaktadır [54]. Bu üçlü, Gemini 1.5 Pro'nun büyük ölçekli eğitimi için optimize edilmiş bir eğitim çerçevesi sunmaktadır.

Model, ön eğitim (pretraining) sürecinde çok geniş ve çeşitli veri kaynaklarıyla

beslenmiştir. Bu kaynaklar arasında, lisanslı metinler, web belgeleri, kod verileri, görseller, ses ve video içerikleri gibi multimodal veri yer almaktadır [54]. Ön eğitim sonrasında model, insan tercihleri ile hizalanmış görevler üzerinde ince ayar (fine-tuning) işlemine tabi tutulmuştur. Bu süreçte görev özelinde bağlamsal anlam çıkarma, talimat izleme, kod üretimi ve mantıksal muhakeme gibi beceriler güçlendirilmiştir.

Gemini 1.5 Pro'nun öne çıkan teknik özelliklerinden biri, uzun bağlamları işleyebilme kapasitesidir. Model, 10 milyon tokene kadar olan girişleri performans kaybı olmadan işleyebilmektedir. Yapılan testlerde, modelin hatırlama oranı 1 milyon token'da %99,7; 10 milyon token'da ise %99,2 olarak rapor edilmiştir [54]. Bu kapasite, özellikle bağlamsal tutarlılığın kritik olduğu metin üretimi görevlerinde modelin güçlü performans sergilemesini sağlar. Her ne kadar bu tezde kullanılan bağlam uzunlukları yüzlerce veya binlerce token ile sınırlı kalsa da, modelin geniş bağlam hafızası içerik sürekliliğine katkı sunmuştur.

Güvenlik ve etik kullanım açısından da Gemini 1.5 Pro, önceki modellere kıyasla önemli iyileştirmeler içermektedir. Model, memorization oranı açısından daha düşük risk sergilemekte; yanıtlarda kullanıcı verisinin tekrar edilmesi olasılığı azalmıştır [54]. Ayrıca, jailbreak gibi kötüye kullanım senaryolarına karşı daha dirençli davranış göstermekte ve politik açıdan hassas içerik üretiminde daha sıkı kontrol mekanizmaları uygulanmaktadır.

Bu tez çalışmasında Gemini 1.5 Pro, iki farklı üretim konfigürasyonu ile kullanılmıştır:

- **Deterministik Üretim:** `top_k=1`, `top_p=0`, `temperature=0`
  Bu ayarlar, modelin yalnızca en yüksek olasılıklı çıktıyı üretmesini sağlayarak tutarlı ve tekrar edilebilir sonuçlar elde edilmesine imkân tanımıştır. Bu konfigürasyon, istem varyasyonlarının göreli başarımını karşılaştırmak için tercih edilmiştir.
- **Yaratıcı Üretim:** `temperature=0.8`
  Daha doğal, akıcı ve ilgi çekici metin üretimi hedeflendiğinde bu yapılandırma kullanılmıştır. Modelin içsel üretkenliğini artırarak farklı ve yaratıcı bildirim alternatifleri sunması amaçlanmıştır.

Gemini 1.5 Pro, kısa metin üretiminde semantik yoğunluk, dilsel akıcılık ve bağlamsal tutarlılık gibi metriklerde üstün performans göstermektedir. Ayrıca, çok dilli üretim, matematiksel muhakeme, kod üretimi ve talimat takibi gibi

görevlerde de önemli başarılar rapor edilmiştir [54]. Bu bağlamda, kısa form içerik olan bildirim metinlerinin üretiminde modelin bu güçlü özelliklerinden doğrudan yararlanılmıştır.

Sonuç olarak, Gemini 1.5 Pro; modern mimarisi, dağıtık eğitimi destekleyen hesaplama altyapısı, uzun bağlam taşıma kapasitesi ve etik güvenlik önlemleri sayesinde bu tezde önerilen bildirim üretim senaryosuna hem teknik hem de uygulama düzeyinde üst düzey katkı sunmuştur.

3.6 Üretici Dil Modeli: Turkish-Gemma-9b-v0.1

Google DeepMind tarafından geliştirilen açık kaynaklı Gemma 2 mimarisi üzerine inşa edilmiş ve Türkçe için özel olarak ince ayar uygulanmış olan Turkish-Gemma-9b-v0.1 modeli bu tez çalışmasında bildirim metinlerinin üretimi amacıyla kullanılmıştır. Kullanılan modelin mimari temelleri, eğitim süreci ve kullanım bağlamı aşağıda bütüncül olarak açıklanmıştır.

Gemma 2 modelleri, açık kaynaklı ve hafif yapıda sunulmak üzere tasarlanmış olup, boyutlarına kıyasla yüksek üretim kalitesi sağlamayı hedeflemektedir [55]. Bu modeller, mimari olarak decoder-only Transformer yapısını temel alır ve önceki sürüm olan Gemma 1 ile birçok özelliği paylaşsa da önemli yapısal ve eğitimsel iyileştirmeler içermektedir. En belirgin yeniliklerden biri, her Transformer katmanında ardışık olarak uygulanan yerel kayar pencere dikkat (sliding window attention) ve global dikkat (global attention) mekanizmalarıdır. Yerel dikkat, 4096 tokenlık bir bağlam penceresi içinde çalışırken; global dikkat, 8192 tokena kadar uzanan daha geniş bir bağlam bilgisi sağlar. Bu dikkat mekanizmaları, kısa ve bağlamsal olarak yoğun metin üretimi için optimize edilmiştir. Ayrıca, Grouped-Query Attention (GQA) gibi hesaplama verimliliğini artıran yöntemler tercih edilmiştir. Bu yaklaşım, geleneksel multi-head attention mekanizmasına kıyasla daha düşük hesaplama maliyetiyle benzer doğruluk sağlar. Modelde ayrıca RMSNorm normalizasyonu, GeGLU aktivasyon fonksiyonu ve logit değerlerine soft-capping uygulanarak üretim kararlılığı artırılmıştır.

Gemma 2 modelleri, yaklaşık 2 trilyon token içeren büyük ölçekli İngilizce veri kümesi üzerinde ön eğitimden geçirilmiş; klasik next-token prediction yerine, daha büyük öğretici modellerin çıktılarından faydalanarak bilgi damıtma (knowledge distillation) uygulanmıştır. Bu sayede bağlamsal tutarlılık ve semantik yoğunluk artırılmış; sıfırdan eğitilen benzer boyuttaki modellere kıyasla %10'a kadar daha iyi performans gösterdiği raporlanmıştır [55].

Tez kapsamında kullanılan Turkish-Gemma-9b-v0.1 modeli, yukarıda açıklanan Gemma 2 mimarisi üzerine inşa edilmiş olup, Yıldız Teknik Üniversitesi Cosmos LLM araştırma grubu tarafından geliştirilmiştir [56]. Modelin geliştirilme süreci; continual pre-training, supervised fine-tuning (SFT), direct preference optimization (DPO) ve model merging adımlarını içermektedir. Model, özellikle Türkçe metin üretimi görevlerine uyumlu hale getirilmiştir. Eğitim sürecinde geniş ölçekli Türkçe metinler, görev odaklı prompt-response verileri ve insan tercihlerine dayalı örnekler kullanılmıştır. Böylece model, bildirim gibi kısa, yaratıcı ve bağlama duyarlı metin üretiminde başarılı sonuçlar vermektedir.

Modelin başarımı, 1.450 sorudan oluşan çeşitli kategorilere ait bir değerlendirme kümesinde test edilmiş; 18 insan değerlendirici tarafından yapılan karşılaştırmalı incelemelerde %57,30 oranında tercih edilmiştir. Bu oran, GPT-4o-mini ve Gemma 2-27B-IT gibi büyük modellerin hemen arkasında yer almakta olup, modelin pratik değerini ortaya koymaktadır. Bu tez çalışmasında model yalnızca bildirim üretimi amacıyla kullanılmıştır. Üretim süreci aşağıdaki yapılandırmalarla gerçekleştirilmiştir:

- Deterministik Üretim: `temperature=0`, $\texttt{top}_k = 1$,

Modelin çıkarım süreci, Hugging Face üzerinde yayınlanan `GGUF` formatındaki sürüm [57] kullanılarak, yerel ortamda çalışan Ollama platformu üzerinden gerçekleştirilmiştir. Bu süreçte herhangi bir quantization uygulanmamış, model 16-bit floating point (16F) biçiminde çalıştırılmıştır. Bu tercih, çıktı kalitesini koruyarak inference sürecinin doğruluğunu artırmış ve tekrarlanabilirliğe katkı sağlamıştır.

3.7 Değerlendirici Dil Modeli: GPT-4o

Bu çalışmada, bildirim üretimi sürecinde elde edilen metinlerin değerlendirilmesinde, üretim için kullanılan modelden farklı bir büyük dil modeli kullanılması tercih edilmiştir. Bu amaçla, OpenAI tarafından geliştirilen çok modlu yapay zekâ modeli olan GPT-4o (`gpt-4o-2024-08-06`) değerlendirme modeli olarak kullanılmıştır. GPT-4o, metin, ses, görüntü ve video gibi farklı modaliteleri tek bir sinir ağı üzerinden işleyebilen omni model sınıfına ait, otoregresif bir mimariye sahip dil modelidir [58].

GPT-4o, OpenAI'nin geliştirdiği önceki modellerin aksine, tüm modaliteleri işlemek üzere uçtan uca eğitilmiş tek bir ağdan oluşur. Bu yapı sayesinde multimodal görevlerde çapraz-modal bağlamları daha başarılı biçimde analiz

edebilmektedir. Model, eğitim sürecinde 2023 Ekim ayına kadar olan geniş kapsamlı bir veri kümesiyle eğitilmiştir. Bu veri kümesi, web belgeleri, kod verileri, yapılandırılmış mantık içeriği, matematiksel problemler, çeşitli lisanslı metinler ve görsel-işitsel içeriklerden oluşmaktadır [58]. Bu sayede model, yalnızca yüzeysel dil işleme değil, aynı zamanda mantıksal akıl yürütme, içerik kıyaslama, kod ve matematik temelli analizler gibi yüksek seviyeli dilsel görevlerde de güçlü performans göstermektedir.

Modelin teknik açıdan dikkat çekici özelliklerinden biri, konuşma girdilerine verdiği yanıt süresinin insan seviyesine yakın olmasıdır. GPT-4o, ses girişlerine ortalama 320ms gibi düşük bir gecikme ile cevap verebilmekte, minimumda ise 232ms'lik tepkime süresi sunabilmektedir [58]. Bu düşük gecikme, metin dışı modalitelerde avantaj sağlasa da bu çalışmada yalnızca metin değerlendirme kapasitesinden yararlanılmıştır.

Bu tez kapsamında GPT-4o, üretilen bildirimlerin dil bilgisel doğruluğu, doğal akıcılığı, bağlamsal tutarlılığı ve kullanıcı etkileşimine yatkınlığı gibi ölçütlerle değerlendirilmesinde kullanılmıştır. Değerlendirme sırasında model çağrıları `temperature = 0.0` ve `max_tokens = 4000` parametreleri ile gerçekleştirilmiştir. `temperature = 0.0` kullanımı, modelin aynı girdilere tutarlı yanıtlar üretmesini sağlayarak değerlendirme sürecinin tekrarlanabilir ve nesnel olmasına olanak tanımıştır. Öte yandan, yüksek `max_tokens` değeri, yalnızca kısa skorlar değil, açıklamalı ve gerekçeli değerlendirme çıktıları alınmasına imkân tanımıştır.

GPT-4o, bu çalışmada "LLM-as-a-judge" paradigması çerçevesinde değerlendirme amaçlı kullanılmıştır. Modelin farklı metin varyasyonlarını semantik olarak kıyaslama, dilsel tutarlılık açısından analiz etme ve gerektiğinde metin düzeltme gibi görevlerde yüksek başarı sağladığı gözlemlenmiştir. Örneğin:

- Dil bilgisel doğruluk değerlendirmesinde, modelden bildirimi düzeltmesi istenmiş, ardından orijinal ve düzeltilmiş metinler arasında Levenshtein mesafesi hesaplanarak sayısal skor üretilmiştir.
- Başlık ve içerik uyumu, doğallık ve emoji uygunluğu gibi daha öznel nitelikler ise GPT-4o'nun bağlamsal analiz kabiliyetine dayalı olarak yapılmıştır.

Sonuç olarak, GPT-4o, bu tezde yalnızca metin üretim yöntemlerinin değerlendirilmesinde kullanılmakla kalmamış, aynı zamanda dilsel metriklerin

hibrit biçimde geliştirilmesine de katkı sunmuştur. Tek sinir ağı üzerinden multimodal yeteneklere sahip olması ve yüksek bağlamsal hassasiyeti sayesinde, ölçeklenebilir, tutarlı ve güvenilir bir otomatik değerlendirici olarak işlev görmüştür.

4
DENEYLER VE SONUÇLAR

Bu bölümde, çalışmada önerilen üç yöntemin performansını karşılaştırmak amacıyla gerçekleştirilen deneylerin tasarımı ve elde edilen sonuçlar detaylı olarak sunulmaktadır. Öncelikle deneylerin nasıl yapılandırıldığı açıklanmakta, ardından her bir yöntemin çeşitli metrikler açısından performansı değerlendirilmektedir.

4.1 Deney Tasarımı

Bu çalışmada, bildirim üretimi için üç farklı yöntem — Elle, Otomatik ve Optimize — performans açısından karşılaştırılmıştır. 18 farklı kampanya ele alınmış ve bu kampanyaların her biri için 5'er adet özelleştirilmiş varyasyon üretilmiştir. Her bir varyasyon için söz konusu üç yöntem uygulanmıştır.

Üretilen bildirimler, Yöntem bölümünde detaylandırılan otomatik değerlendirme metrikleri ile analiz edilmiştir. Yöntemlerin performans karşılaştırmaları, Tablo 4.1 ve Tablo 4.2'de sunulmaktadır.

Deneylerde kullanılan veri kümesi %60 eğitim, %20 doğrulama ve %20 test olmak üzere üçe bölünmüştür. Optimize Yöntem kapsamında, istem optimizasyonu için eğitim verisi kullanılırken; doğrulama verisi optimizasyon sürecini izlemek amacıyla, test verisi ise nihai performans karşılaştırmalarında kullanılmıştır.

4.2 Sonuçlar

Deneyler sonucunda, üç yöntem arasında belirgin ve modele bağlı farklılıklar tespit edilmiştir. Aşağıda yorumlar, yalnızca LLM-odaklı metrikler (Benefit, Grammar, SoundsNatural, Emoji ailesi vb.) dikkate alınarak yapılmıştır; klasik benzerlik metrikleri (BLEU, ROUGE, BERTScore) ayrı ele alınmış ve metin sonunda kısaca özetlenmiştir.

Tablo 4.1 Gemma Türkçe Sonuçları — Yöntem bazında metrik karşılaştırması. En iyi değerler kalın gösterilmiştir.

Görev Odaklı LLM Metrikleri			
Metrik	**Elle Yöntem**	**Otomatik Yöntem**	**Optimize Yöntem**
BenefitMetrik	**4,95**	4,80	4,75
EmojiCountMetrik	**0,50**	0,05	0,10
EmojiSemanticMetrik	0,288	0,319	**0,394**
EmojisMetrik	4,45	4,85	**4,95**
GrammarMetrik	4,45	3,95	**4,55**
TitleGrammarMetrik	4,70	4,60	**4,75**
SoundsNaturalMetrik	**4,90**	4,85	**4,90**
TitleSoundsNaturalMetrik	**5,00**	4,20	4,95
Klasik metrikler			
Metrik	**Elle Yöntem**	**Otomatik Yöntem**	**Optimize Yöntem**
BLEU	**0,011**	0,000	0,000
METEOR	**0,191**	0,136	0,118
RougeL	**0,220**	0,138	0,157
RougeN	**0,262**	0,164	0,183
BERTScore	**0,587**	0,563	0,559

Tablo 4.2 Gemini 1.5 Pro Sonuçları — Yöntem bazında metrik karşılaştırması. En iyi değerler kalın gösterilmiştir.

Görev Odaklı LLM Metrikleri			
Metrik	**Elle Yöntem**	**Otomatik Yöntem**	**Optimize Yöntem**
BenefitMetrik	**5,00**	4,55	4,85
EmojiCountMetrik	0,50	0,05	**0,80**
EmojiSemanticMetrik	0,233	**0,297**	0,291
EmojisMetrik	3,88	4,60	**4,90**
GrammarMetrik	**4,80**	4,75	3,90
TitleGrammarMetrik	**4,75**	4,35	4,35
SoundsNaturalMetrik	**5,00**	**5,00**	4,95
TitleSoundsNaturalMetrik	**5,00**	4,95	**5,00**
Klasik metrikler			
Metrik	**Elle Yöntem**	**Otomatik Yöntem**	**Optimize Yöntem**
BLEU	0,013	**0,023**	0,015
METEOR	0,140	**0,217**	0,157
RougeL	0,190	**0,233**	0,186
RougeN	0,245	**0,272**	0,244
BERTScore	0,576	**0,611**	0,555

Optimize Yöntem, İngilizce temelli DSpy instruction optimizasyonu sayesinde Elle Yöntem'e yaklaşmış, hatta Gemma 2 Türkçe modelinde GrammarMetrik ve TitleGrammarMetrik gibi dil-bilgisi odaklı ölçütlerde birinciliğe yerleşmiştir. Buna karşın aynı yöntem Gemini 1.5 Pro tarafında dil-bilgisi skorlarında gerilemiş ve liderliği Elle Yöntem'e bırakmıştır. Bu asimetri, optimizasyon talimatlarının hedef dil ile tam uyumlu olmamasının performansı iki yönde de etkileyebileceğini göstermektedir. Nitekim (i) Gemma 2 Türkçe'nin ince-ayar verisinde emoji yoğunluğunun görece düşük olabileceği, dolayısıyla modelin a priori olasılık dağılımında "emoji ekle" sinyalini zayıflatıyor olabileceği; (ii) DSpy'nin İngilizce kalıplarla yaptığı istem optimizasyonunun Türkçe bağlamda aynı kuvvetle yansımayabileceği düşünülmektedir. Ayrıca Optimize Yöntem'in "riskli" ya da alıcının dikkatini olumsuz çekebilecek ifadelerden kaçınma eğilimi, metin çeşitliliğini daraltmış ve özellikle kelime kökü zenginliğini belirgin biçimde azaltmıştır. Bununla birlikte, Gemini 1.5 Pro sonuçlarında EmojiCountMetrik değerinin 0,80'e yükselmesi, hedeflenen 1,00 değerine en çok yaklaşan sonuçtur ve talimatların "daha fazla emoji, daha çekici içerik" varsayımını kısmen başarılı uyguladığını göstermektedir.

Elle Yöntem ile üretilen metinler, her iki modelde de en geniş kelime ve (tekrarsız) emoji çeşitliliğini sağlamış; özellikle Gemma 2 Türkçe üzerinde EmojiCountMetrik'te (0,50) en yüksek başarıyı göstermiştir. Few-shot örneklerin Türkçe içeriği, modelin kampanya dilini ve bildirim formatını sezgisel olarak yakalamasına yardımcı olmuş; SoundsNaturalMetrik ve TitleSoundsNaturalMetrik skorlarında yüksek—ancak birbirine çok yakın—puanlar elde edilmiştir. BenefitMetrik'te gözlenen üstünlük, rastgele seçilen örnekler üzerinde yapılan insan doğrulamasıyla da tutarlı çıkmış, LLM-as-a-judge yaklaşımının bu metrik özelinde hedef "faydayı" doğru yansıtabildiğini göstermiştir. Bu durum, otomatik değerlendirme sistemlerinin belirli metrikler için güvenilir olabileceğini, ancak yine de periyodik insan denetiminin metodolojik sağlamlık açısından önemini koruduğunu vurgular.

Otomatik Yöntem, optimizasyon aşamasına girmediği için çoğu LLM metrikte geride kalmıştır. Bununla birlikte, Gemini deneylerinde EmojiSemanticMetrik puanının görece yüksek çıkması dikkat çekicidir. Rastgele seçilen bildirim örnekleri incelendiğinde, eklenen emojilerin içerikle tematik olarak uyumlu olduğu görülmüştür; bu da yüksek semantik skoru açıklamaktadır. Ancak aynı örneklerde başlık + gövde toplamında genellikle üçten fazla emoji kullanıldığı için EmojiCountMetrik değeri Tablo 4.2'deki gibi çok düşüktür (0,05). Kısacası semantik skor artışı kısmen "bol emoji" stratejisinden beslenmekte, fakat metrikler arasındaki tanım farkı yüzünden kullanıcıya önerilen ideal emoji sınırını

karşılamamaktadır.

SoundsNatural ve TitleSoundsNatural metrikleri, her iki model ve yöntem kombinasyonunda birbirine çok yakın ($\leq$ 0,05 puan) seyretmiştir. Ancak bu metriklerin mutlak doğruluğu, insan değerlendirmeleriyle düzenli olarak kalibre edilmediği sürece sınırlı kalabilir; dolayısıyla doğallık puanlarının tek başına yeterli kabul edilmemesi önerilmektedir.

Ortaya çıkan genel tablo, metriklerin sayısı ve ölçeği arttıkça çok-amaçlı optimizasyon dengesinin hassaslaştığını göstermektedir. Özellikle 0-1 aralıklı EmojiCountMetrik gibi dar ölçekli göstergeler, geniş bantlı göstergeler (0-5'lik Benefit veya SoundsNatural) yanında kolayca ağırlık kaybedebilmekte; bu da tekil bir metrikte hedef davranışı (örneğin emoji kullanımı) zorlaştırmaktadır. Dolayısıyla üretim ortamına geçişte metrik ağırlıklarının normalizasyonu ve gerektiğinde hard constraint tanımlarıyla desteklenmesi önerilmektedir. Ayrıca GrammarMetrik gibi hibrit çözümler (kural tabanlı + LLM) dil-bilgisi denetiminde, saf LLM-as-a-judge yaklaşımına göre daha tutarlı sonuç vermiştir.

Klasik benzerlik metrikleri açısından bakıldığında, Gemma 2 tarafında Elle Yöntem açık ara en yüksek skorları alırken, Gemini deneylerinde Otomatik Yöntem'in öne çıkması dikkat çekicidir. Bu asimetrik durum, yüzeysel metin benzerliğinin fayda veya doğallık gibi gerçek kullanıcı deneyimi göstergeleriyle her zaman korele olmadığını ve benzerlik skorlarının karar verici tek ölçüt olarak kullanılmaması gerektiğini belirtmektedir.

Sonuç olarak, yalnızca tek bir model ya da tek bir metrik üzerinden yapılan değerlendirmeler, LLM tabanlı bildirim üretiminde yöntemlerin gerçek potansiyelini yansıtmakta yetersiz kalabilir. Bu nedenle, dilsel dağılımı farklı modeller arası çapraz doğrulama, çok-ölçekli metrik setleri ve insan denetimli kalibrasyon katmanlarının birlikte ele alınması kritik görünmektedir. Ek olarak, Tablo 4.3, Tablo 4.4 ve Tablo 4.5, farklı kampanya senaryoları için (sırasıyla *B_ALL_CHURN_400_500*, *X_HML_ACQ_400_475*, ve *ALL_PD_MILKA_1E1_1510*) üretilen bildirimleri her iki dil modeli ve üç yöntem özelinde yan yana sunarak; üslup, emoji kullanımı, mesaj yapısı ve akıcılık gibi nitel farkları somutlaştırmakta ve metrik bulgularının anlamlılığını görselleştirmektedir.

Tablo 4.3 `B_ALL_CHURN_400_500` kampanyası için yöntem ve LLM bazında bildirim örnekleri

LLM	Yöntem	Başlık	Metin
Gemini	Elle	Rahatına bak, biz hallederiz	X'ten alışverişin rahatlığını hatırlıyor musun? Seni tekrar aramızda görmek için ilk siparişine 100 TL indirim seni bekliyor.
Gemini	Otomatik	Seni Özledik! 👋🏻😔🎁	Sana özel 100 TL indirimle geri dön! Hızlı ve kolay alışveriş deneyimi seni bekliyor. ✨
Gemini	Optimize	Bizi özledin mi? 😔	Hızlı ve kolay alışveriş deneyimiyle zaman kazanmak için geri dön! Sana özel 100 TL indirim seni bekliyor. ✨
Gemma	Elle	X'le alışveriş keyfi geri döndü! 🛒	100 TL indirim seni bekliyor. Evde kal, X ile ihtiyacın olan her şeyi kapına getirelim! 😊
Gemma	Otomatik	Geri döndüğüne sevindik!	Seni özledik! 🛍️🚀 Geri dön ve 100 TL indirim kazan! Hızlı teslimat ve kolay alışveriş seni bekliyor. 😊
Gemma	Optimize	Hemen Alışverişe Başla	Seni özledik! 🎁 100 TL indirimle geri dön ve alışveriş keyfini yeniden yaşa!

Tablo 4.4 `X_HML_ACQ_400_475` kampanyası için yöntem ve LLM bazında bildirim örnekleri

LLM	Yöntem	Başlık	Metin
Gemini	Elle	Seni özledik! 👋🏻	X'e geri dön! İlk siparişine özel 75 TL indirim ve ücretsiz teslimat seni bekliyor.
Gemini	Otomatik	👋🏻 75 TL İndirim Bekliyor! 🎉	İlk siparişine özel bedava teslimat ve 75 TL indirim fırsatını kaçırma! Hoş geldin avantajlarından yararlanmak için hemen sipariş ver! ✨
Gemini	Optimize	Seni Özledik! 👋🏻	Geri dönüşünü özel bir indirimle kutlayalım! 🎉 İlk siparişine özel 75 TL indirim ve ücretsiz teslimat seni bekliyor. Hemen uygulamayı aç ve avantajları yakala! ✨
Gemma	Elle	75 TL indirim seni bekliyor!	X'e geri dön, ilk siparişinde 75 TL indirim kazan! Hızlı ve pratik alışverişin tadını çıkar. 😊
Gemma	Otomatik	Hoş Geldin, Avantajları Yakala!	👋🏻 Hoş Geldin! İlk siparişinde 75 TL indirim seni bekliyor. Ücretsiz teslimat ile avantajları yakala! 🎉
Gemma	Optimize	👋🏻 Hoş Geldin! 75 TL İndirim Senin!	İlk siparişinde ücretsiz teslimat ve 75 TL indirim seni bekliyor! 🎉 Hemen keşfet.

Tablo 4.5 `ALL_PD_MILKA_1E1_1510` kampanyası için yöntem ve LLM bazında bildirim örnekleri

LLM	Yöntem	Başlık	Metin
Gemini	Elle	Yeniden Merhaba 👋🏻	X'e geri dön, özlediğin lezzetler seni bekliyor! Milka Süt Dolgulu Kek'te 2 al 1 öde fırsatını kaçırma 😉
Gemini	Otomatik	Seni Özledik! 👋🏻🧨😋	Milka Süt Dolgulu Kek'te (29 g) harika bir fırsat seni bekliyor: 2 al 1 öde! Geri dön ve kaçırma.
Gemini	Optimize	Bizi Özlediğinizi Biliyoruz! 😊	Milka Süt Dolgulu Kek'lerde (29 g) kaçırılmayacak fırsat! 2 al 1 öde! 🎉 Yeni özelliklerimizle seni tekrar aramızda görmek harika olurdu. 👋🏻
Gemma	Elle	Milka keyfi iki katına çıksın!	X'e geri dön, 2 Milka Süt Dolgulu Kek al 1 öde! 😊 Hemen sipariş ver.
Gemma	Otomatik	Hoş Geldin Teklifi!	👋🏻 Yeni özellikler seni bekliyor! 🧨 Milka Süt Dolgulu Kek (29 g) al, 1 öde! 😊 Hemen geri dön ve avantajlardan yararlan!
Gemma	Optimize	Hoş Geldin Teklifi	🧨 Milka Kek 2 al 1 öde! 🎉 Seni özledik, yeni özellikler seni bekliyor!

5
SONUÇ

Bu çalışmada, Büyük Dil Modelleri (LLM) kullanarak bildirim metinlerini üretmek üzere Elle Yöntem, Otomatik Yöntem ve Optimize Yöntem yaklaşımları karşılaştırılmıştır. Hem Turkish-Gemma-9b-v0.1 hem de Gemini 1.5 Pro 001 modelleri üzerindeki ölçümler, Görev Odaklı LLM metrikleri (Benefit, Grammar, SoundsNatural, Emoji ailesi vb.) ile klasik benzerlik metriklerini (BLEU, ROUGE, BERTScore) birlikte değerlendirmiştir.

Elle Yöntem uzman emeğine dayansa da Benefit, SoundsNatural ve EmojiCount gibi kullanıcı faydası göstergelerinde ve dil-bilgisi/doğallık skorlarında en yüksek değerleri elde etmiş; ayrıca en geniş kelime ve emoji çeşitliliğini koruyarak yaratıcılıkta üstünlüğünü göstermiştir.

Optimize Yöntem, DSpy tabanlı istem iyileştirmesi sayesinde Turkish-Gemma-9b-v0.1 modelinde GrammarMetrik ve TitleGrammarMetrik'te liderliğe yükselmiştir. Bununla birlikte, aynı İngilizce odaklı talimat kalıplarının Gemini 1.5 Pro'da dil-bilgisi skorlarını geriye çekmesi, optimizasyon stratejilerinin hedef dile özgü biçimde tasarlanmasının kritik olduğunu ortaya koymaktadır.

Otomatik Yöntem, LLM metriklerinde görece geride kalmasına karşın Gemini 1.5 Pro üzerinde klasik benzerlik ölçütlerinde en iyi skorları almış ve EmojiSemanticMetrik'te kayda değer başarı göstermiştir. Ancak bunun sebebi olan aşırı emoji kullanımı, EmojiCount puanlarını sınırlı tutmuş ve metrik tanımlarının hassaslığını öne çıkarmıştır.

Sonuçlar, dar ölçekli $(0-1)$ göstergeler ile geniş bantlı $(0-5)$ göstergelerin birlikte optimize edilmesinde metrik ağırlıklarının dengeli tutulmaması hâlinde hedef davranışların gölgelenebileceğini göstermektedir. LLM tarafından düzeltilmiş metin ile özgün metin arasındaki uzaklığa dayanan GrammarMetrik'in, saf LLM-as-a-judge yaklaşımına kıyasla daha kararlı sonuçlar vermesi, hibrit denetim yaklaşımlarının dil-bilgisi doğruluğu açısından avantaj sağladığını göstermiştir. Öte

yandan, klasik benzerlik metrikleri yalnızca önceki insan yazımı bildirimlere olan yüzeysel benzerliği yakaladığından kullanıcı faydasını doğrudan yansıtmaz; bu nedenle bu metriklere aşırı güvenmek yerine çok-ölçekli bir metrik seti benimsemek ve sonuçları düzenli insan kalibrasyonuyla desteklemek kritik görünmektedir.

Çalışma, sınırlı kampanya varyasyonları ve iki model üzerinde yürütülmüştür; bulguların genellenebilirliğini artırmak için farklı dil modelleri, sektörler ve metin türleriyle ek testler yapılmalıdır. Gelecek çalışmalarda şunlar yapılabilir:

- Bildirim performansı, gerçek kullanıcı etkileşimleriyle (çevrim-içi A/B testleri) doğrulanabilir.
- Çeşitlilik ve yaratıcılığı ölçen ek metrikler geliştirilebilir.
- Düşünce Zinciri (Chain of Thought) gibi istem mühendisliği teknikleri kullanılarak çok adımlı düşünme süreçlerinin katkısı araştırılabilir.
- LLM metrik çıktılarının insan yargılarıyla korelasyonu, ölçeklenebilir anket ve crowd-sourcing düzenekleriyle ölçülebilir.
- Açık kaynaklı modeller için ayrıntılı maliyet/performans analizleri yapılarak küçük ayak iziyle yüksek kalite sağlayan alternatifler belirlenebilir.

Sonuç olarak, bu çalışma LLM'lerin yalnızca içerik üretiminde değil, aynı zamanda görev odaklı metriklere göre istem üretiminde de etkin şekilde kullanılabileceğini göstermektedir. Bu da üretken yapay zekâ uygulamalarında daha kontrollü, kullanıcı faydası odaklı ve sürdürülebilir sistemlerin geliştirilmesine olanak tanımaktadır.

KAYNAKÇA

[1] K. Fraser O. Conlan, "Enticing notification text & the impact on engagement," *Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers*, 2020, ss. 444–449.

[2] F. Muench A. Baumel, "More than a text message: dismantling digital triggers to curate behavior change in patient-centered health interventions," *Journal of medical Internet research*, c. 19, no. 5, e147, 2017.

[3] N. Fatima, A. S. Imran, Z. Kastrati, S. M. Daudpota A. Soomro, "A systematic literature review on text generation using deep neural network models," *IEEE Access*, c. 10, ss. 53 490–53 503, 2022.

[4] W. X. Zhao ve diğ., "A survey of large language models," *arXiv preprint arXiv:2303.18223*, 2023.

[5] J. Li, T. Tang, W. X. Zhao, J.-Y. Nie J.-R. Wen, "Pre-trained language models for text generation: A survey," *ACM Computing Surveys*, c. 56, no. 9, ss. 1–39, 2024.

[6] A. Celikyilmaz, E. Clark J. Gao, "Evaluation of text generation: A survey," *arXiv preprint arXiv:2006.14799*, 2020.

[7] E. Montahaei, D. Alihosseini M. S. Baghshah, "Jointly measuring diversity and quality in text generation models," *arXiv preprint arXiv:1904.03971*, 2019.

[8] A. Visuri, N. van Berkel, T. Okoshi, J. Goncalves V. Kostakos, "Understanding smartphone notifications' user interactions and content importance," *International Journal of Human-Computer Studies*, c. 128, ss. 72–85, 2019.

[9] P. Saikia, M. Cheung, J. She S. Park, "Effectiveness of mobile notification delivery," *2017 18th IEEE international conference on mobile data management (MDM)*, IEEE, 2017, ss. 21–29.

[10] H. Zhao, L. Si, X. Li Q. Zhang, "Recommending complementary products in e-commerce push notifications with a mixture model approach," *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2017, ss. 909–912.

[11] L. G. Morrison ve diğ., "The effect of timing and frequency of push notifications on usage of a smartphone-based stress management intervention: an exploratory trial," *PloS one*, c. 12, no. 1, e0169162, 2017.

[12] M. Gençer, G. Bilgin, Ö. Zan T. Voyvodaoğlu, "A new framework for increasing user engagement in mobile applications using machine learning techniques," *Design, User Experience, and Usability. Web, Mobile, and Product Design: Second International Conference, DUXU 2013, Held as Part of HCI International 2013, Las Vegas, NV, USA, July 21-26, 2013, Proceedings, Part IV 2*, Springer, 2013, ss. 651–659.

[13] D. Y. Wicaksana, R. J. S. Z. Aisyah D. U. Q. A'yun, "Designing Effective Mobile Push Notifications: Machine Learning Insights into User Engagement," *Journal of Digital Business and Innovation Management*, c. 3, no. 2, 2024.

[14] F. Corno, L. De Russis T. Montanaro, "A context and user aware smart notification system," *2015 IEEE 2nd World Forum on Internet of Things (WF-IoT)*, IEEE, 2015, ss. 645–651.

[15] C. O'Brien, H. Wu, S. Zhai, D. Guo, W. Shi J. J. Hunt, "Should I send this notification? Optimizing push notifications decision making by modeling the future," *arXiv preprint arXiv:2202.08812*, 2022.

[16] G. NEBBIAI, "Exploring Customer Behavior Through Push Notifications: A Comparative Analysis of Machine Learning and Deep Learning Approaches,"

[17] S. Minaee ve diğ., "Large language models: A survey," *arXiv preprint arXiv:2402.06196*, 2024.

[18] M. Reisenbichler, T. Reutterer, D. Schweidel D. Dan, "Supporting content marketing with natural language generation," *Marketing Science Institute (February 21), https://www. msi. org/working-papers/supporting-content-marketing-withnatural-language-generation-2*, 2021.

[19] B. Kilichev, N. Abbasova, N. Gafurova, S. Dadabayeva, M. Nurmatova S. Fazildinova, "A Development of Personalized Content Creation Technology Model using NLP and AI Integrated System," *2024 4th International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, IEEE, 2024, ss. 595–600.

[20] F. Zubair, A. Jafir ve diğ., "Natural Language Generation: Recent Advances and Applications," *International Journal of Advanced Engineering Technologies and Innovations*, c. 1, no. 1, ss. 50–74, 2021.

[21] H. K. Skrodelis, A. Romanovs, N. Zenina H. Gorskis, "The latest in natural language generation: Trends, tools and applications in industry," *2023 IEEE 10th Jubilee Workshop on Advances in Information, Electronic and Electrical Engineering (AIEEE)*, IEEE, 2023, ss. 1–5.

[22] S. Schulhoff ve diğ., "The Prompt Report: A Systematic Survey of Prompting Techniques," *arXiv preprint arXiv:2406.06608*, 2024.

[23] X. Amatriain, "Prompt design and engineering: Introduction and advanced methods," *arXiv preprint arXiv:2401.14423*, 2024.

[24] Q. Dong ve diğ., "A survey on in-context learning," *arXiv preprint arXiv:2301.00234*, 2022.

[25] T. Brown ve diğ., "Language models are few-shot learners," *Advances in neural information processing systems*, c. 33, ss. 1877–1901, 2020.

[26] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo Y. Iwasawa, "Large language models are zero-shot reasoners," *Advances in neural information processing systems*, c. 35, ss. 22 199–22 213, 2022.

[27] X. Wang, W. Zhu, M. Saxon, M. Steyvers W. Y. Wang, "Large language models are latent variable models: Explaining and finding good demonstrations for in-context learning," *Advances in Neural Information Processing Systems*, c. 36, 2024.

[28] S. Min ve diğ., "Rethinking the role of demonstrations: What makes in-context learning work?" *arXiv preprint arXiv:2202.12837*, 2022.

[29] S. Min, M. Lewis, L. Zettlemoyer H. Hajishirzi, "Metaicl: Learning to learn in context," *arXiv preprint arXiv:2110.15943*, 2021.

[30] Y. Zhou, Y. Zhao, I. Shumailov, R. Mullins Y. Gal, "Revisiting Automated Prompting: Are We Actually Doing Better?" *arXiv preprint arXiv:2304.03609*, 2023.

[31] J. Yang, S. Ma F. Wei, "Auto-ICL: In-Context Learning without Human Supervision," *arXiv preprint arXiv:2311.09263*, 2023.

[32] C. Yang ve diğ., "Large language models as optimizers," *arXiv preprint arXiv:2309.03409*, 2023.

[33] T. Shin, Y. Razeghi, R. L. Logan IV, E. Wallace S. Singh, "Autoprompt: Eliciting knowledge from language models with automatically generated prompts," *arXiv preprint arXiv:2010.15980*, 2020.

[34] Y. Zhou ve diğ., "Large language models are human-level prompt engineers," *arXiv preprint arXiv:2211.01910*, 2022.

[35] M. Deng ve diğ., "Rlprompt: Optimizing discrete text prompts with reinforcement learning," *arXiv preprint arXiv:2205.12548*, 2022.

[36] K. Opsahl-Ong ve diğ., "Optimizing Instructions and Demonstrations for Multi-Stage Language Model Programs," *arXiv preprint arXiv:2406.11695*, 2024.

[37] L. Yin Z. Wang, "Auto-Differentiating Any LLM Workflow: A Farewell to Manual Prompting," *arXiv preprint arXiv:2501.16673*, 2025.

[38] J. Shen, J. J. Dudley, J. Zheng, B. Byrne P. O. Kristensson, "Promptor: A conversational and autonomous prompt generation agent for intelligent text entry techniques," *arXiv preprint arXiv:2310.08101*, 2023.

[39] C. Li, M. Zhang, Q. Mei, W. Kong M. Bendersky, "Automatic Prompt Rewriting for Personalized Text Generation," *arXiv preprint arXiv:2310.00152*, 2023.

[40] C. Van Der Lee, A. Gatt, E. Van Miltenburg, S. Wubben E. Krahmer, "Best practices for the human evaluation of automatically generated text," *Proceedings of the 12th International Conference on Natural Language Generation*, 2019, ss. 355–368.

[41] K. Papineni, S. Roukos, T. Ward W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, ss. 311–318.

[42] E. Sulem, O. Abend A. Rappoport, "BLEU is not suitable for the evaluation of text simplification," *arXiv preprint arXiv:1810.05995*, 2018.

[43] Y. Zhang, S. Vogel A. Waibel, "Interpreting bleu/nist scores: How much improvement do we need to have a better system?" *LREC*, 2004.

[44] C. Callison-Burch, M. Osborne P. Koehn, "Re-evaluating the role of BLEU in machine translation research," *11th conference of the european chapter of the association for computational linguistics*, 2006, ss. 249–256.

[45] S. Gehrmann ve diğ., "The gem benchmark: Natural language generation, its evaluation and metrics," *arXiv preprint arXiv:2102.01672*, 2021.

[46] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," *Text summarization branches out*, 2004, ss. 74–81.

[47] S. Banerjee A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005, ss. 65–72.

[48] T. Kocmi C. Federmann, "Large language models are state-of-the-art evaluators of translation quality," *arXiv preprint arXiv:2302.14520*, 2023.

[49] J. Fu, S.-K. Ng, Z. Jiang P. Liu, "Gptscore: Evaluate as you desire," *arXiv preprint arXiv:2302.04166*, 2023.

[50] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu C. Zhu, "G-eval: Nlg evaluation using gpt-4 with better human alignment," *arXiv preprint arXiv:2303.16634*, 2023.

[51] B. Murugadoss ve diğ., "Evaluating the Evaluator: Measuring LLMs' Adherence to Task Evaluation Instructions," *arXiv preprint arXiv:2408.08781*, 2024.

[52] A. Panickssery, S. R. Bowman S. Feng, "Llm evaluators recognize and favor their own generations," *arXiv preprint arXiv:2404.13076*, 2024.

[53] O. Khattab ve diğ., "Dspy: Compiling declarative language model calls into self-improving pipelines," *arXiv preprint arXiv:2310.03714*, 2023.

[54] G. Team ve diğ., "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context," *arXiv preprint arXiv:2403.05530*, 2024.

[55] G. Team ve diğ., "Gemma 2: Improving open language models at a practical size," *arXiv preprint arXiv:2408.00118*, 2024.

[56] Y. T. U. C. E. D. C. R. Group, *Turkish-Gemma-9b-v0.1*, `https://huggingface.co/ytu-ce-cosmos/Turkish-Gemma-9b-v0.1`, Erişim tarihi: 7 Haziran 2025, 2025.

[57] M. Radermacher, *Turkish-Gemma-9b-v0.1-GGUF*, Erişim tarihi: 7 Haziran 2025, 2025. erişim adresi: `https://huggingface.co/mradermacher/Turkish-Gemma-9b-v0.1-GGUF`.

[58] A. Hurst ve diğ., "Gpt-4o system card," *arXiv preprint arXiv:2410.21276*, 2024.

A VURGU TANIMLARI

Churn

Emphasize the benefits of returning, personalized offers, and incentives to persuade them to come back. Highlight any new features or improvements that might interest them., e.g., We Miss You!, New Features Await!, Welcome Back Offer!

Convenience Seekers

Emphasize fast and convenient service for a hassle-free, time-saving shopping experience. Highlight quick delivery and efficiency. e.g., Seamless Convenience, Delivered Fast!, Effortless Shopping, Instant Gratification!

Generic

Generate enthusiasm and highlight unique offerings and exciting deals. e.g. Discover Exciting Offers!, Don"t Miss Out! Explore new and exciting products today., Hot Deals Inside! Find amazing discounts and exclusive offers just for you., Join the Fun! Enjoy top picks and trending items now.

Loyal and Weekly Shoppers

Focus on a warm, friendly tone that makes them feel appreciated and valued. Reinforce their positive shopping behavior and create a sense of connection and community. e.g., Together, We Make Your Week Easier and Brighter!, Your Weekly Shopping, Made Special Just for You!

Promo-Driven Cluster

Highlight significant savings, exclusive offers, and limited-time discounts. The focus should be on conveying the value of the promotion and the urgency to take advantage of it. e.g., Big Savings Alert!, Enjoy 30% Off Today!, Exclusive Offer!, Limited Time Discount Available Now!

B
ÜSLUP TANIMLARI

Friendly

Be approachable, warm, and kind. Emphasize relationship-building. Help build trust and rapport with users, making them feel valued.

Humorous

Be playful, light-hearted, and often incorporate puns or jokes. Engage users in a fun way, making the brand memorable and relatable.

Appreciative

Express gratitude or show recognition towards the user's actions, loyalty, or engagement. Demonstrate that the company values its users, fostering a sense of loyalty and encouraging continued interaction with the service.

Exciting

Convey enthusiasm, celebration, or excitement about a product, offer, or news. Generate excitement in users and entice them to engage with the content or offer.

Reassuring

Alleviate concerns, doubts, or fears. Comfort and affirm. Maintain trust, especially during service interruptions, delivery delays, or significant changes.

Urgent

Convey a sense of immediacy and time-sensitivity. Encourage swift action, capitalize on impulsive buying, or highlight time-bound opportunities.

Direct

Be straightforward, without embellishment. Get to the point promptly. Ideal for users who prefer clear, no-nonsense communication, ensuring they receive essential information without distraction.

TEZDEN ÜRETİLMİŞ YAYINLAR

Konferans Bildirisi

1. H. Taşköprü and B. Diri, “Büyük Dil Modelleri ile Bildirim Metni Üretimi,” 2025 33rd Signal Processing and Communications Applications Conference (SIU), İstanbul, Türkiye, 25–28 Haziran 2025. (Kabul edildi, yayım aşamasında)