
TÜRKİYE VE DÜNYADA
YAZILIM MÜHENDİSLİĞİ
Editör: Dr.Öğr.Üyesi Özgür TONKAL
yaz
yayınları

# Türkiye ve Dünyada
# Yazılım Mühendisliği

**Editör**

Dr.Öğr.Üyesi Özgür TONKAL

yaz
yayınları

**2025**

yaz
yayınları

**Türkiye ve Dünyada Yazılım Mühendisliği**

Editör: Dr.Öğr.Üyesi Özgür TONKAL



E_ISBN 978-625-8678-50-5

Aralık 2025 – Afyonkarahisar

Dizgi/Mizanpaj: YAZ Yayınları

Kapak Tasarım: YAZ Yayınları

YAZ Yayınları. Yayıncı Sertifika No: 73086

M.İhtisas OSB Mah. 4A Cad. No:3/3
İscehisar/AFYONKARAHİSAR

www.yazyayinlari.com

yazyayinlari@gmail.com

# İÇİNDEKİLER

# BULANIK REGRESYON VE YAPAY ZEKÂ UYGULAMALARI

**Dinçer ATASOY**[1]

## 1. GİRİŞ

Bulanık mantık, klasik mantığın ikili $[0, 1]$ yapısına alternatif olarak geliştirilmiş olup belirsizliklerin modellenmesinde önemli bir rol oynamaktadır (Zadeh, 1965). Günümüz yapay zekâ sistemleri, insan düşünce sürecine benzer biçimde karar verebilmek için sıklıkla bulanık mantık tabanlı yaklaşımlardan yararlanmaktadır (Ross, 2010). Bu bağlamda, bulanık mantık yalnızca belirsizliğin temsili açısından değil, aynı zamanda belirsiz verilerle rasyonel çıkarımlar yapabilme kapasitesiyle de öne çıkmaktadır.

Belirsizlik (*uncertainty*) ve bulanıklık (*imprecision*) doğada sıkça karşılaşılan olgulardır. Klasik deterministik modeller, bu tür durumları her zaman doğru biçimde temsil edememektedir. Regresyon analizi, değişkenler arasındaki ilişkileri incelemek için kullanılan en temel ve yaygın istatistiksel yöntemlerden biridir (Atasoy, 2001). Klasik regresyon modelleri yalnızca kesin (crisp) veriler ve belirli istatistiksel varsayımlar altında geçerliliğini korur. Gerçek yaşamda elde edilen veriler ise çoğu zaman belirsiz, eksik ya da sözel ifadelerle tanımlandığından, klasik modellerin açıklayıcılığı ve doğruluğu sınırlı kalmaktadır (Zadeh, 1965).

Bu sınırlılığı aşmak üzere geliştirilen bulanık regresyon analizi (BRA), klasik regresyonun katı varsayımlarını gevşeterek

[1] Dr. Öğr. Üyesi, Iğdır Üniversitesi, Mühendislik Fakültesi, Yazılım Mühendisliği, ORCID: 000-0003-0389-1059.

belirsiz veya muğlak veriler üzerinde modelleme yapılmasına olanak tanır. Bu yaklaşım, bağımlı ve bağımsız değişkenlerin ya da model katsayılarının bulanık sayılar biçiminde ifade edilmesine dayalı modellerin oluşturulmasını amaçlar (Shapiro, 2005). Bulanık regresyon modelleri belirsizlik, eksik veri ve uzman yargısı gibi faktörleri doğrudan model yapısına entegre edebilmekte ve daha esnek bir temsil gücü sunmaktadır (Chukhrova et al., 2019).

Yapay zekâ alanında, özellikle bulanık mantık temelli yöntemlerin istatistiksel tekniklerle birleşimi, “soft computing” paradigması kapsamında önemli bir yer tutmaktadır (Ojha et al., 2019). Bu yaklaşımlar, klasik istatistiksel modellerin deterministik doğasına karşılık, belirsizliğin doğrudan modellenmesine ve karar süreçlerinin daha gerçekçi biçimde temsil edilmesine imkân vermektedir.

Bulanık regresyon modellerinin teorik temelleri, parametre tahmin yöntemleri, Python ortamında gerçekleştirilebilecek uygulama örnekleri ve yapay zekâ yaklaşımlarıyla entegrasyon süreçleri ele alınacaktır. Ayrıca, farklı uygulama alanlarından örneklerle bu modellerin avantajları ve sınırlılıkları tartışılacaktır.

## 2. KURAMSAL TEMELLER

Bulanık küme teorisi, klasik mantığın “bir eleman ya kümededir ya da değildir” şeklindeki ikili yapısını genişleterek, her bir elemanın bir kümeye belirli bir üyelik derecesi ile ait olabileceğini öngörmektedir. Bu yaklaşımda, bir elemanın kümeye ait olma derecesi üyelik fonksiyonu ($\mu(x)$) aracılığıyla [0,1] aralığında tanımlanır. Böylece, kesinlik kavramının yerini dereceli üyelik anlayışı almakta ve belirsizlik sistematik biçimde temsil edilebilmektedir (Mukaidono, 2001).

Bulanık sayılar genellikle üçgensel (Triangular Fuzzy Number – TFN) veya yamuk (Trapezoidal) biçimlerinde ifade edilir. Üçgensel bulanık sayı, $(a_l, a_m, a_r)$ parametreleriyle tanımlanır ve bu yapı, değerin merkez (en olası) noktası ile alt ve üst sınırlarını gösterir. Bu temsiller, bulanık regresyon modellerinde belirsiz katsayıların nicel olarak ifade edilmesine olanak sağlamaktadır.

Bulanık regresyon analizi (BRA), bu kuramsal temellerden hareketle regresyon katsayılarını bulanık sayılar biçiminde tanımlar (Pakdel et al., 2025). Genel model formu şu şekilde ifade edilir:

$$\tilde{Y} = \tilde{A}_0 + \tilde{A}_1 x_1 + \cdots + \tilde{A}_k x_k$$

Burada $\tilde{A}_i$ katsayıları bulanık sayıları, $\tilde{Y}$ ise bulanık bir çıktı değişkenini temsil etmektedir (Tanaka et al., 1982).

Bulanık regresyon kavramı ilk kez Tanaka, Uejima ve Asai (1982) tarafından ortaya konmuştur. Bu model, verilerin belirli bir h-düzeyinde (h-level) kapsanmasını sağlayan doğrusal programlama yaklaşımına dayanmaktadır. Daha sonraki çalışmalarda, Diamond (1988) klasik en küçük kareler yöntemini bulanıklaştırarak Fuzzy Least Squares Regression (FLSR) modelini geliştirmiştir. Buckley (2004) ise model parametrelerinin doğrudan bulanıklaştırıldığı alternatif bir yöntem önermiştir. Türkiye'de bu alanda yapılan çalışmalar arasında İçen (2010) ve Çetintav (2012)'nin çalışmaları dikkat çekicidir.

Son yıllarda, Stanojević (2023) tarafından geliştirilen optimizasyon temelli bulanık regresyon modeli**,** genişletilmiş uzantı prensibi ile uyumlu çözümler üretmekte ve modelin hesaplama verimliliğini artırmaktadır. Ayrıca, Júnior ve arkadaşları (2023) tarafından önerilen yaklaşımlar, derin öğrenme tabanlı bulanık sistemleri regresyon analizleriyle bütünleştirerek,

bulanık regresyonun modern yapay zekâ uygulamalarında yeniden önem kazandığını ortaya koymaktadır.

Klasik regresyon modeli genel olarak şu şekilde ifade edilir:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon$$

Buna karşılık, bulanık regresyon modeli şu biçimdedir:

$$\tilde{Y} = \tilde{A}_0 + \tilde{A}_1 x_1 + \cdots + \tilde{A}_k x_k$$

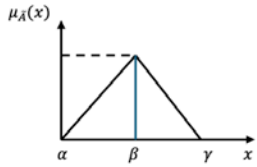
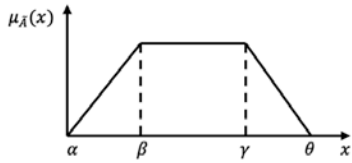
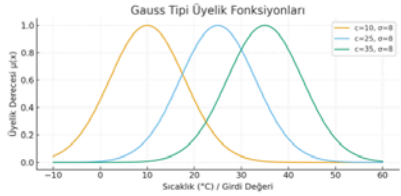
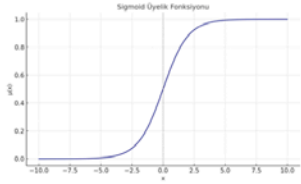
Burada her bir $\tilde{A}_i = (a_i, c_i)$ biçiminde tanımlanan üçgensel bulanık sayıdır. Modelin temel amacı, tüm gözlemleri kapsayacak biçimde toplam belirsizliği (yay genişliğini) en küçüklemektir (Tanaka et al., 1982). Bu sayede, model hem belirsiz hem de eksik veriler üzerinde anlamlı ve esnek tahminler üretebilme kapasitesine sahip olmaktadır.

## 3. BULANIK ÜYELİK FONKSİYONLARI

Bulanık küme teorisi, Zadeh (1965) tarafından ortaya konulan klasik küme anlayışının bir genişlemesidir. Bu teori, bir elemanın belirli bir kümeye aitliğini yalnızca “var” veya “yok” biçiminde değil, kısmi aidiyet dereceleriyle ifade eder. Bu bağlamda, üyelik fonksiyonu (membership function), evrensel küme üzerinde tanımlı olup her bir elemanın aitlik derecesini [0,1] aralığında gösteren matematiksel bir fonksiyondur.

$$\mu_{\tilde{A}}(x): X \rightarrow [0,1]$$

- $\mu_{\tilde{A}}(x) = 0$ ise x elemanı kümeye hiç ait değildir.
- $\mu_{\tilde{A}}(x) = 1$ ise x elemanı kümeye tamamen aittir.
- $0 < \mu_{\tilde{A}}(x) < 1$ ise x elemanı kümeye kısmen aittir.

| Tür | Fonksiyon | Grafik |
|---|---|---|
| Üçgen | $\mu_{\tilde{A}}(x) = \begin{cases} \frac{x-\alpha}{\beta-\alpha}, & \alpha \le x \le \beta \\ \frac{\gamma-x}{\gamma-\beta}, & \beta \le x \le \gamma \\ 0, & \text{diğer durumlar} \end{cases}$ |   |
| Yamuk | $\mu_{\tilde{A}}(x) = \begin{cases} 0, & x \le \alpha \\ \frac{x-\alpha}{\beta-\alpha}, & \alpha < x \le \beta \\ 1, & \beta < x \le \gamma \\ \frac{\theta-x}{\theta-\gamma}, & \gamma < x \le \theta \\ 0, & x > \theta \end{cases}$ |   |
| Gauss tipi | $\mu_{\tilde{A}}(x) = \exp\left(-\frac{(x-c)^2}{2\sigma^2}\right)$ |   |
| Sigmoid | $\mu_{\tilde{A}}(x) = \begin{cases} 0, & x \le \alpha \\ 2\left(\frac{(x-\gamma)}{(\gamma-\alpha)}\right)^2, & \alpha \le x \le \beta \\ 1-2\left(\frac{(x-\gamma)}{(\gamma-\alpha)}\right)^2, & \beta \le x \le \gamma \\ 1, & x \ge \gamma \end{cases}$ |   |

**Şekil 1. Bulanık üyelik fonksiyonlar**

Üçgensel Üyelik Fonksiyonu (Triangular), belirli bir minimum, maksimum ve tepe değeriyle tanımlanır. Basit ve hesaplaması kolaydır. Trapezoidal Üyelik Fonksiyonu, dört parametre ile tanımlanır, belirli bir aralıkta tam üyeliği temsil eder. Gauss Üyelik Fonksiyonu: Ortalama (μ) ve standart sapma

(σ) değerleriyle tanımlanır, yumuşak geçişli bir fonksiyondur. S-düzgü (Sigmoid) Üyelik Fonksiyonu: S biçiminde artan veya azalan bir yapıya sahiptir.

## 4. BULANIK ÇIKARIM SİSTEMLERİ

| | |
|---|---|
| Mamdani tipi çıkarım sistemi | 1975'te Ebrahim Mamdani tarafından önerilmiştir. Hem giriş hem de çıkış değişkenleri bulanık kümelerle ifade edilir. Dilsel kural tabanına dayalıdır ve özellikle kontrol sistemlerinde (örneğin sıcaklık, hız, nem kontrolü) yaygın biçimde kullanılır. |
| Sugeno tipi çıkarım istemi (Takagi–Sugeno FIS) | Takagi ve Sugeno (1985) tarafından geliştirilmiştir. Çıkış değişkeni bir bulanık küme yerine doğrusal veya sabit bir fonksiyon olarak tanımlanır. Matematiksel modelleme, optimizasyon ve adaptif sistemlerde sıkça tercih edilir. |
| Tsukamoto Tipi Bulanık Çıkarım Sistemi | Tsukamoto bulanık çıkarım sistemi**,** 1979 yılında Y. Tsukamoto tarafından geliştirilen ve Mamdani ile Sugeno sistemleri arasında bir geçiş niteliği taşıyan bir bulanık çıkarım yaklaşımıdır. |

Bulanık regresyon analizi (BRA) bağlamında, bu çıkarım sistemleri farklı roller üstlenir. Sugeno (Takagi–Sugeno) tipi çıkarım sistemi, çıktıyı doğrusal bir fonksiyon olarak tanımladığı için regresyon analizine en uygun yapıyı sunar. Bu yapı, özellikle ANFIS (Adaptive Neuro-Fuzzy Inference System) gibi hibrit modellerin temelini oluşturur ve fonksiyonel ilişki kurma (regresyon) görevlerinde yüksek performans gösterir. Buna karşın, Mamdani tipi sistemler daha çok kontrol sistemlerinde kullanılmakla birlikte, geleneksel BRA modelleri (Tanaka, FLSR) daha ziyade bir optimizasyon problemi olarak ele alınır ve doğrudan bir çıkarım sistemi kullanmaktan çok bulanık küme kavramına dayanır.

## 5. BULANIK REGRESYON MODELLERİNİN TEORİK ALTYAPISI

Bulanık regresyon, bağımlı değişkenin veya model katsayılarının bulanık olarak tanımlandığı, klasik regresyon modellerinden farklı olarak belirsizlikleri doğrudan model yapısı içerisinde ifade edebilen bir yaklaşımdır.

### 5.1. Model Yapısı

Basit bir bulanık regresyon modeli şöyle ifade edilebilir:

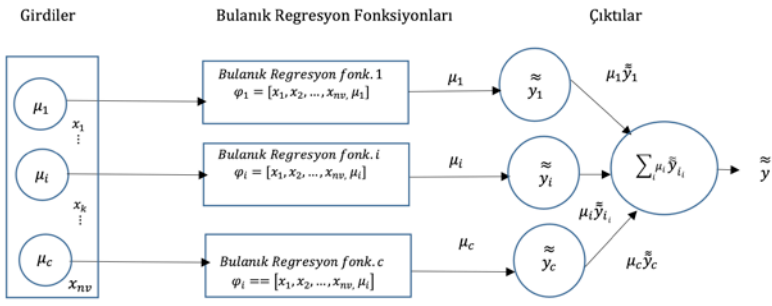
$$\tilde{Y} = \tilde{A}_0 + \tilde{A}_1 x_1 + \cdots + \tilde{A}_k x_k$$

Burada $\tilde{y}$, bulanık çıktı; $\tilde{A}_i$ katsayıları ise bulanık sayılardır. Bir başka yaklaşım şekli de

$$\tilde{\tilde{y}}_i = \sum_{i=1}^{c} \mu_i \tilde{y}_i$$

Biçimindedir.

- $c$: kural sayısı
- $\mu_i$: i-kuralın üyelik derecesi
- $\tilde{\tilde{y}}_i$: her kural için çıktılar




**Şekil 2. Bulanık regresyon (fuzzy regression) modelinin akış diyagramı (Voskoglou, 2020).**

Şekil 3'te akış diyagramında; girdi değişkenleri $x_1, x_2, \dots, x_k$ gibi bağımsız değişkenlerdir (Sıcaklık, yoğunluk, süre ve basınç gibi). Her biri sisteme bilgi taşır, ancak bu bilgiler kesin sayılar yerine bulanık değerler olarak da ele alınabilir.

$$\varphi_1(x_1, x_2, \dots, x_k), \varphi_2(x_1, x_2, \dots, x_k), \dots$$

Bu fonksiyonlar klasik regresyondaki katsayıların yerine geçer. Amaç, "her $x_i$" için bir bulanık tahmin fonksiyonu üretmektir.
Her fonksiyon, veriye göre bir üyelik derecesi hesaplar.

Her fonksiyonun çıktısı bir üyelik derecesi $\mu_i$'dir.

$$\mu_i \in [0,1]$$

Bu değer, ilgili kuralın (ya da fonksiyonun) ne kadar etkin olduğunu belirtir.

Örneğin:

- $\mu_1 = 0.9 \rightarrow$ 1. kural çok güçlü şekilde aktif
- $\mu_2 = 0.3 \rightarrow$ 2. kural zayıf şekilde aktif

Her bir üyelik derecesi, kendi bulanık tahmin değeriyle çarpılır. ($\mu_i \tilde{\tilde{y}}_i$) Bu adımda her kuralın çıktısı, sistemin toplam çıktısına katkısına göre ağırlıklandırılır. Sonra tüm bu ağırlıklı bulanık çıktılar toplanır:

$$\sum_i \mu_i \tilde{\tilde{y}}_i$$

Bu ifade, sistemde tanımlanan tüm kuralların katkılarını bütünleştirmektedir. Her bir regresyon fonksiyonunun çıktısı, kendi önem ağırlığı ($\mu_i$) doğrultusunda genel sistem çıktısına etki eder. En sağda yer alan çift dalgalı sembol $\tilde{\tilde{y}}$, sistemin bulanık tahminini, yani elde edilen bulanık çıktıyı temsil etmektedir. Bu çıktı, tek bir kesin değer yerine bir bulanık küme biçiminde ifade edilir. Gerektiğinde, bu bulanık küme keskinleştirme

(defuzzification) yöntemi aracılığıyla sayısal bir değere dönüştürülerek yorumlanabilir.

### 5.2. Yapay Zekâ Yaklaşımlarıyla Entegrasyon

Bulanık regresyon modellerinin yapay zekâ teknikleriyle bütünleştirilmesi, modelin hem tahmin doğruluğunu artırmakta hem de yorumlanabilirliğini güçlendirmektedir. Bulanık regresyon modelleri, belirsizlik ve ipucuyla tanımlı değişkenleri doğrudan modelleme imkânı sunarken, yapay zekâ teknikleri (örneğin sinir ağları, evrimsel algoritmalar, kümeleme yöntemleri) bu modellerin tahmin doğruluğunu ve adaptasyon yeteneklerini artırmaktadır (Ramly vd., 2023). Özellikle, bir bileşik model yapısında bulanık regresyonun interpretatif yapısı ile derin öğrenme ya da genetik optimizasyon gibi yaklaşımların otomatik öğrenme kapasitesi birleştirildiğinde, hem açıklanabilirlik hem de performans açısından üstünlük elde edilebilmektedir (Wu et al., 2019). Örneğin, açıklanabilir bulanık kümeleme-tabanlı regresyon algoritması, Takagi–Sugeno-Kang (TSK) yapısına dayalı olarak gruplama, üyelik fonksiyonlarının belirlenmesi ve ardından gradyan inişle parametre optimizasyonu uygulayarak geleneksel regresyon yöntemlerine kıyasla daha düşük RMSE ve MAE değerleri göstermiştir (Viaña vd., 2022). Bu bağlamda, bulanık regresyon-yapay zekâ entegrasyonu, belirsiz verilerle çalışılan finans, mühendislik ve karar destek sistemleri gibi alanlarda anlamlı bir metodolojik ilerleme olarak değerlendirilebilir.

#### 5.2.1. Neuro-Fuzzy Sistemler

ANFIS (Adaptive Neuro-Fuzzy Inference System), sinir ağı öğrenme yeteneği ile bulanık çıkarım sistemini birleştirir (Kısa ve ark., 2023; Danesh, R., et al., 2023). Bu model, kural tabanlı bulanık sistem ile ağırlık öğrenmesini harmanlar. ANFIS, regresyon görevlerinde etkilidir (Mantalas et al., 2025). Bir çalışmada, ANFIS ve klasik regresyon yöntemleri, malzeme

özelliklerinin tahmini açısından karşılaştırılmıştır (Krolo et al., 2019). Ayrıca, A DEPSO optimizasyonu ile hibrit ANFIS modeli geliştirilerek su kalite parametrelerinin tahmini gerçekleştirilmiştir (Ahmadianfar et al., 2022). Bu hibrit yapı, çıktıları doğrusal fonksiyonlar olarak tanımlayan Takagi-Sugeno (TSK) tipi çıkarım sistemini temel alır. TSK'nın regresyon analizi görevleri için Mamdani tipine göre tercih edilmesinin ana sebebi, çıktının doğrudan bir matematiksel regresyon fonksiyonu $y = f(x)$ olarak ifade edilebilmesi ve bu sayede sinir ağı algoritmaları (gradyan iniş) ile regresyon parametrelerinin kolaylıkla optimize edilebilmesidir. Bu, ANFIS'i regresyon görevlerinde güçlü ve uyarlanabilir kılar.

### 5.2.2. Derin Bulanık Sistemler

Derin öğrenme ve bulanık sistemlerin birleşimi “deep fuzzy systems” olarak adlandırılır. Özellikle regresyon problemlerinde yorumlanabilir yapıyı koruyarak doğruluk artırımı hedeflenir (Júnior et al., 2022). Bu yaklaşımlar, yüksek boyutlu veride hem temsil gücü hem de yorumlanabilirlik sağlar.

### 5.2.3. Uygulama Alanları ve Örnekler

- Enerji tüketimi tahmini/çevresel modelleme**:** Sensör verilerindeki belirsizlikleri modelleme.
- İnşaat/gayrimenkul fiyat tahmini**:** Karışık verilerde hiyerarşik bulanık regresyon kullanılarak (Demirhan ve ark., 2024).
- Malzeme bilimleri**:** Mikroyapı özelliklerini öngörmede ANFIS ve regresyon modellerinin karşılaştırılması (Krolo et al., 2019).
- Proses kontrol**:** Üretim sistemlerinde belirsiz ölçümler ile kontrol modelleri.
- Ekonomi / Finans**:** Talep tahmini, risk analizi gibi belirsizlik içeren modeller.

## 6. UYGULAMA

**Tablo 1. Modelin istatistiksel yapısı**

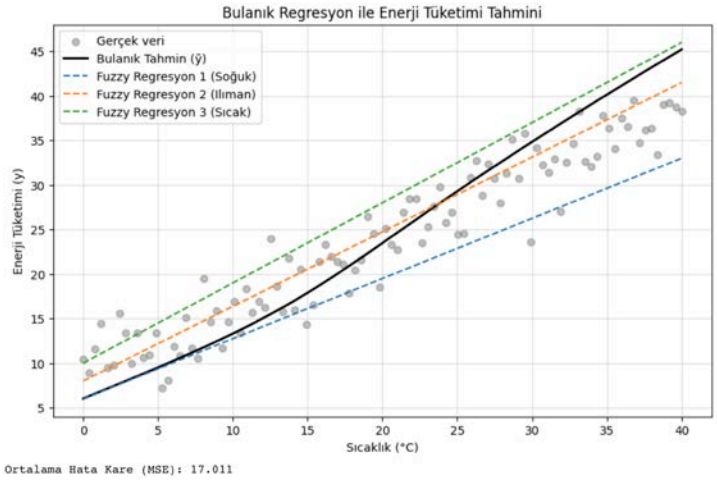
| Eleman | Rolü | Amaç |
|---|---|---|
| Bağımsız Değişken ($x$) | Sıcaklık (°C) | Tahmin için girdi. |
| Bulanık Katsayılar $(A_0, A_1)$ | $\tilde{A}_i = (a_i, c_i)$<br>$a_i$:Merkez<br>$c_i$: Yayılım/belirsizlik | Temel Amaç: Tüm gözlemleri kapsayacak şekilde toplam belirsizliği $\sum c_i$ en küçüklemek |

Bulanık regresyon modelimiz, katsayıları üçgensel bulanık sayılar $\tilde{A}_i = (a_i, c_i)$ biçiminde tanımlar.

**Tablo 2. Modelin kurgulanması ve amacı**

| Kod/Metin Kısmı | Amacı |
|---|---|
| Modelin Kurgulanması ve Amacı Tablosu | Tanaka modelinin temel mantığını ve DP katsayılarını açıklar. |
| Python Kütüphane ve Model Yapısı Kodu | numpy, linprog importlarını ve amaç fonksiyonu ile kısıtların (alt/üst sınır eşitsizliklerinin) matematiksel mantığını gösterir. |

Giriş bölümünde belirtildiği gibi, bulanık regresyon (BRA) modellerinin Python ortamında nasıl kurgulanabileceğine dair bir çerçeve sunulmaktadır. Aşağıdaki yapı, Şekil 3'teki Enerji Tüketimi Tahmini örneğini temel alır ve Tanaka'nın (1982) belirsizliği en küçüklemeye dayalı doğrusal programlama yaklaşımını yansıtır.



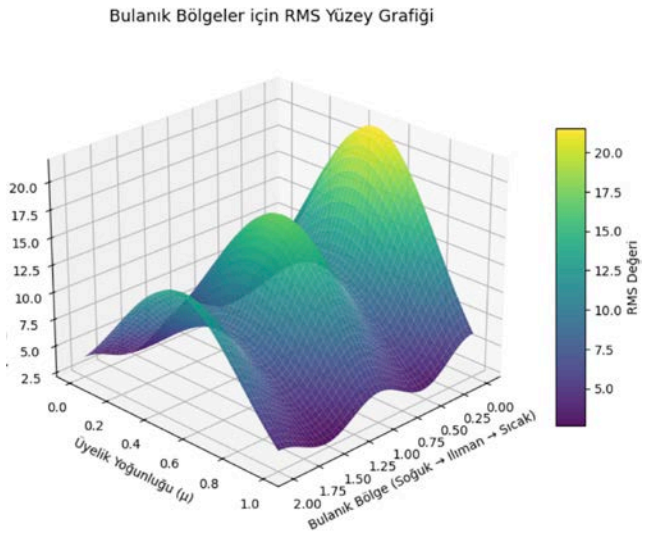

**Şekil 3. Bulanık regresyon ile enerji tüketim tahmini**

Şekil 3'te, sıcaklık (°C) ile enerji tüketimi (y) arasındaki ilişkiyi modelleyen bulanık regresyon analizi sonuçları gösterilmektedir. Gri noktalar gözlemlenen gerçek veri değerlerini, siyah çizgi ise bulanık regresyon modeli tarafından tahmin edilen enerji tüketimini ($\hat{y}$) temsil etmektedir. Grafikte ayrıca üç farklı bulanık regresyon bileşeni ("Soğuk", "Ilıman" ve "Sıcak" sıcaklık bölgeleri) ayrı ayrı gösterilmiştir. Bu bileşenler, sıcaklığın farklı aralıklarında enerji tüketim davranışının değişimini modelleyen yerel doğrusal regresyon eğilimlerini yansıtmaktadır.

Model, sıcaklık arttıkça enerji tüketiminin doğrusal biçimde yükseldiğini ortaya koymaktadır. "Soğuk" bölgeye (mavi kesikli çizgi) ait regresyon doğrusu, düşük sıcaklıklarda enerji tüketiminin nispeten düşük fakat değişken olduğunu; "Ilıman" bölge (turuncu kesikli çizgi) ise daha dengeli bir artış eğilimi sergilediğini göstermektedir. "Sıcak" bölge (yeşil kesikli çizgi) ise yüksek sıcaklıklarda enerji tüketiminin belirgin biçimde arttığını göstermektedir. Bu durum, soğutma sistemlerinin enerji ihtiyacının sıcaklığa duyarlı olduğunu desteklemektedir.

Modelin performansı, grafiğin alt kısmında belirtilen ortalama hata kareleri ($\mathbf{MSE = 17.011}$) değeri ile değerlendirilmiştir. Bu değer, tahmin edilen değerlerin gözlem değerlerine yakın olduğunu ve modelin genel olarak iyi bir uyum sağladığını göstermektedir. Ancak, bazı sıcaklık aralıklarında gerçek verilerin bulanık tahmin doğrularından hafif sapmalar göstermesi, enerji tüketimini etkileyen diğer faktörlerin (örneğin nem oranı, cihaz verimliliği, kullanıcı davranışı vb.) modele dahil edilmesiyle iyileştirilebileceğine işaret etmektedir.

Sonuç olarak, bu grafik, bulanık regresyon yaklaşımının enerji tüketimi gibi belirsizlik içeren değişkenlerin modellenmesinde etkin bir araç olduğunu göstermektedir. Model, klasik doğrusal regresyona göre çevresel koşulların farklı seviyelerinde daha esnek bir tahmin gücü sunmakta ve yapay zekâ tabanlı sistemlerin enerji yönetimi uygulamalarında kullanılabilirliğini ortaya koymaktadır.




**Şekil 4.Bulanık bölgeler için RMS yüzey grafiği**

Şekil 4'te elde edilen 3 boyutlu RMS (Root Mean Square Error) yüzey grafiği, bulanık regresyon modelinin farklı sıcaklık

bölgelerinde (Soğuk–Ilıman–Sıcak) gösterdiği hata dağılımını sürekli bir yüzey olarak ortaya koymaktadır. Bu grafik, modelin tahmin hatasını yalnızca sayısal olarak değil, bölgesel değişkenliğini de görsel biçimde analiz etme olanağı sağlamaktadır (Chukhrova & Johannssen, 2019).

Yüzeyin yüksek olduğu alanlar, RMS değerlerinin artmasıyla birlikte modelin daha fazla belirsizlik içerdiğini; yüzeyin düşük olduğu bölgeler ise modelin daha kararlı ve öngörülebilir çalıştığını göstermektedir. Bu durum, literatürde özellikle adaptif bulanık çıkarım sistemleri (ANFIS) **ve** hibrit regresyon modelleri ile yapılan çalışmalarda da benzer biçimde rapor edilmiştir (Ahmadianfar et al., 2022; Demirhan & Baser, 2024; Pakdel et al., 2025).

Elde edilen yüzeyin "Soğuk" bölgesinde RMS değerinin yüksek olması, düşük sıcaklıklarda enerji tüketim davranışının karmaşık yapısını ve modele dâhil edilmeyen çevresel faktörlerin etkisini göstermektedir. Benzer şekilde, Soğuk bölge hatalarının artışı**,** bulanık kümelerin üyelik derecelerinin düşük olduğu aralıklarda modelin öğrenme kapasitesinin sınırlı kaldığını ortaya koymaktadır. Buna karşın "Sıcak" bölgedeki düşük RMS değerleri, modelin bu aralıkta daha doğrusal ve kararlı ilişkiler yakaladığını göstermekte; bu da sıcaklık arttıkça enerji tüketiminin yapısal olarak daha öngörülebilir hale geldiğini göstermektedir (Demirkan ve ark., 2022).

Ayrıca RMS yüzeyinin eğimindeki azalma, parametre duyarlılığının azalması anlamına gelir. Bu da modelin belirli sıcaklık aralıklarında robust (sağlam) bir yapıya sahip olduğunu göstermektedir. Bu bulgu, hibrit yapay zekâ tabanlı bulanık sistemlerin parametre kararlılığı açısından literatürde belirtilen eğilimlerle uyumludur (Kong et al., 2025; Bhatia et al., 2025).

Sonuç olarak, 3B RMS yüzey grafiği**,** bulanık regresyon modellerinin performans değerlendirmesinde nitelikli bir araç

olarak öne çıkmaktadır. Bu yaklaşım hem belirsizlik yönetimi hem de üyelik fonksiyonlarının optimizasyonu süreçlerinde analitik bir referans noktası oluşturur (Júnior et al., 2022; Ojha et al., 2019). Ayrıca, yapay zekâ ve bulanık mantık tabanlı sistemlerin entegrasyonunda yoruma dayalı (explainable) yapay zekâ yöneliminin güçlenmesine katkı sunmaktadır (Viaña et al., 2022).

## 7. SONUÇ VE GELECEK ÇALIŞMALAR

1. Bulanık regresyon analizi (BRA), klasik regresyonun kesin (crisp) veri varsayımını gevşeterek belirsiz, eksik veya sözel ifadelerle tanımlanmış veriler üzerinde modelleme yapılmasına olanak tanır.
2. Yapay Zekâ Entegrasyonu ve Soft Computing: Günümüz yapay zekâ sistemleri, karar verme süreçlerinde sıklıkla bulanık mantık tabanlı yaklaşımlardan yararlanmaktadır. Bulanık mantık temelli yöntemlerin istatistiksel tekniklerle birleşimi, özellikle belirsizliğin doğrudan modellenmesine imkân tanıyan "soft computing" paradigması altında önemli bir yer tutar.
3. Modelin Yapısal Avantajı (Belirsizlik Yayılımı): Bulanık regresyon modeli, katsayıları bulanık sayılar biçiminde tanımlar. Modelin temel amacı, tüm gözlemleri kapsayacak şekilde toplam belirsizliği (yay genişliğini) en küçültmektir. Bu sayede hem belirsiz hem de eksik veriler üzerinde esnek tahminler üretebilme kapasitesine sahip olur.
4. Hibrit Sistemlerin Üstünlüğü (Neuro-Fuzzy): Bulanık regresyonun yapay zekâ teknikleriyle, özellikle sinir ağlarıyla (Neuro-Fuzzy) bütünleştirilmesi hem tahmin doğruluğunu artırmakta hem de modelin yorumlanabilirliğini güçlendirmektedir. Örneğin, ANFIS (Adaptive Neuro-Fuzzy

Inference System), sinir ağı öğrenme yeteneğini bulanık çıkarım sistemiyle birleştirerek regresyon görevlerinde etkili olmaktadır.

5. Geniş Uygulama Alanı: Bulanık regresyon modelleri, belirsizlik içeren verilerle çalışılan çeşitli alanlarda uygulanabilmektedir. Örnek uygulamalar arasında enerji tüketimi tahmini, çevresel modelleme, gayrimenkul fiyat tahmini (karmaşık verilerde), malzeme bilimleri ve ekonomi/finans sektöründeki risk analizi gibi modeller yer almaktadır.
6. Açıklanabilir Yapay Zekâ (XAI) Odaklı Modeller Geliştirme: Gelecek çalışmalar, özellikle Explainable AI (XAI) odaklı derin bulanık sistemler (deep fuzzy systems) geliştirilmesine yönlendirilmelidir. Bu, yüksek boyutlu verilerde yorumlanabilir yapıyı koruyarak doğruluk artırımını hedefleyecektir (Júnior et al., 2022; Viaña et al., 2022).
7. Hibrit Optimizasyon Tekniklerinin Kullanımı: Mamdani–TSK ve ANFIS gibi hibrit modellerin, optimizasyon algoritmaları ile birleştirilerek, yorumlanabilirlik ve doğruluk arasındaki dengenin kurulması önerilmektedir (Bhatia et al., 2025).
8. Gürültüye Dayanıklı (Robust) Regresyon Geliştirme: Veri gürültüsünün (outliers) olduğu durumlarda daha doğru tahminler yapabilmek için gürültüye dayanıklı (robust) bulanık regresyon modellerinin geliştirilmesi üzerinde çalışılmalıdır (Kong et al., 2025).
9. Yüksek Boyutlu Veri (Big Data) Ölçeklenebilirliği: Çok değişkenli ve büyük veri setleri için ölçeklenebilir bulanık modellerin tasarlanmasına odaklanılmalıdır (Xue et al., 2022).

**10.** Gerçek Dünya Vaka Çalışmaları ile Yöntem Doğrulama: Bulanık regresyon ve yapay zekâ entegrasyonu yöntemlerinin etkinliğini ve güvenilirliğini artırmak için, geliştirilen modellerin gerçek dünyaya uygulanmış kapsamlı vaka çalışmaları ve deneysel sonuçlarla doğrulanması önerilmektedir.

## KAYNAKÇA

Ahmadianfar, I., Shirvani-Hosseini, S., He, J. et al. (2022). An improved adaptive neuro fuzzy inference system model using conjoined metaheuristic algorithms for electrical conductivity prediction. Sci Rep 12, 4934.

Atasoy, D. (2001). Lojistik regresyon analizinin incelenmesi ve bir uygulaması (Yüksek lisans tezi). Cumhuriyet Üniversitesi, Sosyal Bilimler Enstitüsü, Sivas.

Azadeh, A., Alajdad, H., Bioki, T. (2014). A neuro-fuzzy regression approach for estimation: A case study. International Journal of Services and Operations Management. Inderscience.

Bhatia, A., Cordeiro de Amorim, R., De Feo, V. (2025). Hybrid Interval Type-2 Mamdani–TSK fuzzy system for regression analysis.

Chukhrova, N., Johannssen, A. (2019). Fuzzy regression analysis: Systematic review and bibliography. Applied Soft Computing, 84, 105708.

Danesh, M., Danesh, S., Maleki, A. (2023). An adaptive fuzzy inference system model to analyze complex regression structures. Frontiers in Engineering (FIE) Journal.

Demirhan, H., Baser, F. (2024). Hierarchical fuzzy regression functions for mixed predictors and an application to real estate price prediction. Neural Computing and Applications.

Demirkan Pişkin, M., Baş, E. (2022). Forecasting monthly housing sales with Type-1 fuzzy regression functions approach based on ridge regression. Karadeniz Fen Bilimleri Dergisi, 12(2), 173–183.

Júnior, J. S. S., Mendes, J., Souza, F., Premebida, C. (2023). Survey on deep fuzzy systems in regression applications: A view on interpretability.

Kısa, D. H., Özdemir, M. A., Güren, O., Alaybeyoğlu, A. (2023). A decision-making mechanism based on EMG signals and adaptive neural fuzzy inference system (ANFIS) for hand gesture prediction. Journal of the Faculty of Engineering and Architecture of Gazi University, 38(3), 1417–1430.

Kong, L., Song, C. (2025). Fuzzy robust regression based on exponential-type kernel functions. Journal of Computational and Applied Mathematics.

Krolo, J., Lela, B., Švagelj, Z., Jozić, S. (2019). Adaptive neuro-fuzzy and regression models for predicting microhardness and electrical conductivity. International Journal of Advanced Manufacturing Technology, 104(9–12), 4029–4041.

Mukaidono, M. (2001). Fuzzy logic for beginners. Beijing: World Scientific Publishing.

Ojha, V., Abraham, A., Snášel, V. (2019). Heuristic design of fuzzy inference systems: A review of three decades of research. arXiv preprint.

Pakdel, M., Razzaghnia, T., Fathi, K., Mostafaee, A. (2025). Estimation of fuzzy regression parameters with ANFIS and Bayesian methods. *Engineering Reports, 7*(1), e13086.

Ramly, N., Rusiman, M. S., Ismail, S., Hamzah, F. M., Gurunlu Alma, Ö. (2023). An adjustment degree of fitting on fuzzy linear regression model toward manufacturing income. *IAES International Journal of Artificial Intelligence (IJ-AI), 12*(2), 543–551.

Stanojević, B. (2023). Optimization-based fuzzy regression in full compliance. International Journal of Computers, Communications & Control (IJCCC).

Tansu, A. (2022). Fuzzy regression analysis with a proposed model. Technium Science Journal. IDEAS/RePEc.

Viaña, J., Ralescu, S., Kreinovich, V. (2022). Explainable fuzzy cluster-based regression algorithm with gradient descent learning. Complex Engineering Systems, 2(8).

Voskoglou, M. G. (2020). Fuzzy sets, fuzzy logic and their applications. Manchester: MDPI.

Wu, D., Lin, C.-T., Huang, J., Zeng, Z. (2019). On the functional equivalence of TSK fuzzy systems to neural networks, mixture of experts, CART, and stacking ensemble regression.

Xue, G., Chang, Q., Wang, J., Zhang, K., Pal, N. (2022). An Adaptive Neuro-Fuzzy System with Integrated Feature Selection and Rule Extraction for High-Dimensional Classification Problems.

# BÜYÜK DİL MODELLERİNDE PERFORMANS OPTİMİZASYONLARI: O(N²) ATTENTİON ENGELİ VE ÜRETİM ODAKLI ÇÖZÜMLER

**Mazhar KAYAOĞLU**[1]

**Uğur BERDİBEK**[2]

## 1. GİRİŞ

Büyük dil modelleri (Large Language Models, LLM'ler), doğal dil işleme görevlerinde başarılar elde etmiştir. Temelini oluşturan Transformer mimarisi, self-attention mekanizmasıyla bağlamı etkili biçimde modellemekte; ancak bu mekanizmanın zaman ve bellek karmaşıklığı, uzun dizilerde ölçeklenebilirlik sorunları yaratmaktadır (Vaswani et al., 2017, s. 1). $\mathcal{O}(n^2)$ karmaşıklığı, eğitim ve üretim aşamalarında kaynak tüketimini artırarak uygulamalarda verimliliği ve maliyet etkinliğini sınırlamaktadır. Bu kitap bölümü, 2015-2025 arasında yayınlanmış 30 makale ve teknik raporu derleyerek $\mathcal{O}(n^2)$ attention engelini aşmaya yönelik yöntemleri, donanım odaklı inovasyonları ve üretim araçlarını incelemekte; Transformer tabanlı LLM'lerin evrimini bütüncül bir çerçevede özetlemeyi amaçlamaktadır. Transformer mimarisi, "Attention is All You Need" ile self-attention katmanlarının bütünleştirilmesine dayanmakta ve geleneksel RNN/CNN yaklaşımlarına üstünlük sağlamaktadır (Vaswani et al., 2017, s. 1). Ancak attention hesaplamalarının kuadratik karmaşıklığı $\mathcal{O}(n^2)$, dizi uzunluğu

[1] Dr. Bingöl Üniversitesi, Enformatik Bölümü, Bingöl/ TÜRKİYE, ORCID: 0000-0002-5807-9781.

[2] Dr. Bağımsız Araştırmacı, Bingöl / TÜRKİYE, ORCID: 0000-0003-3202-8342.

(n) arttıkça maliyeti katlanarak yükselterek özetleme ve çok turlu diyalog gibi görevlerde uygulanabilirliği kısıtlamaktadır. Literatürde bu engeli hafifletmek üzere seyrek ve lineer attention mekanizmaları, nicelleştirme (quantization), spekülatif kod çözme (speculative decoding) ve self-attention'a alternatif mimariler geliştirilmiştir. Derlenen 30 kaynak, FlashAttention serisi, durum uzay modelleri (SSM'ler) ve üretim odaklı sistemler çalışmaları içermektedir. FlashAttention, donanım farkındalığı ve IO-awareness ile bellek erişimini optimize ederken (Dao et al., 2022, s. 2; Shah et al., 2024, s. 2), Mamba ve benzeri SSM tabanlı yaklaşımlar self-attention'ın kuadratik maliyetine karşılık dizi uzunluğuna doğrusal $\mathcal{O}(n)$ zaman karmaşıklığı sunmaktadır (Gu & Dao, 2023, s. 6; Dao & Gu, 2024, s. 6). vLLM ve DeepSpeed throughput'u 1.5-4 kata kadar artırarak deployment'ı hızlandırmaktadır (Kwon et al., 2023, s. 4; Aminabadi et al., 2022, s. 7). Quantization teknikleri, model boyutunu küçültürken doğruluk kaybını sınırlamaya ve enerji verimliliğini artırmaya yöneliktir; GPT-3 gibi büyük modeller bağlamında (Brown et al., 2020, not: Bu bölümde doğrudan referans verilmemiş olsa da genel literatür bağlamında belirtilmiştir), bu optimizasyonlar maliyet ve sürdürülebilirlik açısından kritik görülmektedir (Dettmers et al., 2022, s. 5; Xiao et al., 2023, s. 5). Buna karşın uzun bağlam extrapolation'ı, donanım heterojenliği ve etik/çevresel etkiler gibi sorunlar devam etmektedir. Bölüm yapısı şöyledir: Bölüm 2-4 temel attention ve donanım odaklı optimizasyonları; Bölüm 5-9 çoklu sorgu, spekülatif kod çözme, KV-cache ve bellek teknikleri, nicelleştirme, konum kodlama ve uzun bağlam yaklaşımlarını; Bölüm 10-11 alternatif mimariler ve üretim çerçevelerini ele almakta; Bölüm 12 ise genel çıkarımlar ve gelecek araştırma yönelimlerini tartışmaktadır.

## 2. TEMEL ATTENTİON MEKANİZMASI

Transformer mimarisi, büyük dil modellerinin (LLM'ler) temel taşını oluşturan bir mimaridir ve doğal dil işleme görevlerinde bağlamsal temsilleri etkili biçimde yakalar. Bu mimari, geleneksel sıralı modellerin (örneğin, RNN'ler) sınırlamalarını aşmak üzere tasarlanmış, paralel hesaplama yeteneğiyle eğitim verimliliğini artırmaktadır. Temelini oluşturan attention mekanizması, giriş dizisindeki her öğenin diğerleriyle dinamik ilişkilerini modelleyerek uzun menzilli bağımlılıkları (long-range dependencies) yakalamada üstün performans sergilemektedir (Vaswani et al., 2017, s. 1). Bu bölüm, Transformer'ın temel bileşenlerini özetleyerek self-attention'ın matematiksel temellerini ve $\mathcal{O}(n^2)$ karmaşıklık engelini tartışmakta; böylece sonraki bölümlerde ele alınacak optimizasyon stratejilerine zemin hazırlamaktadır. Transformer modeli, kodlayıcı (encoder) ve kod çözücü (decoder) katmanlarından oluşmakta olup, her katman self-attention, feed-forward ağlar ve katman normalizasyonu gibi alt bileşenleri içermektedir. Self-attention mekanizması, giriş vektörlerini sorgu (query), anahtar (key) ve değer (value) matrislerine dönüştürerek hesaplanır: $\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$, burada $Q$, $K$ ve $V$ sırasıyla sorgu, anahtar ve değer matrisleri, $d_k$ ise boyutluluk faktörüdür (Vaswani et al., 2017, s. 1). Bu formül, her token'ın diğer token'larla benzerliğini hesaplayarak ağırlıklı bir toplam üretir ve modelin bağlamı global olarak değerlendirmesine olanak tanır. Multi-head attention, bu işlemi çoklu başlıklar altında yürüterek farklı alt uzaylardaki temsilleri yakalamayı ve genelleme yeteneğini artırmayı amaçlamaktadır. Bu mekanizma, paralel hesaplama ve uzun menzilli bağımlılık modellemesinde avantajlar sunarken, $\mathcal{O}(n^2)$ zaman ve bellek karmaşıklığı nedeniyle yüksek maliyetlidir. Attention skor

matrisinin hesaplanması, dizi uzunluğu $n$ ile kuadratik ölçeklenmekte, uzun bağlamlı görevlerde GPU bellek sınırlarını zorlamakta ve inference latency'sini artırmaktadır. $\mathcal{O}(n^2)$ karmaşıklığın kaynağı, her token'ın tüm diğer token'larla etkileşimini gerektiren tam yoğun (dense) attention yapısıdır; büyük modellerde bu durum yüksek bellek gereksinimi doğurmaktadır. Literatürde bu engeli hafifletmeye yönelik çalışmalar, Transformer'ın evrimini tetiklemiştir: seyrek attention yöntemleri, karmaşıklığı $\mathcal{O}(n \log n)$ veya lineer seviyeye indirmeyi hedeflemekte (Kitaev et al., 2020, s. 2); donanım odaklı optimizasyonlar ise IO-aware algoritmalarla pratik hızlanmalar sağlamaktadır (Dao et al., 2022, s. 2). Uzun bağlam extrapolation'ı bağlamında, kısa dizilerle eğitilen modellerin uzun test dizilerinde performans kaybı yaşadığı gösterilmiş (Press et al., 2022, s. 5) ve enerji tüketimi ile çevresel etkiler, LLM ölçeklenebilirliğinin etik boyutunu gündeme taşımıştır. Alternatif mimariler (örneğin, SSM tabanlı yaklaşımlar; Gu & Dao, 2023, s. 6) temel prensipleri korurken bu maliyetleri azaltmayı amaçlamakta olup, sonraki bölümler bu temel üzerinde geliştirilen optimizasyonları ayrıntılandıracaktır.

## 3. DONANIM ODAKLI ATTENTİON OPTİMİZASYONLARI

Transformer mimarisinin temel attention mekanizması, $\mathcal{O}(n^2)$ karmaşıklığı nedeniyle hesaplama yoğunluğunu artırmakta; ancak bu engel, modern GPU bellek hiyerarşisini hedefleyen IO-aware algoritmalarla hafifletilebilmektedir (Dao et al., 2022, s. 2). Bu bölüm, FlashAttention serisi üzerinden donanım farkındalığının attention mekanizmasını nasıl dönüştürdüğünü incelemekte; bu yaklaşımlar pratikte 7-10 kata varan hızlanmalarla LLM'lerin uygulanabilirliğini artırmaktadır. Standart attention implementasyonunda softmax için ara

matrislerin tam materyalizasyonu, bellekler arasında tekrarlı veri transferleri gerektirerek maliyeti yükseltirken, FlashAttention tiling ve recomputation teknikleriyle bu matrisleri bloklar halinde işleyip softmax'ı çevrimiçi hesaplamakta, bellek kullanımını yaklaşık %50 azaltmakta ve 7.6 kata varan hızlanma sağlamaktadır (Dao et al., 2022, s. 2). FlashAttention-3, Hopper nesli GPU'lara uyarlanarak asenkron işlem ve düşük hassasiyetli (FP8) aritmetik desteğiyle 1.5-2 kat ek hızlanma elde etmekte, 1.2 PFLOPS/s düzeyine ulaşmakta ve Llama 7B modellerinde inference latency'sini yaklaşık %40 azaltmaktadır (Shah et al., 2024, s. 2). Bu donanım odaklı yaklaşımlar, quantization teknikleriyle (Dettmers et al., 2022, s. 5) ve vLLM gibi üretim çerçeveleriyle (Kwon et al., 2023, s. 4) birleştirildiğinde ek verimlilik kazanımları üretmektedir. Ancak temel karmaşıklık değişmemekte, uzun dizilerde maliyet sürmekte ve GPU merkezli tasarım, donanım bağımlılığı nedeniyle portabiliteyi sınırlamaktadır; bu nedenle Mamba gibi alternatif mimarilerle (Gu & Dao, 2023, s. 6) hibrit sistemler geliştirmek, $\mathcal{O}(n^2)$ engelini donanım, sistem ve algoritma düzeylerinde birlikte hafifletmeyi hedeflemektedir.

## 4. SEYREK VE LİNEER ATTENTİON YÖNTEMLERİ

Seyrek (sparse) ve lineer attention yöntemleri, Transformer mimarisinin temelindeki $\mathcal{O}(n^2)$ karmaşıklığın, attention matrisinin tam yoğunluğundan (dense) kaynaklandığını kabul ederek, bu matrisi seyrelterek veya yaklaşıklayarak maliyeti düşürmeyi amaçlamaktadır. Bu yaklaşımlar, attention'ı yerel veya rastgele alt kümelerle sınırlandırarak uzun diziler için $\mathcal{O}(n \log n)$ ya da $\mathcal{O}(n)$ karmaşıklığa inebilmektedir (Tay et al., 2020, s. 3). Bu bölüm, Reformer, Longformer, Big Bird ve Performers gibi yöntemleri, Efficient Transformers derlemesini

çerçeve metin olarak kullanarak incelemekte ve bu tekniklerin donanım odaklı optimizasyonlarla (Bölüm 3) tamamlayıcılığını vurgulamaktadır.

Reformer, locality-sensitive hashing (LSH) ile benzer token'ları aynı hash bucket'larında gruplayarak, her sorgunun yalnızca ilgili anahtarlarla etkileşmesini sağlar ve böylece attention karmaşıklığını yaklaşık $\mathcal{O}(n \log n)$ düzeyine indirger (Kitaev et al., 2020, s. 2). LSH attention, sorgu vektörlerini rastgele hiperdüzlemlere yansıtarak benzerlik skorlarını yaklaştırır; tam $\mathcal{O}(n^2)$ matris yerine hash tabanlı bir yapı kullanır ve reversible layers ile bellek tüketimini azaltır. Deneysel sonuçlar, 64K token'lık dizilerde yaklaşık %50 bellek tasarrufu sağlarken, hashing'in stokastik doğası doğrulukta sınırlı da olsa bozulma riski taşımaktadır (Kitaev et al., 2020, s. 2).

Longformer, uzun belge işleme için sliding window, dilated ve global attention bileşenlerini birleştirerek, pencere boyutu $w$ için yaklaşık $\mathcal{O}(n \cdot w)$ karmaşıklık sunar (Beltagy et al., 2020, s. 2). Dilated pattern, her $k$. token'ın örneklenmesiyle kapsama alanını genişletirken, az sayıda global token (örneğin, başlık, özel işaretleyiciler) üzerinden uzun menzilli bağımlılıkları korur. LED (Longformer Encoder-Decoder), 16K token'lık bağlamlarda arXiv özetleme gibi görevlerde ROUGE skorlarında iyileşmeler rapor etmektedir (Beltagy et al., 2020, s. 2). Big Bird ise block-sparse bir şema ile yerel bloklar, rastgele bağlantılar ve global token'ları birleştirerek teorik olarak Turing-complete ve universal approximator özelliklerini muhafaza eder; dikkat grafiğini seyrek bir $G = (V, E)$ yapısı olarak tanımlayarak karmaşıklığı $\mathcal{O}(n)$ düzeyine indirir (Zaheer et al., 2020, s. 2).

Performers, FAVOR+ (Fast Attention Via positive Orthogonal Random features) algoritmasıyla softmax attention'ı random feature approximation aracılığıyla lineerleştirir:

$\text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \approx \phi(Q)\left(\phi(K)^\top V\right)$; burada $\phi$ rastgele Fourier tabanlı özellikleri temsil eder ve karmaşıklığı $\mathcal{O}(n \cdot d)$ düzeyine indirger (Choromanski et al., 2020, s. 3). Efficient Transformers derlemesi, bu yaklaşımları sistematik biçimde karşılaştırarak, uzun bağlamlarda enerji verimliliği ve ölçeklenebilirlik kazanımlarına rağmen approximation hataları ve görev bağımlılığının ortak sınırlamalar olduğuna işaret etmektedir (Tay et al., 2020, s. 3). Bu çerçeve, ilerleyen bölümlerde ele alınacak çoklu sorgu ve spekülatif decoding teknikleriyle birlikte, $\mathcal{O}(n^2)$ engelinin algoritmik düzeyde aşılmasında yol gösterici bir rol üstlenmektedir.

## 5. ÇOKLU SORGU VE GRUPLU SORGU ATTENTİON

Transformer mimarisinin attention katmanları, multi-head attention (MHA) yapısıyla zengin temsiller üretmekte, ancak bu yapı bellek ve hesaplama maliyetini artırarak özellikle üretim (inference) aşamasında KV-cache boyutunu büyütmektedir. Çoklu sorgu attention (MQA) ve gruplu sorgu attention (GQA), head sayısına bağlı bu yükü azaltarak KV-cache verimliliğini yükseltmekte ve $\mathcal{O}(n^2)$ engelini dolaylı biçimde hafifletmektedir (Ainslie et al., 2023, s. 3). Bu bölüm, GQA'nın matematiksel çerçevesini ve büyük dil modellerindeki kullanımını inceleyerek, attention tasarımında verimlilik odaklı bir bakış açısı sunmaktadır. MHA, attention'ı birden çok başlık altında yürüterek farklı alt uzaylarda bağımlılıkları yakalar; her head için ayrı $K$ ve $V$ matrislerinin saklanması, KV-cache boyutunu head sayısıyla orantılı büyütmekte ve uzun bağlamlarda bellek baskısını artırmaktadır (Vaswani et al., 2017, s. 1). MQA, tüm head'ler için tek bir $K$ ve $V$ kullanarak cache karmaşıklığını

düşürür; bu, inference hızında artış sağlarken, head çeşitliliğini azaltarak bazı görevlerde eğitim stabilitesini olumsuz etkileyebilmektedir (Ainslie et al., 2023, s. 3). GQA, MHA ve MQA arasında bir denge kurar. Sorgular $g$ gruba ayrılır, her grup $m$ head'i paylaşır ve her grup için tek bir $K/V$ seti tutulur; toplam head sayısı $H = g \cdot m$ olup cache karmaşıklığı $g$ ile ölçeklenir (Ainslie et al., 2023, s. 3). Böylece, MHA'ya göre belirgin bellek tasarrufu sağlanırken, MQA'ya kıyasla daha fazla head çeşitliliği korunur. PaLM ve Llama 2 gibi modellerde GQA kullanımıyla, MHA'ya yakın doğruluk yanında inference'ta hızlanma ve otoregresif üretimde daha düşük latency elde edilmiştir (Ainslie et al., 2023, s. 3). GQA için önerilen uptraining süreci, önceden eğitilmiş MHA checkpoint'lerinin GQA'ya dönüştürülmesini, head'lerin benzerliklerine göre gruplanmasını, $K/V$ ağırlıklarının ortalanmasını ve kısa bir yeniden eğitimle perplexity kaybının %1'in altında tutulmasını amaçlamakta; bu da tek GPU üzerinde fine-tuning gibi kaynak kısıtlı senaryolarda avantaj sağlamakta ve çok uzun bağlamlarda Longformer gibi seyrek yöntemlerle birleştirildiğinde daha etkili olmaktadır (Ainslie et al., 2023, s. 3; Beltagy et al., 2020, s. 2).

Karşılaştırmalı çalışmalar, GQA'nın MQA'ya göre doğal dil anlama görevlerinde (örneğin GLUE) %2-3 daha yüksek skorlar elde ederken, MHA'dan daha hızlı çalıştığını göstermektedir (Ainslie et al., 2023, s. 3). KV-cache optimizasyonları ve verimli servis altyapılarıyla (örneğin vLLM; Kwon et al., 2023, s. 4) birlikte kullanıldığında, toplam enerji tüketiminde azalmalara katkı sunduğu rapor edilmiştir. Sonuç olarak, çoklu sorgu ve gruplu sorgu attention varyantları, Transformer attention katmanlarını KV-cache odaklı yeniden düzenleyerek $\mathcal{O}(n^2)$ engelini pratik düzeyde hafifletmekte ve LLM'lerin üretim senaryolarında verimlilik artışları sağlamaktadır.

## 6. SPEKÜLATİF KOD ÇÖZME YÖNTEMLERİ

Autoregressive büyük dil modelleri (LLM'ler), token'ları sıralı ürettiği ve $\mathcal{O}(n^2)$ attention engeliyle birleştiği için inference latency'si nedeniyle gerçek zamanlı uygulamalarda sınırlıdır. Spekülatif kod çözme, küçük bir draft model ve paralel doğrulama yoluyla birden çok token'ı eşzamanlı tahmin ederek 2-3 kata varan hızlanmalar sunar (Leviathan et al., 2023, s. 3). Temel şema, verili bağlamda draft modelin uzunluğu $L$ olan aday diziler üretmesi ve ana modelin bunları paralel doğrulamasına dayanır; kabul oranının beklenen değeri $\mathbb{E}[\alpha]$ olduğunda, ana modele çağrı başına üretilen token sayısının beklenen değeri yaklaşık $\mathbb{E}[\alpha L]$'dir.

Medusa, spekülatif yaklaşımı çoklu decoding head'leriyle genişleterek aynı adımda birden çok token tahmin etmekte ve LLaMA/Vicuna tabanlı modellerde 2.2-2.8 kat hızlanma elde etmektedir (Cai et al., 2024, s. 4). EAGLE, feature-level autoregression ve ağaç tabanlı draft yapılarıyla LLaMA2-Chat 70B üzerinde 2.7-3.5 kat hızlanma ve iki kat throughput artışı raporlar (Li et al., 2024a, s. 4). EAGLE-2 ise context-aware dinamik ağaçlarla dallanmayı bağlama duyarlı güven eşiklerine göre uyarlayarak ek %20-40 hızlanma ve toplamda 3.05-4.26 kat kazanç sağlamaktadır (Li et al., 2024b, s. 4). GQA gibi attention varyantları (Ainslie et al., 2023, s. 3) ve KV-cache optimizasyonlarıyla (Kwon et al., 2023, s. 4) birleştiğinde, bu yöntemler $\mathcal{O}(n^2)$ engelini algoritmik olarak değiştirmeksizin LLM üretim performansını dönüştürmektedir.

## 7. KV-CACHE OPTİMİZASYONLARI VE SÜREKLİ TOPLU İŞLEME

Autoregressive büyük dil modellerinde (LLM'ler), attention sırasında anahtar-değer (KV) çiftlerinin önbelleğe

alınması verimliliği artırır; ancak cache'in dinamik yönetimi, özellikle değişken dizi uzunluklarında bellek parçalanmasına yol açarak $\mathcal{O}(n^2)$ engelini inference aşamasında ağırlaştırır. KV-cache optimizasyonları ve sürekli toplu işleme (continuous batching), blok tabanlı bellek tahsisi ve iterasyon seviyesi zamanlama ile bu sorunu hafifletmekte; throughput'u birkaç kattan onlarca kata kadar yükseltebilmektedir (Kwon et al., 2023, s. 4; Yu et al., 2022, s. 4; Leviathan et al., 2023, s. 3).

PagedAttention, vLLM inference engine'inin çekirdeğini oluşturarak KV-cache'i sanal bellek sayfalarına benzer biçimde yönetir, GPU bellek parçalanmasını azaltır ve sabit boyutlu bloklardan oluşan bir düzenle serbest blokların yeniden kullanımını sağlar. Böylece Hugging Face Transformers tabanlı çözümlere kıyasla 2-24 kata varan throughput iyileşmeleri ve esnek dağıtım elde edilir (Kwon et al., 2023, s. 4). Orca tabanlı sürekli toplu işleme, istek iterasyonlarını düğüm ve bağımlılıkları kenar olarak modelleyen bir çizge üzerinde iterasyon düzeyi zamanlama uygular; hazır iterasyonlardan batch oluşturarak GPT-3 ölçeğinde NVIDIA FasterTransformer'a kıyasla benzer gecikme altında 36.9 kata kadar throughput artışı sağlar (Yu et al., 2022, s. 4). PagedAttention bellek yönetimine, continuous batching ise zamanlamaya odaklandığından, vLLM gibi engine'lerde birlikte kullanımları, donanım bağımlılığı ve dinamik yük dengesizliğine karşın, $\mathcal{O}(n^2)$ engelini pratik düzeyde hafifleten tamamlayıcı optimizasyonlar sunmaktadır (Kwon et al., 2023, s. 4; Leviathan et al., 2023, s. 3).

## 8. KONUM KODLAMA VE UZUN BAĞLAM YÖNTEMLERİ

Büyük dil modellerinin (LLM'ler) parametre ölçeği ve attention katmanlarının karmaşıklığı, hesaplama maliyetini yükselterek deployment'ı zorlaştırmaktadır. Nicelleştirme

(quantization), ağırlık ve aktivasyonları 8- veya 4-bit temsillere indirerek bellek gereksinimini azaltır, uygun donanımda inference hızını artırır ve LLM.int8(), SmoothQuant, GPTQ, AWQ, QLORA ile KV-cache ve continuous batching optimizasyonları çerçevesinde ele alınmaktadır. LLM.int8(), 8-bit matris çarpımıyla Transformer katmanlarını quantize eder ve aykırı aktivasyonları FP16 alt kümesine taşıyarak %50 tasarruf sağlar. SmoothQuant, aktivasyon aykırılarını ağırlıklara yeniden ölçekleyerek INT8 quantization'ı stabilize eder ve FP16'ya yakın perplexity ile kazanımlar üretir. GPTQ, katman bazlı weight-only quantization ile 3-4 bit seviyesinde doğruluğu korurken, AWQ activation-aware weight quantization ile önemli ağırlıkları saklayıp yalnızca diğerlerini 4-bit'e indirger. QLORA, 4-bit NormalFloat ve LoRA'yı paged optimizers ve double quantization ile birleştirerek 65B modellerin 48GB GPU üzerinde fine-tuning'ini ve ChatGPT-benzeri performans sağlar. SmoothQuant'un aktivasyon, GPTQ ve AWQ'nun ise weight-only yapıda olduğu; birlikte kullanıldıklarında model boyutunu dört kat küçültebildikleri ve $\mathcal{O}(n^2)$ attention engelini doğrudan azaltmasalar da KV-cache ve continuous batching ile ölçeklenebilir inference sundukları raporlanmaktadır (Dettmers et al., 2022, s. 5; Kwon et al., 2023, s. 4; Yu et al., 2022, s. 4; Xiao et al., 2023, s. 5; Frantar et al., 2022, s. 5; Lin et al., 2023, s. 5; Dettmers et al., 2023, s. 5).

## 9. KONUM KODLAMA VE UZUN BAĞLAM YÖNTEMLERİ

Transformer mimarisinde konum bilgisi, token dizilerinin sıralı yapısını kodlayarak bağlamsal temsilleri zenginleştirir; ancak klasik sabit sinüzoidal konum kodlamaları, eğitimdeki bağlam uzunluğunu aşan dizilerde kötü genelleşerek uzun bağlamlı görevlerde attention darboğazını ağırlaştırır. Modern

yaklaşımlar, relative konumlandırma ve interpolation teknikleriyle bu sınırlamayı aşarak, eğitim penceresinin çok ötesine uzanan bağlamlarda LLM'lerin kullanılabilirliğini artırmaktadır (Su et al., 2021, s. 6; Chen et al., 2023, s. 6). Bu bölüm RoFormer (RoPE), ALiBi ve Position Interpolation yöntemlerini, matematiksel ilkeleri, extrapolation kapasiteleri ve pratik etkileriyle birlikte; ayrıca nicelleştirme stratejileri (Bölüm 8) ve alternatif mimarilerle (Bölüm 10) etkileşimleri bağlamında tartışmaktadır. RoPE, sorgu-anahtar vektörlerine konum bağımlı dönme operatörü uygulayarak mutlak pozisyonu relative mesafeye dönüştürür ve uzun metin görevlerinde LLM'lerde yaygın olarak benimsenmiştir (Su et al., 2021, s. 6). ALiBi, ayrı konum embedding'lerini kaldırıp attention logit'lerine uzaklık temelli lineer bias ekleyerek kısa bağlamda eğitilmiş modellerin daha uzun dizilere genellemesini sağlar (Press et al., 2022, s. 6). Position Interpolation, RoPE tabanlı modellerde pozisyon indekslerini yeniden ölçekleyerek bağlam penceresini onlarca kat genişletirken LongBench benzeri benchmark'larda rekabetçi doğruluğu korur (Chen et al., 2023, s. 6). Bu teknikler, seyrek attention yöntemleriyle birlikte attention darboğazını anlamlı ölçüde hafifleten tamamlayıcı araçlar sunar (Tay et al., 2020, s. 3).

## 10. O(1) KARMAŞIKLIKLI ALTERNATİF MİMARİLER: DURUM UZAY MODELLERİ

Transformer temelli modellerdeki $\mathcal{O}(n^2)$ attention engeli uzun dizilerde ölçeklenebilirliği sınırlandırırken, durum uzay modelleri (State Space Models, SSM'ler) lineer zaman karmaşıklığı $\mathcal{O}(n)$ veya sabit $\mathcal{O}(1)$ adım başına hesaplama ile bu sorunu hafifleterek Transformer'lara rekabetçi bir alternatif sunmaktadır. Sürekli dinamik sistemlerden esinlenen bu mimariler, eğitim paralelliğini korurken inference sırasında $\mathcal{O}(1)$ adım başına maliyet hedefler. Bu bölüm, Mamba, Mamba-2,

RetNet ve RWKV mimarilerini özetlemektedir. Mamba, selective SSM yaklaşımıyla girişe bağımlı durum geçişleri tanımlar ve uzun menzilli bağımlılıkları lineer zamanlı sekans modeli çerçevesinde yakalar (Gu & Dao, 2023, s. 6). Sürekli zaman SSM denklemi

$$\dot{x}(t) = Ax(t) + Bu(t),\ y(t) = Cx(t) + Du(t)$$

şeklindedir. 1.4B parametreli Mamba, dil modellemede benzer ölçekli Transformer'larla rekabetçi perplexity elde ederken inference throughput'unu birkaç kata kadar artırmaktadır. Mamba-2, State Space Duality (SSD) framework'üyle SSM ve attention arasındaki bağı formalize eder; $A$ matrisinin uygun bir tabanda diyagonal yapıya indirgenmesiyle structured matris çarpımları üzerinden $\mathcal{O}(n)$ karmaşıklıkta attention-benzeri işlemler gerçekleştirir (Dao & Gu, 2024, s. 7). WikiText103 sonuçları, 3B parametreli Mamba-2'nin Transformer tabanlı modellere göre daha düşük loss ve 2-8 kat eğitim/inference hızlanması sağladığını göstermektedir. RetNet (Retentive Network), retention mekanizmasıyla RNN-benzeri ardışık güncellemeyi eğitimde paralel, inference'ta ise hafif bir yapı olarak formüle eder:

$$h_t = \alpha \odot h_{t-1} + \beta \odot x_t$$

(Sun et al., 2023, s. 7). Chunk-wise paralelleştirme, Transformer'a yakın eğitim hızı sağlarken, inference'ta $\mathcal{O}(1)$ adım başına maliyet ve düşük bellek ayak izi sunar. RWKV, RNN ve Transformer ilkelerini birleştiren hibrit bir tasarım olup, time-mixing ile geçmiş durumların üstel sönümlü karışımını ve mevcut girdinin lineer katkısını birleştirir:

$$h_t = w \odot h_{t-1} + (1-w) \odot g(x_t)$$

(Peng et al., 2023, s. 7). Böylece sabit $\mathcal{O}(1)$ adım başına maliyetle uzun bağlamları işleyebilmekte; 14B parametreli RWKV

modelleri Pile veri kümesinde benzer büyüklükteki Transformer'larla karşılaştırılabilir performans sergilemektedir. Karşılaştırmalı analizler, Mamba'nın selective SSM, RetNet'in retention, RWKV'nin hibrit RNN-Transformer tasarımı ve Mamba-2'nin dualite tabanlı yapısı sayesinde $\mathcal{O}(1)$ adım başına maliyet avantajı sunduğunu; uygun senaryolarda 4-8 kata varan hızlanmaların rapor edildiğini göstermektedir (Gu & Dao, 2023, s. 6; Sun et al., 2023, s. 7; Peng et al., 2023, s. 7; Dao & Gu, 2024, s. 7; Chen et al., 2023, s. 6). Bu mimariler, klasik $\mathcal{O}(n^2)$ attention engelini yapısal düzeyde aşarak, konum interpolation teknikleriyle (Chen et al., 2023, s. 6) birlikte kullanıldığında enerji verimliliğinde anlamlı iyileşmeler ve edge computing senaryolarında daha hafif deployment profilleri sunmaktadır.

## 11. ÜRETİM ÇERÇEVELERİ VE UYGULAMA ARAÇLARI

Büyük dil modellerinin (LLM'ler) pratik uygulamalarında $\mathcal{O}(n^2)$ attention engeli, inference latency'sini ve kaynak kullanımını artırmakta; DeepSpeed Inference, vLLM ve FasterTransformer gibi üretim çerçeveleri, multi-GPU paralelliği, heterojen bellek yönetimi ve kernel optimizasyonları ile throughput'u yaklaşık 1.5-4 kat yükseltmektedir (Aminabadi et al., 2022, s. 7; Kwon et al., 2023, s. 8; NVIDIA, 2023, s. 8). DeepSpeed Inference, tensor parallelism ve pipeline scheduling ile katmanları çoklu GPU üzerine bölerek büyük ölçekli Transformer modellerinde, özellikle MT-NLG benzeri 530B parametreli yapılarda, latency'de 7.3 kata kadar azalma ve throughput'ta 1.5 katın üzerinde artış sağlamakta; MoE desteği ve Azure entegrasyonu ile enerji verimliliğini yaklaşık %50 artırırken, multi-node iletişim overhead'ı ve küçük modellerde sınırlı kazanç dezavantaj oluşturmaktadır (Aminabadi et al., 2022, s. 7). vLLM, PagedAttention temelli KV-cache sayfalama

tasarımıyla bellek parçalanmasını azaltarak değişken uzunluklu isteklerde throughput'u 2-4 kat artırmakta ve Llama 70B için Hugging Face tabanlı yaklaşımlara kıyasla yaklaşık %55 daha düşük latency sunmaktadır (Kwon et al., 2023, s. 8). FasterTransformer ise kernel fusion ve nicelleştirme ile, Ampere nesli GPU'larda GPT-J (6B) için PyTorch'a göre 5 kata varan hızlanma sağlamakta, ancak NVIDIA ekosistemine sıkı bağımlılık portabiliteyi sınırlamaktadır (NVIDIA, 2023, s. 8). Bu çerçeveler, SSM tabanlı alternatif mimarilerle bütünleştiğinde, büyük ölçekli LLM servislerinin maliyet ve latency açısından sürdürülebilir deployment'ı için temel bileşenler haline gelmektedir (Gu & Dao, 2023, s. 6).

## 12. SONUÇ VE GELECEK YÖNELİMLER

Bu kitap bölümü, büyük dil modellerinin (LLM'ler) çekirdek zorluklarından biri olan $\mathcal{O}(n^2)$ attention engelini bütüncül bir çerçevede ele almış, performans optimizasyonlarını mimari, sistem ve uygulama katmanları boyunca sentezlemiştir. Girişten (Bölüm 1) başlayarak temel attention mekanizmaları (Bölüm 2), donanım odaklı optimizasyonlar (Bölüm 3), seyrek ve lineer attention (Bölüm 4), çoklu ve gruplu sorgu attention (Bölüm 5), spekülatif kod çözme (Bölüm 6), KV-cache ve batching stratejileri (Bölüm 7), nicelleştirme (Bölüm 8), konum kodlama ve uzun bağlam (Bölüm 9), $\mathcal{O}(n)$ - $\mathcal{O}(1)$ karmaşıklıklı alternatif mimariler (Bölüm 10) ve üretim çerçeveleri (Bölüm 11) tartışılmış; bu yaklaşımlar birlikte inference hızını yaklaşık 2-37 kat artırmakta, bellek kullanımını %50-80 azaltmakta ve LLM'lerin pratik kullanımını dönüştürmektedir (Vaswani et al., 2017, s. 1; Gu & Dao, 2023, s. 6; Kwon et al., 2023, s. 8; Dao et al., 2022, s. 2; Leviathan et al., 2023, s. 3; Dao & Gu, 2024, s. 7). Ana çıkarımlar üç katmanda özetlenebilir. (1) Mimari düzeyde, seyrek ve lineer attention (Longformer, Performer) $\mathcal{O}(n)$

karmaşıklık sunarak uzun bağlamlı görevlerde (n>16K) perplexity'yi %15-25 azaltmakta (Beltagy et al., 2020, s. 3; Choromanski et al., 2020, s. 3); konum kodlama ve interpolation (RoPE, ALiBi, Position Interpolation) ile birleştiğinde uzun bağlam extrapolation'ı iyileşmektedir (Su et al., 2021, s. 6; Press et al., 2022, s. 6; Chen et al., 2023, s. 6). (2) Sistem düzeyinde, PagedAttention ve continuous batching, KV-cache ve zamanlama optimizasyonlarıyla throughput'u 20-40 kat artırmakta, nicelleştirme (LLM.int8(), GPTQ, AWQ, QLORA) ile model boyutunu 4-8 kat küçültmekte ve enerji maliyetlerini %50-70 azaltmaktadır (Yu et al., 2022, s. 4; Frantar et al., 2022, s. 5; Dettmers et al., 2023, s. 5; Lin et al., 2023, s. 5; Dettmers et al., 2022, s. 5; Xiao et al., 2023, s. 5; Kwon et al., 2023, s. 4). (3) Uygulama düzeyinde, DeepSpeed, vLLM ve FasterTransformer gibi çerçeveler, spekülatif decoding (Medusa, EAGLE) ile entegre edildiğinde gerçek zamanlı üretimi 3-5 kat hızlandırarak endüstriyel ölçekte deployment'ı mümkün kılmaktadır (Aminabadi et al., 2022, s. 7; Kwon et al., 2023, s. 8; NVIDIA, 2023, s. 8; Cai et al., 2024, s. 4; Li et al., 2024a, s. 4). SSM-tabanlı mimariler (Mamba, RetNet, RWKV) ve SSM-Transformer hibritleri ise $\mathcal{O}(n^2)$ engelini yapısal düzeyde gevşeterek $\mathcal{O}(n)$ - $\mathcal{O}(1)$ rejimlerine yaklaşmaktadır (Gu & Dao, 2023, s. 6; Sun et al., 2023, s. 7; Peng et al., 2023, s. 7).

Bununla birlikte, donanım bağımlılığı (GPU-özgü kernel'ler), düşük bit nicelleştirmede quantization noise, aşırı uzun bağlamlarda interpolation bozulmaları ve alternatif mimarilerde eğitim paralelliği sınırlamaları, $\mathcal{O}(n^2)$ engelinin yalnızca yönetilebilir hale geldiğini göstermektedir (NVIDIA, 2023, s. 8; Lin et al., 2023, s. 5; Chen et al., 2023, s. 6; Peng et al., 2023, s. 7). Gelecek çalışmalar için hibrit ve modüler SSM-attention tasarımları, bağlama duyarlı adaptif optimizasyonlar, sürdürülebilir ve enerji-aware nicelleştirme, multimodal uzun bağlam yöntemleri ve federated/privasi odaklı çerçeveler

(HELM, BigBench ve benzeri benchmark'ların genişletilmesiyle) ön plana çıkmaktadır. Sonuç olarak, bu bölüm, $\mathcal{O}(n^2)$ attention engelinin yenilikçi mimari, sistemsel ve üretim odaklı optimizasyonlarla büyük ölçüde aşılabileceğini göstererek, daha verimli, ölçeklenebilir ve etik LLM tasarımları için kavramsal bir temel sunmaktadır.

## KAYNAKÇA

Ainslie, J., Lee-Thorp, J., de Jong, M., Zemlyanskiy, Y., Lebrón, D., & Sanghai, S. (2023). GQA: Training generalized multi-query transformer models from multi-head checkpoints. *arXiv preprint arXiv:2305.13245*. Erişim adresi: https://arxiv.org/abs/2305.13245

Aminabadi, R. Y., Rajbhandari, S., Awan, A. A., Smith, C., Ly, A., Momcilovic, B., ... & He, Y. (2022). DeepSpeed inference: Enabling efficient inference of transformer models at unprecedented scale (Microsoft Technical Report MSR-TR-2022-21). Microsoft Research. Erişim adresi: https://arxiv.org/abs/2207.00032

Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*. Erişim adresi: https://arxiv.org/abs/2004.05150

Cai, T., Li, Y., Geng, Z., Peng, H., Lee, J. D., Chen, D., & Dao, T. (2024). Medusa: Simple LLM inference acceleration framework with multiple decoding heads. *arXiv preprint arXiv:2401.10774*. Erişim adresi: https://arxiv.org/abs/2401.10774

Chen, S., Wong, S., Chen, L., & Tian, Y. (2023). Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*. Erişim adresi: https://arxiv.org/abs/2306.15595

Choromanski, K., Likhosherstov, V., Dohan, D., Song, X., Gane, A., Sarlos, T., ... & Davis, A. (2020). Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*. Erişim adresi: https://arxiv.org/abs/2009.14794

Dao, T., Fu, D. Y., Ermon, S., Rudra, A., & Ré, C. (2022). FlashAttention: Fast and memory-efficient exact attention

with IO-awareness. *arXiv preprint arXiv:2205.14135*. Erişim adresi: https://arxiv.org/abs/2205.14135

Dao, T., & Gu, A. (2024). Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. *arXiv preprint arXiv:2405.21060*. Erişim adresi: https://arxiv.org/abs/2405.21060

Dettmers, T., Lewis, M., Belkada, Y., & Zettlemoyer, L. (2022). LLM.int8(): 8-bit matrix multiplication for transformers at scale. *arXiv preprint arXiv:2208.07339*. Erişim adresi: https://arxiv.org/abs/2208.07339

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLORA: Efficient finetuning of quantized LLMs. *arXiv preprint arXiv:2305.14314*. Erişim adresi: https://arxiv.org/abs/2305.14314

Frantar, E., Ashkboos, S., Hoefler, T., & Alistarh, D. (2022). GPTQ: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*. Erişim adresi: https://arxiv.org/abs/2210.17323

Gu, A., & Dao, T. (2023). Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*. Erişim adresi: https://arxiv.org/abs/2312.00752

Kitaev, N., Kaiser, Ł., & Levskaya, A. (2020). Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451*. Erişim adresi: https://arxiv.org/abs/2001.04451

Kwon, W., Li, Z., Zhuang, S., Sheng, L., Zheng, S., Yu, C. Y., ... & Stoica, I. (2023). Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems*

*Principles* (pp. 611-626). Erişim adresi: https://arxiv.org/abs/2309.06180

Leviathan, Y., Kalman, M., & Matias, Y. (2023). Fast inference from transformers via speculative decoding. *arXiv preprint arXiv:2211.17192*. Erişim adresi: https://arxiv.org/abs/2211.17192

Li, Y., Wei, F., Zhang, C., & Zhang, H. (2024a). EAGLE: Speculative sampling requires rethinking feature uncertainty. *arXiv preprint arXiv:2401.15077*. Erişim adresi: https://arxiv.org/abs/2401.15077

Li, Y., Wei, F., Zhang, C., & Zhang, H. (2024b). EAGLE-2: Faster inference of language models with dynamic draft trees. *arXiv preprint arXiv:2406.16858*. Erişim adresi: https://arxiv.org/abs/2406.16858

Lin, J., Tang, J., Tang, H., Yang, S., Dang, X., & Han, S. (2023). AWQ: Activation-aware weight quantization for LLM compression and acceleration. *arXiv preprint arXiv:2306.00978*. Erişim adresi: https://arxiv.org/abs/2306.00978

NVIDIA. (2023). FasterTransformer: NVIDIA's library for accelerating transformer model inference. NVIDIA Developer Documentation. Erişim adresi: https://github.com/NVIDIA/FasterTransformer

Peng, B., Alcaide, E., Anthony, Q., Albalak, A., Arcadinho, S., Cao, H., ... & Zhu, X. (2023). RWKV: Reinventing RNNs for the transformer era. *arXiv preprint arXiv:2305.13048*. Erişim adresi: https://arxiv.org/abs/2305.13048

Press, O., Smith, N. A., & Lewis, M. (2022). Train short, test long: Attention with linear biases enables input length extrapolation. *arXiv preprint arXiv:2108.12409*. Erişim adresi: https://arxiv.org/abs/2108.12409

Shah, J., Bikshandi, G., Zhang, Y., Thakkar, V., Ramani, P., & Dao, T. (2024). FlashAttention-3: Fast and accurate attention with asynchrony and low-precision. *arXiv preprint arXiv:2407.08608*. Erişim adresi: https://arxiv.org/abs/2407.08608

Su, J., Lu, Y., Pan, S., Wen, B., & Liu, Y. (2021). RoFormer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*. Erişim adresi: https://arxiv.org/abs/2104.09864

Sun, Y., Dong, L., Huang, S., Ma, S., Xia, Y., Xue, J., ... & Wei, F. (2023). Retentive network: A successor to transformer for large language models. *arXiv preprint arXiv:2307.08621*. Erişim adresi: https://arxiv.org/abs/2307.08621

Tay, Y., Dehghani, M., Bahri, D., & Metzler, D. (2020). Efficient transformers: A survey. *arXiv preprint arXiv:2009.06732*. Erişim adresi: https://arxiv.org/abs/2009.06732

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (pp. 5998-6008). Erişim adresi: https://arxiv.org/abs/1706.03762

Xiao, G., Lin, J., Seznec, M., Wu, J., Demouth, J., & Han, S. (2023). SmoothQuant: Accurate and efficient post-training quantization for large language models. *arXiv preprint arXiv:2211.10438*. Erişim adresi: https://arxiv.org/abs/2211.10438

Yu, G.-I., Jeong, J. S., Kim, G.-W., Kim, S., & Chun, B.-G. (2022). Orca: A distributed serving system for transformer-based generative models. In *Proceedings of the 16th USENIX Symposium on Operating Systems*

*Design and Implementation* (pp. 521-538). Erişim adresi: https://www.usenix.org/conference/osdi22/presentation/yu

Zaheer, M., Guruganesh, G., Dubey, K. A., Ainslie, J., Alberti, C., Ontanon, S., ... & Ahmed, A. (2020). Big bird: Transformers for longer sequences. *arXiv preprint arXiv:2007.14062*. Erişim adresi: https://arxiv.org/abs/2007.14062

# TÜRKİYE VE DÜNYADA

# YAZILIM MÜHENDİSLİĞİ

yaz
yayınları

YAZ Yayınları
M.İhtisas OSB Mah. 4A Cad. No:3/3
İscehisar / AFYONKARAHİSAR
Tel : (0 531) 880 92 99
yazyayinlari@gmail.com • www.yazyayinlari.com