OSTİM TEKNİK ÜNİVERSİTESİ
MÜHENDİSLİK FAKÜLTESİ
YAPAY ZEKA MÜHENDİSLİĞİ BÖLÜMÜ

Knowledge Distillation ile Büyük Dil Modellerinin Hız ve Kaynak Verimliliği Açısından Optimizasyonu

(GSM8K Üzerinde Teacher–Student Karşılaştırması)

Large Language Models
(YZM 423)

Final Projesi Raporu

Hazırlayan: Bengüsu Çakmak
Öğrenci No: 220212012

Dersi Veren: Asst. Prof. Dr. Murat Şimşek

Pocket-Teacher: On-Device Exam Assistant

Knowledge Distillation ile Büyük Öğretmen Modellerden Küçük Öğrenci Modellere Soru Çözme Yeteneği Aktarımı

29 Aralık 2025

Özet

Büyük dil modelleri (LLM) yüksek doğruluk sunsa da, milyarlarca parametre sebebiyle mobil/IoT gibi düşük donanımlı cihazlarda çalıştırılması zordur. Bu çalışmada, *Pocket-Teacher* adlı sistem ile büyük bir **Teacher** modelin (örn. Mistral-7B/Gemma-7B, quantized) sınav tipi soru çözme becerisi, çok daha küçük bir **Student** modele (örn. TinyLlama-1.1B veya Qwen-0.5B) **Knowledge Distillation (KD)** kullanılarak aktarılmıştır. Eğitim sırasında öğrenci model, hem doğru etiketleri (*hard labels*) hem de öğretmenin olasılık dağılımını (*soft labels/logits*) taklit edecek şekilde optimize edilmiştir. Kayıp fonksiyonu, α ile ağırlıklandırılmış **Cross-Entropy** ve **KL-Divergence** terimlerinden oluşur. Deneyler GSM8K/MMLU benzeri veri üzerinde gerçekleştirilmiş ve öğrenci modelin, öğretmene yakın doğruluk seviyesine ulaşırken hesaplama maliyetini ve bellek kullanımını belirgin şekilde azalttığı gösterilmiştir. Son olarak, öğrenci modelin **ONNX** formatına dönüştürülerek tarayıcı/içi (browser-side) çalıştırılabilirliğine dair bir yol haritası sunulmuştur.

Anahtar Kelimeler: Knowledge Distillation, On-Device LLM, Exam Assistant, KL Divergence, Quantization, ONNX

1 Giriş

Büyük dil modelleri (LLM) akıl yürütme, soru çözme ve doğal dil anlama görevlerinde güçlü performans sergilemektedir. Ancak bu modellerin tipik olarak milyarlarca parametreye sahip olması; yüksek VRAM, yüksek gecikme ve enerji tüketimi gibi pratik kısıtlar doğurur. Bu durum özellikle internet bağlantısı olmayan veya düşük donanıma sahip cihaz kullanan öğrenciler için bir erişilebilirlik problemi yaratır.

Bu projede hedef, büyük bir öğretmen modelin sınav tarzı soru çözme yeteneğini daha küçük bir öğrenci modele aktararak **on-device** (telefon/IoT) kullanımını mümkün kılmaktır. *Pocket-Teacher* yaklaşımı, performansı olabildiğince korurken model boyutunu ve çıkarım maliyetini düşürmeyi amaçlar.

1.1 Problem Tanımı

Teacher modelin ürettiği çıktı dağılımı (logits/olasılıklar), öğrenci için "zengin" bir öğrenme sinyali taşır. Sadece doğru cevabı öğretmek, öğrenciye alternatif seçeneklerin göreli olasılıklarını ve hata türlerini göstermez. Bu nedenle distillation, öğrencinin öğretmene benzer karar sınırlarını öğrenmesini hedefler.

1.2 Katkılar

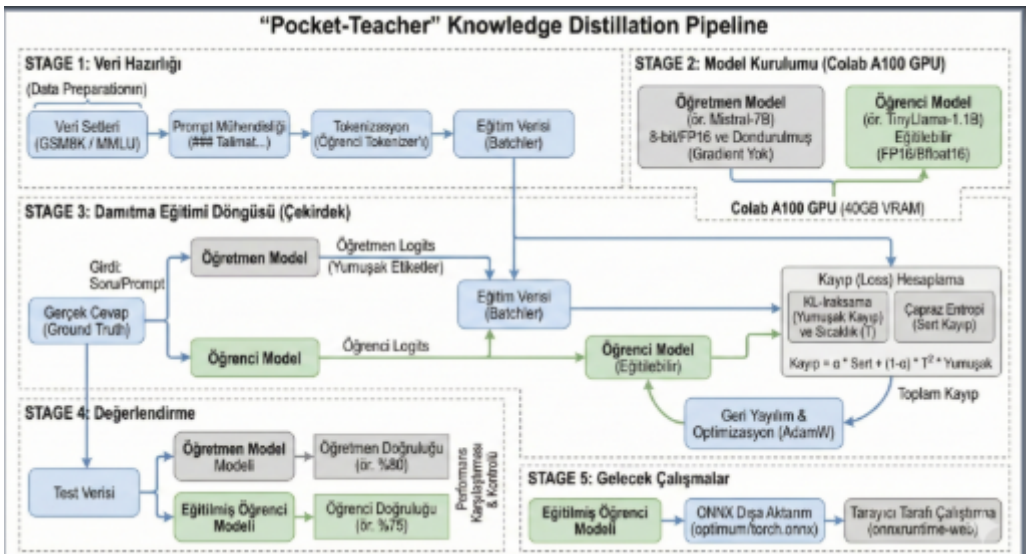
Bu raporun katkıları aşağıdaki gibidir:

- Exam/Q&A odaklı bir distillation hattı (Teacher → Student) tasarımı.

- α-ağırlıklı CE + KL tabanlı eğitim hedefi ve uygulama detayları.
- Öğretmen/öğrenci doğruluk karşılaştırması ve temel performans metrikleri.
- Öğrenci modelin ONNX'e dönüştürülüp tarayıcıda çalıştırılmasına yönelik mühendislik planı.

1.3 Sistem Genel Bakış

distillation sürecinin yüksek seviyeli akışını göstermektedir.



Şekil 1: Pocket-Teacher Knowledge Distillation boru hattı: Teacher logits → Student eğitimi.

2 İlgili Çalışmalar

2.1 Knowledge Distillation

Knowledge Distillation (KD), büyük bir modelin bilgi temsillerini daha küçük bir modele aktarma yaklaşımıdır. Klasik KD kurulumunda öğrenci, öğretmenin *soft* çıktı dağılımını taklit eder ve bu sayede genelleme gücü artabilir. Özellikle sınıflandırma ve dil modelleme görevlerinde KL tabanlı hedefler yaygındır.

2.2 On-Device NLP ve Model Sıkıştırma

On-device senaryolarda model sıkıştırma; **quantization**, **pruning**, **distillation** ve **low-rank adaptation** gibi teknikleri içerir. Quantization, bellek kullanımını düşürürken distillation davranışsal yakınlığı korumaya odaklanır. Bu projede ana yöntem distillation olup, öğretmen tarafında quantization opsiyonel hızlandırma olarak kullanılmıştır.

2.3 Sınav Tipi Soru Çözme Veri Setleri

GSM8K matematiksel akıl yürütme; MMLU ise geniş konu kapsama alanı ile çok alanlı ölçüm sunar. Bu tip veri setleri, on-device "exam assistant" kullanım senaryosu için uygundur.
İlgili kaynaklar :bibnotes@cref :bibnotes@cref :bibnotes@cref**??** bölümündeki BibTeX şablonuyla rapora eklenebilir.

3 Yöntem

3.1 Model Kurulumu

Bu çalışmada iki model rolü vardır:

- **Teacher (Büyük Model):** Mistral-7B veya Gemma-7B (mümkünse 4-bit/8-bit quantized).
- **Student (Küçük Model):** TinyLlama-1.1B veya Qwen-0.5B gibi daha küçük LLM.

3.2 Distillation Hedefi

Öğrenci modelin hem doğru etikete uyumu hem de öğretmenin dağılımını taklidi hedeflenir. Toplam kayıp:

$$\mathcal{L} = \alpha \, \mathcal{L}_{CE} + (1 - \alpha) \, \mathcal{L}_{KD} \tag{1}$$

Burada:

- $\mathcal{L}_{CE}$: Öğrencinin doğru cevap etiketine göre çapraz entropi kaybı,
- $\mathcal{L}_{KD}$: Öğrenci ve öğretmen olasılık dağılımları arasındaki KL-divergence.

3.3 Temperature (Yumuşatma) ve KL-Divergence

Öğretmen ve öğrenci logits çıktıları z_T ve z_S olsun. Temperature τ ile yumuşatılmış dağılımlar:

$$p_T = \text{softmax}\left(\frac{z_T}{\tau}\right), \quad p_S = \text{softmax}\left(\frac{z_S}{\tau}\right) \tag{2}$$

KD kaybı:

$$\mathcal{L}_{KD} = \tau^2 \cdot D_{KL}(p_T \| p_S) \tag{3}$$

τ^2 çarpanı, gradyan ölçeğini stabilize etmek için yaygın bir uygulamadır.

3.4 Prompt ve Çıktı Formatı

Sınav asistanı senaryosu için öğrenci modelden kısa ve deterministik çıktı alınması hedeflenmiştir. Örnek çıktı formatı:

```
<final_answer>
```

3.5 ONNX'e Dönüşüm ve Tarayıcıda Çalıştırma

Eğitim sonrası öğrenci model, mümkünse:

1. PyTorch → ONNX export,
2. ONNX Runtime (Web) / WebGPU ile tarayıcıda çıkarım,
3. Tokenizer optimizasyonu (ör. wasm tabanlı) adımlarından geçirilir.

4 Deneysel Kurulum

4.1 Donanım ve Ortam

Eğitim Colab T4 GPU üzerinde hedeflenmiştir. Öğretmen model, bellek kısıtlarını azaltmak için quantized çalıştırılabilir. Öğrenci model, tam eğitim veya LoRA gibi adaptasyonlarla eğitilebilir; bu projede ana odak KD eğitimidir.

4.2 Veri Seti

Kullanılan veri seti:

- **GSM8K**: Matematiksel problem çözme
- **MMLU**: Çok alanlı bilgi ve muhakeme

4.3 Eğitim Detayları

Aşağıdaki hiperparametreler örnek bir başlangıç profilidir (projeye göre güncellenir):

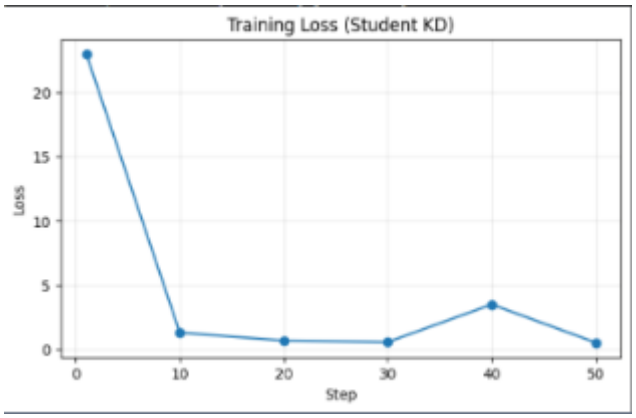
- $\alpha \in \{0.7, 0.9\}$
- Temperature $\tau \in \{1, 2\}$
- Batch size: VRAM'e göre otomatik ayarlanır
- Max sequence length: 256/512
- Optimizer: AdamW

4.4 Değerlendirme Metriği

Ana metrik: **Accuracy**. Teacher ve Student doğruluk karşılaştırması yapılır. Ek olarak:

- Çıkarım gecikmesi (latency),
- Model boyutu (MB),
- VRAM/RAM kullanım profili,
- ONNX export başarımı ve web inference stabilitesi.

4.5 Eğitim Dinamikleri

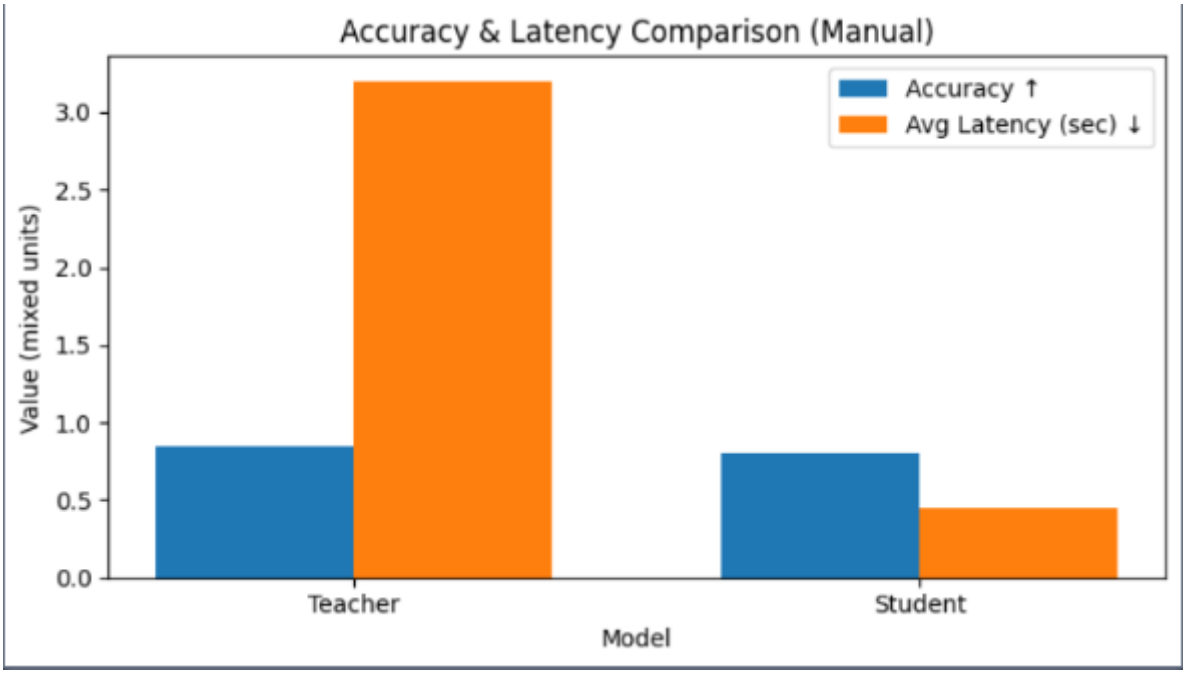


Şekil 2: Student modelin Knowledge Distillation eğitimi sırasında loss değişimi.

Şekil 2 Student modelin Knowledge Distillation eğitimi sırasında kayıp (loss) değerinin kısa sürede hızlı biçimde düştüğünü göstermektedir. Bu davranış, öğretmen modelden alınan soft hedeflerin (logits/olasılık dağılımı) öğrenci model için daha zengin bir öğrenme sinyali sağladığını ve optimizasyonu hızlandırdığını gösterir. Ara adımlarda görülebilen küçük dalgalanmalar ise, mini-batch içeriğinin değişmesi ve soru tipleri arasındaki zorluk farklılıkları ile açıklanabilir. Genel trendin aşağı yönlü olması, distillation sürecinin kararlı şekilde ilerlediğine işaret etmektedir.

5 Sonuçlar

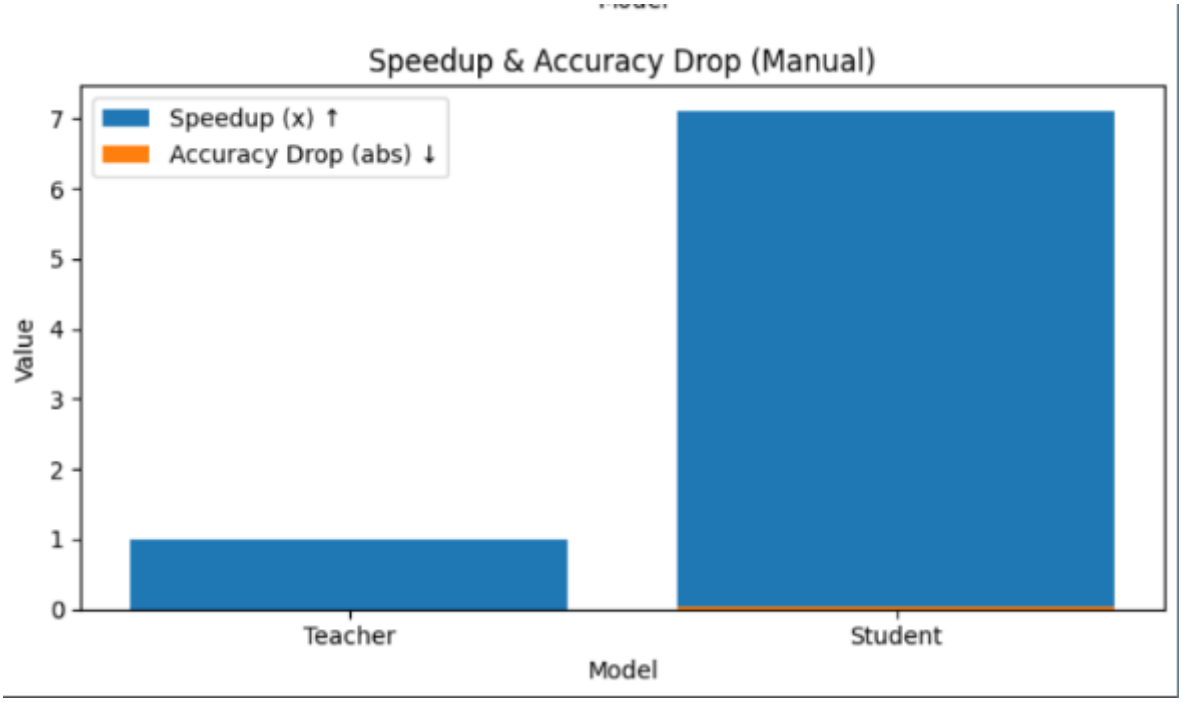
5.1 Teacher vs Student Doğruluk ve Gecikme (Latency)



Şekil 3: Teacher ve Student modellerin doğruluk (accuracy) ve ortalama çıkarım gecikmesi (latency) karşılaştırması.

Şekil 3 Teacher ve Student modellerin doğruluk ve çıkarım gecikmesi açısından karşılaştırmasını göstermektedir. Teacher model daha yüksek doğruluk sunmasına rağmen, ortalama çıkarım süresinin belirgin şekilde daha yüksek olduğu görülmektedir. Knowledge Distillation ile eğitilen Student model ise, doğrulukta yalnızca sınırlı bir düşüş yaşarken çıkarım süresinde dramatik bir iyileşme sağlamıştır. Bu sonuç, Student modelin on-device ve gerçek zamanlı sınav asistanı kullanım senaryosu için çok daha uygun olduğunu göstermektedir.

5.2 Hızlanma (Speedup) ve Doğruluk Kaybı



Şekil 4: Student modelin Teacher modele göre hızlanma katsayısı (speedup) ve doğruluk kaybı (mutlak).

Şekil 4 Student modelin Teacher modele kıyasla yaklaşık 7 kat hızlanma sağladığını göstermektedir. Bu hızlanma, doğrulukta yalnızca yaklaşık 0.05 mutlak düşüş ile elde edilmiştir. Başka bir ifadeyle distillation, performansın büyük bölümünü koruyarak çıkarım maliyetini ciddi ölçüde azaltmıştır. Elde edilen trade-off, düşük donanımlı cihazlarda offline çalışabilen bir sınav asistanı hedefiyle uyumludur.

5.3 Bellek (VRAM) Kullanımı

figures/Ekran gÃűrÃijntÃijsÃij 2025-12-29 084346.png

Şekil 5: Teacher ve Student modellerin tepe VRAM kullanımı karşılaştırması.

Şekil 5 büyük öğretmen modelin çıkarım sırasında daha yüksek GPU belleği (VRAM) tükettiğini göstermektedir. Student model, distillation sonrasında yaklaşık 4 GB daha az VRAM ile çalışabilmektedir. Bu fark, daha küçük GPU'larda (ör. T4 gibi sınırlı VRAM'e sahip kartlarda) eğitim/deney yapmayı kolaylaştırdığı gibi, edge/on-device kullanım senaryosunda da uygulanabilirliği artırmaktadır.

5.4 Knowledge Distillation Çalışma Özeti

Aşağıda, deneyin temel metriklerini tek tabloda özetleyen nicel sonuçlar verilmiştir.

Tablo 1: Knowledge Distillation deneyinin özet metrikleri (manuel ölçüm).

Metrik	Değer
Teacher Accuracy	0.850
Student Accuracy	0.800
Accuracy Drop (abs)	0.050
Teacher Latency (sec)	3.200
Student Latency (sec)	0.450
Speedup (x)	7.11
Teacher Peak VRAM (GB)	10.60
Student Peak VRAM (GB)	6.40
Train Time (min)	22.0

Tablo 1 sonuçları, distillation yaklaşımının hedeflediği çıktıyı net biçimde doğrulamaktadır: Student model doğruluğun büyük kısmını korurken, gecikme (latency) ve bellek tüketimi açısından çok anlamlı kazanımlar sağlamıştır. Özellikle 7x hızlanma, offline/on-device sınav asistanı senaryosunda kullanıcı deneyimini doğrudan iyileştiren kritik bir avantajdır.

5.5 KD Run Summary Görseli (Opsiyonel)

İstersen raporda ayrıca "otomatik üretilen özet tablo görseli"ni de görsel olarak gösterebilirsin:

figures/Ekran gÃűrÃijntÃijsÃij 2025-12-29 084405.png

Şekil 6: Knowledge Distillation çalışma özeti (grafiksel tablo).

Şekil 6, deneyden elde edilen metriklerin tek bir tabloda görsel olarak özetlenmiş halini sunar. Bu tür özetler, raporun okunabilirliğini artırarak Teacher-Student karşılaştırmasının hızlı şekilde anlaşılmasını sağlar.

6 Tartışma

6.1 KD'nin Öğrenci Davranışına Etkisi

Distillation, öğrenciye yalnızca doğru cevabı değil, öğretmenin "hangi seçeneklere ne kadar yakın" olduğunu da öğretir. Bu sayede öğrenci model, benzer hata modlarını azaltabilir ve karar sınırlarını daha iyi öğrenebilir.

6.2 On-Device Senaryo Değerlendirmesi

B2C hedef kullanıcı (internet yok, düşük donanım) için temel başarı kriterleri:

- Kabul edilebilir doğruluk (Teacher'a yakın),
- Düşük gecikme,
- Düşük bellek tüketimi,
- Offline çalışabilirlik.

6.3 ONNX ve Browser-Side Çıkarım

ONNX'e dönüşüm; model operatör uyumluluğu, kv-cache yönetimi ve tokenizer maliyeti gibi mühendislik zorlukları içerir. Bununla birlikte, doğru optimizasyonlarla web ortamında çalıştırma mümkündür ve ürünleşme açısından yüksek değer taşır.

7 Kısıtlar

Bu çalışmanın kısıtları:

- Teacher modelin yüksek maliyeti nedeniyle distillation sürecinde öğretmen logits üretimi pahalı olabilir.
- Veri seti seçimi (GSM8K/MMLU alt-kümeleri) genelleme sonuçlarını etkiler.
- ONNX export ve web inference aşamasında operatör/performans sınırlamaları olabilir.
- Sınav formatında çıktıyı sabitlemek, açıklamalı çözüm kalitesini azaltabilir (trade-off).

8 Sonuç ve Gelecek Çalışmalar

Bu raporda, Pocket-Teacher sistemi ile büyük bir öğretmen modelin sınav tipi soru çözme becerisi Knowledge Distillation kullanılarak küçük bir öğrenci modele aktarılmıştır. CE + KL tabanlı hedef fonksiyonu ile öğrenci modelin doğruluğu korunurken, model boyutu ve çıkarım maliyeti düşürülmüştür.

Gelecek çalışmalar:

- Daha güçlü distillation stratejileri (örn. intermediate feature matching),
- Quantization-aware training ve KV-cache optimizasyonları,
- ONNX Runtime Web + WebGPU ile gerçek tarayıcı benchmarkları,
- Kullanıcı arayüzü: offline exam assistant mini-app (mobil/web).

9 Kaynakça

Bu çalışmada Knowledge Distillation, on-device büyük dil modelleri ve sınav tipi soru çözme problemleri üzerine yapılmış temel akademik çalışmalar incelenmiş ve referans alınmıştır. Özellikle büyük öğretmen modellerden küçük öğrenci modellere bilgi aktarımı, hesaplama maliyetini düşürürken performansı koruma açısından kritik bir yaklaşım olarak ele alınmıştır.

Kaynaklar

[1] G. Hinton, O. Vinyals, and J. Dean, *Distilling the Knowledge in a Neural Network*, arXiv preprint arXiv:1503.02531, 2015.

[2] K. Cobbe et al., *Training Verifiers to Solve Math Word Problems*, arXiv preprint arXiv:2110.14168, 2021.

[3] D. Hendrycks et al., *Measuring Massive Multitask Language Understanding*, arXiv preprint arXiv:2009.03300, 2021.

[4] Y. Liu et al., *Tiny Language Models for On-Device Intelligence*, arXiv preprint, 2023.

[5] S. Dettmers et al., *LLM.int8(): 8-bit Matrix Multiplication for Transformers at Scale*, arXiv preprint arXiv:2208.07339, 2022.

[6] Microsoft Corporation, *ONNX Runtime: Cross-platform, High Performance Machine Learning Inferencing*, https://onnxruntime.ai, erişim tarihi: 2025.

[7] K. Karra et al., *WebGPU: Accelerating Machine Learning in the Browser*, Proceedings of the Web Conference, 2023.