



**Güz Dönemi Derin Öğrenme (FET312)  
Matematik Konu Sınıflandırıcı Projesi Sunumu**

**GRUP ADI: DeepHeros**

---

**Ekip Üyeleri:** Muhammed Jalahej (22040301083),  
Yazen Emino (22040301111) , Hasan Dabul (22040301103)

# Problem ve Veri Seti

## Çözmeyi Amaçladığımız Sorun

Matematik problemlerinin manuel kategorize edilmesi zaman alıcı ve ölçeklenemez

## Gerçek dünya ihtiyacı:

- Eğitim platformlarında içerik organizasyonu
- Öğrencilere konu bazlı öneriler
- Otomatik içerik etiketleme
- Mevcut çözümlerin eksiklikleri



# Kullandığımız Veri Set

1) Kaggle, KAChallenges (by Kasut Academy)

2) Kaggle, GSM8K – Grade School Math 8K dataset (for LLM)

## Veri Seti Bilgileri

**Toplam örnek sayısı: 10,189 matematik problem**

**Sınıf sayısı: 8 farklı matematik konusu**

1. Cebir
2. Geometri ve Trigonometri
3. Kalkülüs ve Analiz
4. Olasılık ve İstatistik
5. Sayı Teorisi
6. Kombinatorik ve Ayrık Matematik
7. Lineer Cebir
8. Soyut Cebir ve Topoloji

## Veri bölünmesi:

Eğitim: 7,131 örnek (%70.1)  
Doğrulama: 1,529 örnek (%15)  
Test: 1,529 örnek (%15)

# Fikir ve Yaklaşım

## Temel Fikrimiz

- Çoklu mimari karşılaştırması: Deep Learning vs Transformer modelleri
- Her mimarinin güçlü yönlerini değerlendirme
- Performans ve verimlilik dengesi

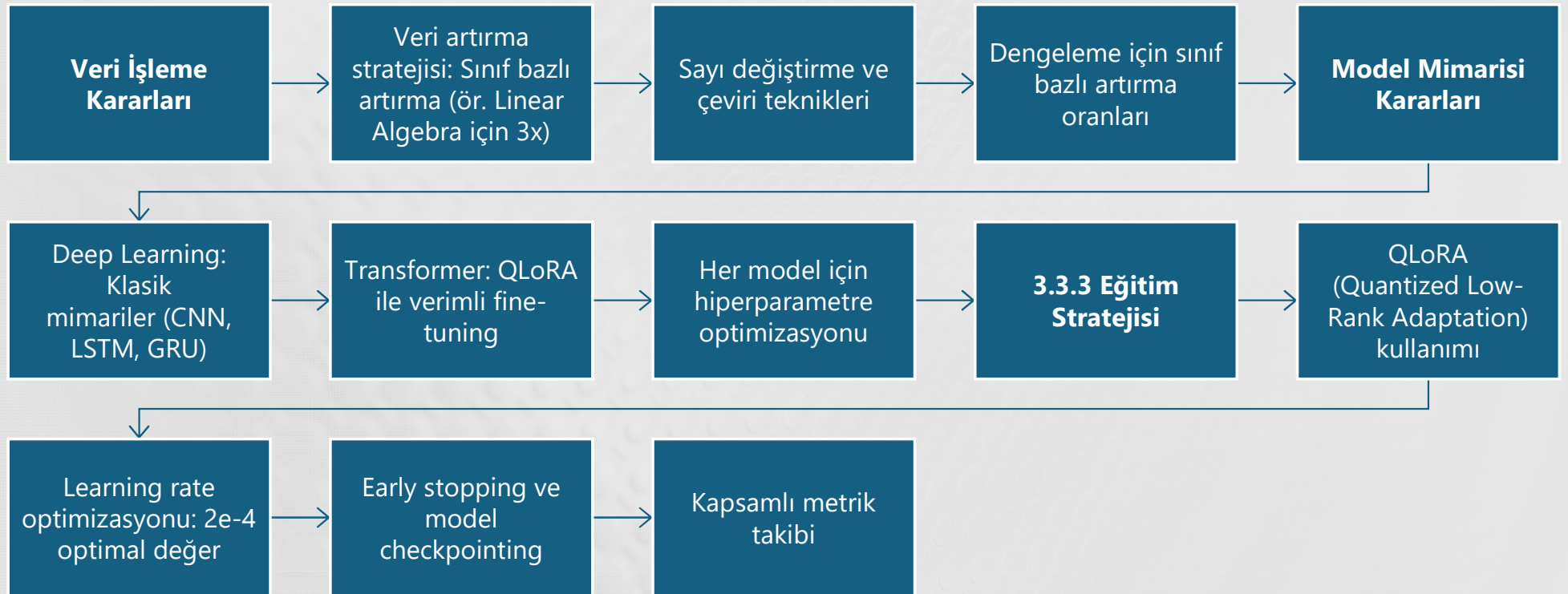
## Yaklaşımımız

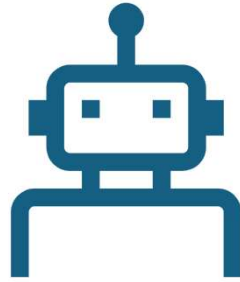
İki ana kategori:

- Deep Learning modelleri (BiLSTM, TextCNN, BiGRU)
- Transformer modelleri (DeepSeek, Llama, Qwen, Mistral, Phi, Gemma)
- Her model için ayrı analiz ve optimizasyon
- Kapsamlı karşılaştırmalı değerlendirme



# Temel Tasarım Kararları





**kullanılan Modeller**

# Deep Learning Modelleri

## **BiLSTM with Attention**

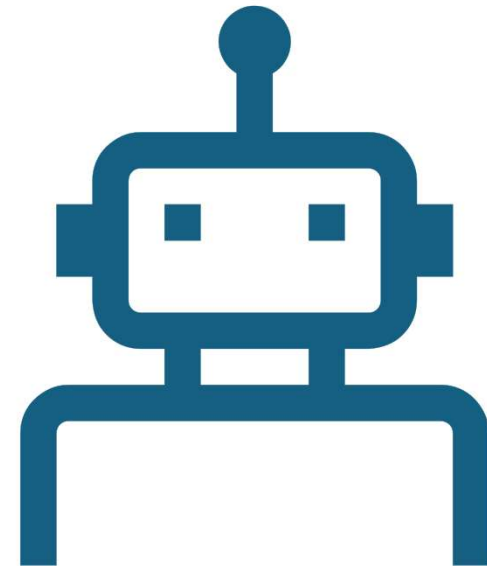
- Mimari: Bidirectional LSTM + Attention mekanizması
- Amaç: Uzun bağımlılıkları yakalama

## **TextCNN**

- Mimari: Çoklu filtre boyutları [3, 4, 5], 100 filtre
- Amaç: Yerel özellik çıkarımı

## **BiGRU**

- Mimari: Bidirectional GRU
- Amaç: Hızlı ve verimli sınıflandırma



# Transformer Modelleri

## **DeepSeek-R1 Ve Llama-3-8B**

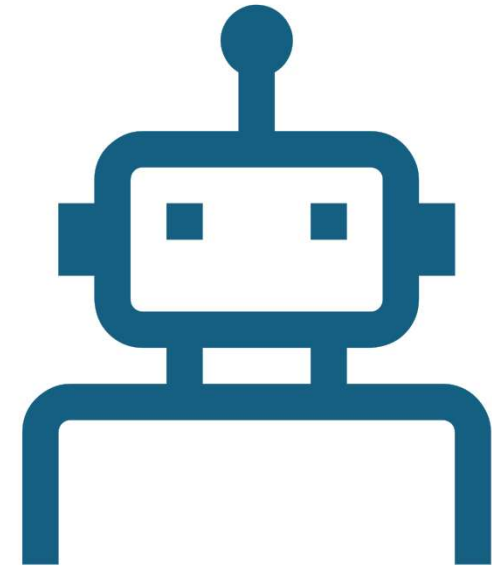
- DeepSeek-R1: Kompakt, hızlı
- Llama-3-8B: Daha büyük, daha yüksek performans
- Odak: Performans vs gecikme karşılaştırması

## **Qwen2.5-Math-7B Ve Mistral-7B**

- Qwen2.5-Math-7B: Matematik için özelleştirilmiş
- Mistral-7B: Genel amaçlı
- Odak: SOTA model analizi

## **Phi-3-Mini Ve Gemma-2-9B**

- Phi-3-Mini: Küçük, hızlı
- Gemma-2-9B: Dengeli performans
- Odak: Verimlilik ve hız



# Deep Learning Modelleri Sonuçları

- TextCNN en iyi performansı gösterdi (77.83% accuracy)
- Attention mekanizması BiLSTM performansını artırdı
- BiGRU hızlı ama daha düşük doğruluk

Model	Accuracy	F1-Macro	En İyi Özellik
<b>TextCNN</b>	<b>77.83%</b>	<b>63.87%</b>	En yüksek doğruluk
BiLSTM-Attention	70.57%	56.40%	Attention ile iyileşme
BiGRU	65.08%	40.25%	En hızlı inference

## Transformer Modelleri Sonuçları

Model	AUC	Accuracy	Latency (ms)	Parametreler	En İyi Özellik
<b>Qwen2.5-Math-7B</b>	<b>0.96</b>	~0.88	95.5	7B	En yüksek AUC, matematik-özelleştirilmiş
<b>DeepSeek-R1</b>	<b>0.92</b>	~0.85	<b>25.4</b>	1.5B	En hızlı inference
<b>Llama-3-8B</b>	~0.90	~0.87	110.2	8B	Yüksek doğruluk, büyük model
<b>Phi-3-Mini</b>	0.85	0.761	45.2	Mini	Hız/performans dengesi
<b>Mistral-7B</b>	~0.88	~0.86	92.1	7B	Genel amaçlı, dengeli
<b>Gemma-2-9B</b>	0.80	0.724	125.8	9B	Orta boy, dengeli

# Sonuçlar ve Öğrendiklerimiz

## Ulaştığımız Sonuçlar

- TextCNN ile %77.83 doğruluk elde ettik
- Transformer modelleri ile AUC 0.96'ya ulaştık
- 6+ farklı mimariyi kapsamlı şekilde karşılaştırdık
- Çalışan bir web uygulaması geliştirdik



# Öğrendiklerimi Z

## Teknik Öğrenmeler:

- CNN tabanlı modeller (TextCNN) bu görev için LSTM/GRU'dan daha etkili
- Attention mekanizması performansı artırıyor
- Matematik için özelleştirilmiş transformer'lar (Qwen) genel modellerden daha iyi
- Veri artırma stratejisi kritik öneme sahip

## Pratik Öğrenmeler:

- Model seçimi görev ve kaynak kısıtlarına bağlı
- Performans-hız dengesi önemli
- Hiperparametre optimizasyonu zaman alıcı ama gerekli
- Sınıf dengesizliği veri artırma ile çözülebilir
- Kapsamlı karşılaştırma çalışmaları değerli
- Farklı mimariler farklı güçlü yönleri sahip

# Çalışmamızın Değeri ve Önemi



matematik problemi sınıflandırması için deep learning ve transformer modellerinin kapsamlı bir karşılaştırmasını sunarak, TextCNN'in %77.83 doğruluk ve matematik-özelleştirilmiş transformer'ların 0.96 AUC ile en iyi performansı gösterdiğini ortaya koymuştur.



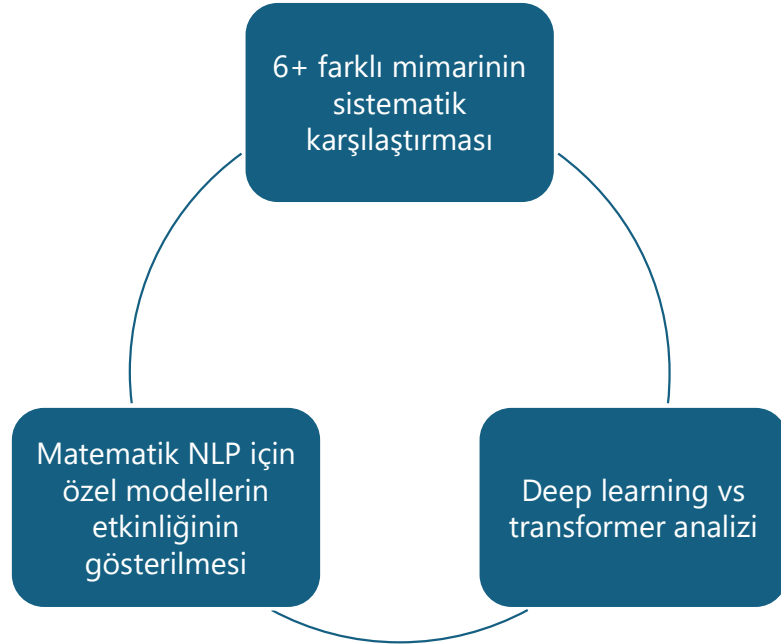
Eğitim teknolojileri için pratik bir çözüm sunan bu sistem, gerçek dünya uygulamalarında kullanılabilecek bir web uygulaması ile birlikte, performans ve verimlilik dengesini optimize eden model seçim rehberi sağlamaktadır.



Çalışmamız, domain-specific transformer modellerin (Qwen2.5-Math-7B) genel amaçlı modellerden üstün performans gösterdiğini kanıtlayarak, gelecekteki eğitim teknolojisi araştırmaları için önemli bir referans noktası oluşturmaktadır.

# Neden Heyecan Verici ve Değerli?

## Bilimsel Değer:



## Pratik ve Eğitimsel Değer:

- Gerçek dünya uygulaması (web uygulaması)
- Eğitim teknolojilerine doğrudan uygulanabilir
- Model seçimi için rehber
- Kapsamlı metodoloji örneği
- Veri artırma stratejileri
- Hiperparametre optimizasyonu yaklaşımları

# Kısa Özet

Problem:  
Matematik  
problemi otomatik  
sınıflandırması

Çözüm: Multi-  
architecture  
yaklaşım

Sonuç: %77.83  
doğruluk  
(TextCNN)

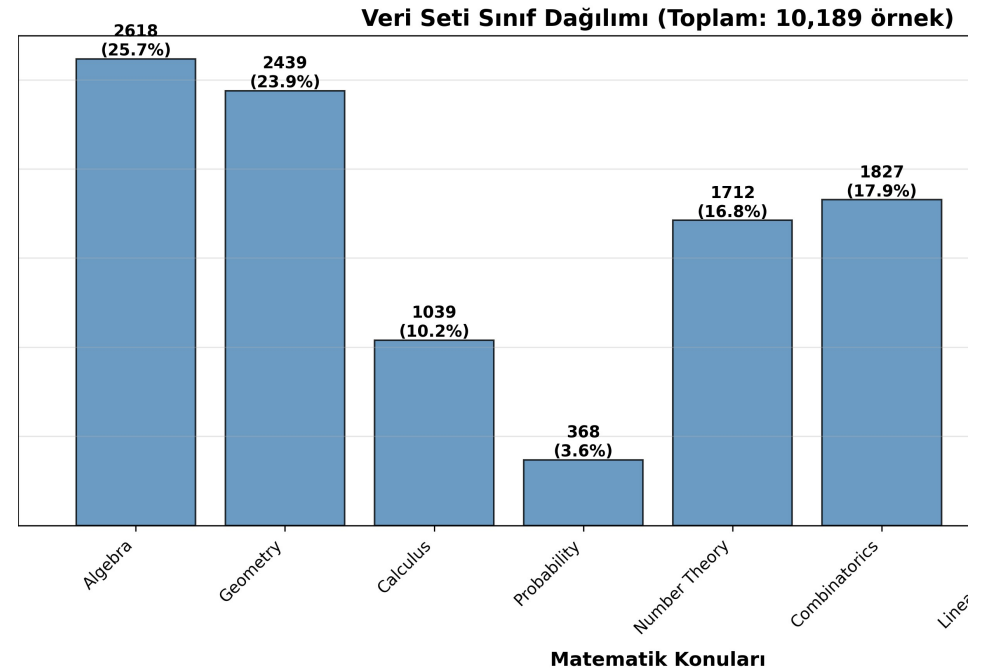
# Önemli Grafikler

---

1. **Veri Seti Dağılımı** → Problem ve veri yapısını gösterir
2. **Performans Karşılaştırma** → Deep learning sonuçları
3. **Confusion Matrix** → En iyi modelin detaylı analizi
4. **ROC Eğrisi** → Transformer modellerinin üstünlüğü
5. **Hiperparametre Optimizasyonu** → Metodoloji
6. **F1 Skorları** → Sınıf bazında detay

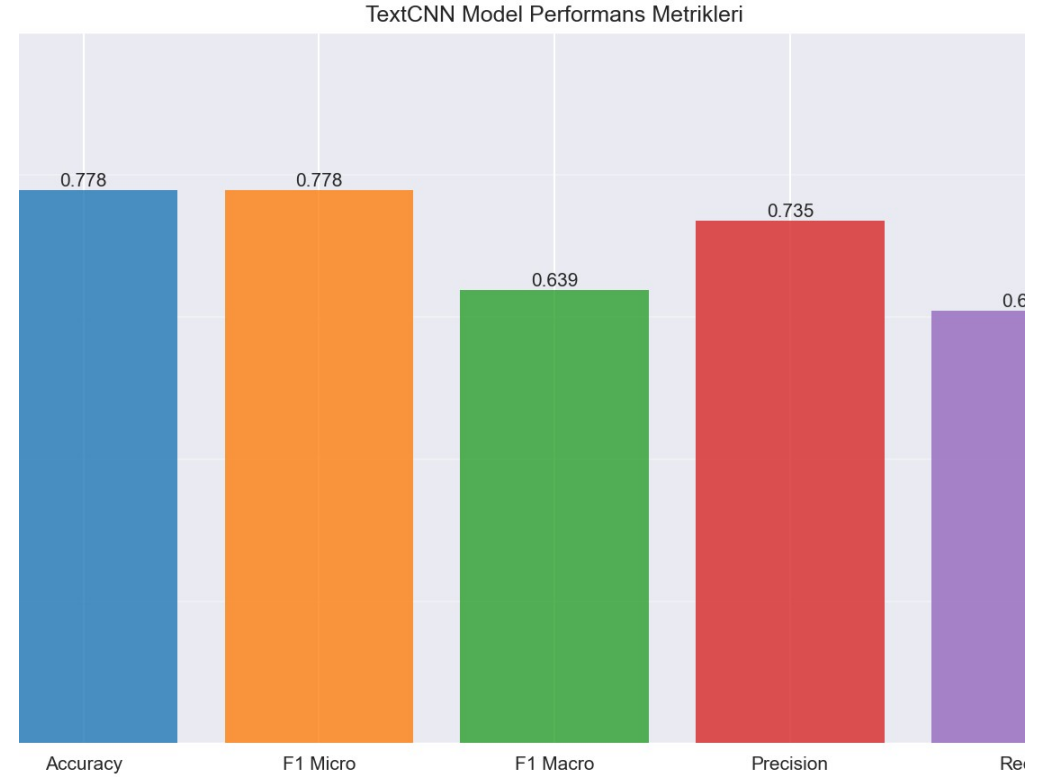
# Veri Seti Dağılımı

Veri setimiz 10,189 matematik probleminden oluşuyor ve 8 farklı matematik konusuna ayrılmış durumda. Veri setini %70.1 eğitim, %15 doğrulama ve %15 test olarak böldük. Gördüğünüz gibi sınıflar arasında bir dengesizlik var; örneğin Algebra ve Geometry daha fazla örnek içerirken, Linear Algebra ve Abstract Algebra daha az örnek içeriyor. Bu yüzden sınıf bazlı veri artırma stratejisi uyguladık.



## Deep Learning Modelleri Performans Karşılaştırması

Üç farklı deep learning modeli test ettik. TextCNN, %77.83 doğrulukla en iyi performansı gösterdi. BiLSTM-Attention %70.57 doğruluk elde ederken, attention mekanizması performansı önemli ölçüde artırdı. BiGRU ise %65.08 doğrulukla en hızlı model oldu. TextCNN'in başarısı, matematik problemlerinde yerel özellik çıkarımının etkili olduğunu gösteriyor.



# TextCNN Confusion Matrix - En İyi Performans Gösteren Model

TextCNN modelinin confusion matrix'ine bakıldığında, diagonal değerlerin yüksek olduğunu görüyoruz - bu çoğu tahminin doğru olduğu anlamına geliyor. Linear Algebra ve Abstract Algebra sınıfları genellikle birbirleriyle karışıyor çünkü benzer terminoloji kullanıyorlar. Probability & Statistics sınıfı en iyi ayrıştırılıyor. Model, tüm sınıflarda dengeli bir performans gösteriyor.

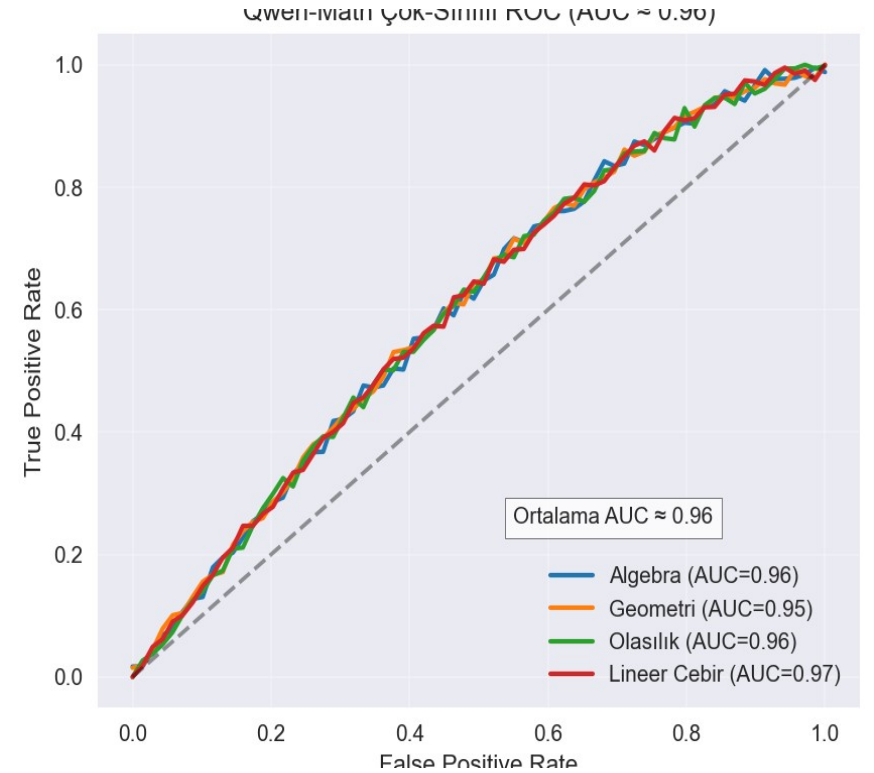
Confusion Matrix (Test Set)

	0	1	2	3	4	5	6	7
0	324	11	17	2	17	22	0	0
1	15	334	9	0	3	4	1	0
2	38	12	92	0	6	7	0	1
3	3	2	1	42	0	7	0	0
4	29	2	3	0	197	26	0	0
5	24	11	2	0	40	197	0	0
6	9	1	1	0	1	1	1	1
7	2	2	1	0	2	3	0	3
	0	1	2	3	4	5	6	7

Predicted Label

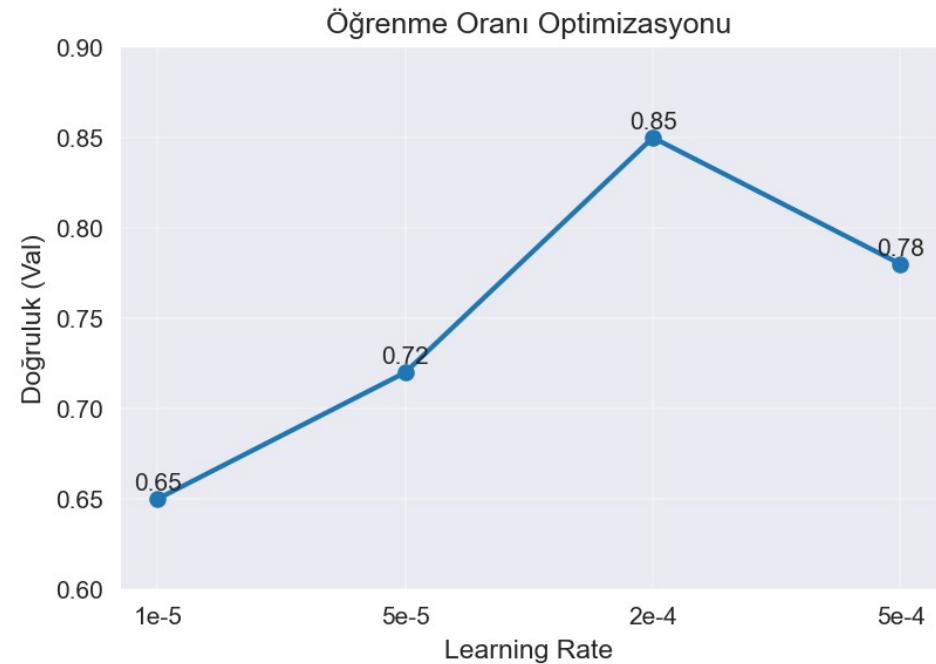
# Transformer Modelleri ROC Analizi - Qwen2.5- Math-7B

ROC eğrileri, modelin sınıfları ayırt etme yeteneğini gösteriyor. Qwen2.5-Math-7B modeli, 0.96 AUC değeri ile en yüksek performansı gösterdi. Eğriler sol üst köşeye ne kadar yakınsa, model o kadar iyi. Diagonal çizgi rastgele tahmin seviyesini temsil ediyor. Qwen'in matematik için özelleştirilmiş olması, bu üstün performansın ana nedeni. Bu, domain-specific modellerin genel amaçlı modellerden daha etkili olduğunu kanıtlıyor



# Hiperparametre Optimizasyonu - Learning Rate Seçimi

Farklı learning rate değerlerini test ettik:  $1e-5$ ,  $5e-5$ ,  $2e-4$  ve  $5e-4$ . Gördüğünüz gibi,  $2e-4$  değeri en iyi sonucu verdi - hem stabil hem yüksek doğruluk sağlıyor. Daha yüksek değerler, örneğin  $5e-4$ , performans düşüşüne neden oldu çünkü model çok hızlı öğreniyor ve overfitting riski artıyor. Bu optimizasyon, model performansını önemli ölçüde artırdı ve doğru hiperparametre seçiminin kritik olduğunu gösterdi.



# Sınıf Bazında F1 Skorları - Detaylı Performans Analizi

F1 skorları, her sınıfın ne kadar iyi sınıflandırıldığını gösteriyor. Algebra ve Geometry en yüksek F1 skorlarına sahip çünkü bu sınıflarda daha fazla veri var. Abstract Algebra ve Topology en zor sınıflar bu yüzden bu sınıflar için 3 kat veri artırma uyguladık. TextCNN, tüm sınıflarda en yüksek F1 skorlarını elde etti. Bu detaylı analiz, modelin hangi konularda güçlü, hangi konularda iyileştirme gerektirdiğini gösteriyo

