modeller değerlendirmeye alınmıştır. Değerlendirmeye alınan modeller arasında, kullanıcı tarafından verilen talimatı yerine getirebilme, verilen soruya karşı bir cevap çıktısı üretme, eksik verilen cümlenin devamını getirebilme gibi çoklu metin dil kabiliyetlerini yerine getirebilme şartına bakılmıştır. Bu görevlerde bozuk ve anlamsız çıktı üreterek bariz şekilde başarısız ve yetersiz olarak ayırt edilebilen modeller karşılaştırmaya dahil edilmemiştir. Bu kapsamda, Türkçe Bert modelleri [6], [7] ve TURNA [8] GPT tabanlı olmadıkları ve MASK tabanlı oldukları için; Main ise sürümünün açık kaynaklı olmaması nedeniyle bu karşılaştırmaya dahil edilmemiştir.

B. Seçilen Modeller

Performansları karşılaştırılan dil modelleri Tablo I'de verilmiştir.

TABLO I: KARŞILAŞTIRILAN DİL MODELLERİ

Model	Parametre Sayısı	Yayınlanma Tarihi	TR Fine-Tuned	Base Model	Açıklamalar
Mistral-7B-Instruct-v0.2-turkish [9]	7.24 Milyar	05/01/2024	✓	Mistral-7B-Instruct-v0.2	SFT Training ve Freeze yöntemleri kullanılarak alpaca-gpt4-tr talimatlarına göre finetune edilmiş bir versiyondur.
Llama-2-7b-chat-turkish-instructions [10]	6.74 Milyar	11/08/2023	✓	Llama-2-7B-chat	Türkçe talimatlar veri kümesinde finetune edilmiş bir versiyondur.
Trendyol-LLM-7b-chat-v0.1 [11]	6.84 Milyar	07/02/2024	✓	Trendyol-LLM-7b-base-v0.1	Base modeli temel alınarak 180 binlik Türkçe talimat veri seti üzerinde LoRa kullanılarak finetune edilmiş bir versiyondur.
Trendyol-LLM-7b-base-v0.1 [12]	6.84 Milyar	07/02/2024	✓	Llama-2-7B	Optimize edilmiş bir transformer mimarisi kullanan autoregressive dil modelidir. LoRa kullanılarak 10 milyar token üzerinde finetune edilmiştir.
mGPT [13]	1.3 Milyar	15/04/2022	—	—	Wikipedia ve C4 Corpus kullanılarak 25 dil ailesinden dilbilimsel olarak çeşitli 61 dilde pretrain edilmiş bir multilingual dil modelidir.
Deepseek-llm-7b-chat [14]	7.0 Milyar	29/11/2023	—	deepseek-llm-7b-base	Base model kullanılarak Talimat veri kümesinde finetune edilmiş multilingual bir dil modelidir.
open chat_3.5 [15]	7.0 Milyar	20/09/2023	—	Mistral-7B-v0.1	Pekiştirmeli öğrenmeden ilham alan C-RLFT tekniği ile finetune edilmiş , multilingual bir dil modelidir.

III. KARŞILAŞTIRMA VERI KÜMELERI

Soru cevaplama çok farklı görevler için ortak bir format sağlamaktadır. Bu sebeple dil modellerinin karşılaştırılmasında en yaygın olarak kullanılan yöntemlerdendir. Seçilmiş talimat veri kümesi, çok farklı alanlardan soru içermektedir. Bu şekilde, her bir modelin bu sorulardaki cevaplarını hem otomatik değerlendirmelerle hem de oylama yolu ile karşılaştırılmaktadır. Model, aynı veri kümesiyle eğitilmediği sürece çıktıların cevaplarla örtüşmesi her zaman söz konusu değildir. Soru-cevap kümesinden örnekler Tablo II'de verilmiştir.

TABLO II: SORU-CEVAP KÜMESİ ÖRNEĞİ

Soru	Cevap
Bir elma ağacında 10 elma var. Bir rüzgar esiyor ve ağaçtan 2 elma düşüyor. Kaç elma ağaçta kalmış olur?	Ağaçta 8 elma kalmış olur.
Saç bakımı için üç basit yöntemi açıklayın.	Düzenli kesim, doğal ürünlerle nazik temizlik, derin nemlendirme tedavileri.
Sosyal ilişkileri geliştirmek ve yeni arkadaşlar edinmek için stratejiler sun.	İlgi alanlarına uygun etkinliklere katıl, güleryüzle iletişim kur, aktif dinleme pratiği yap, sosyal medyayı etkili kullan ve inisiyatif al.

Bağlamda öğrenme veri kümelerinde modelin eğitilmeden sadece örnekler göstererek cevaplama kabiliyeti ölçülmüştür. Veri kümesinden örnekler Tablo III'de verilmiştir.

TABLO III: BAĞLAMDA ÖĞRENME VERİ KÜMESİ ÖRNEĞİ

Soru	Seçenekler	Cevap
Güneş sistemindeki hangi gezegen en uzun yıla sahiptir?	"Güneş'e en yakın gezegen.", 'En uzun güne sahip gezegen.', 'En çok uydusu olan gezegen.', "Güneş'e en uzak gezegen."	Güneş'e en uzak gezegen.
Singapur'da sakız çiğnemek yasa dışı mı?	"Hayır, Singapur'da yalnızca sakız ithal etmek ve satmak yasa dışıdır.", "Evet, Singapur'da sakız çiğnemek yasaktır."	Hayır, Singapur'da yalnızca sakız ithal etmek ve satmak yasa dışıdır.

Bağlamda öğrenme performanslarının ölçümünde ARC [16], HellaSwag [17], TruthfulQA [18], MMLU [19] veri setleri yaygın olarak kullanılmaktadır. Bu veri setleri ile ilgili bilgiler Tablo IV'de verilmiştir. Veri setlerinin ingilizce olmasından dolayı her veri kümesinden örnekleme yapılıp Türkçeye çevrilmiştir.

Soru cevaplama veri kümesi ise açık kaynak olan bir Türkçe veri kümesinden [20] düzgün ve anlamlı 1000 tane örnek ayıklanarak oluşturulmuştur. Bu veri kümesinin modellerin eğitiminde kullanılmış olma ihtimali yüzünden ayrıca, 300 tane yeni soru-cevap ikilisi hazırlanıp bir veri kümesi daha oluşturulmuştur. Oluşturulan veri kümeleri, araştırmacılarla paylaşılacaktır.

TABLO IV: BAĞLAMDA ÖĞRENME VERİ KÜMELERİ

İsmi/Türü	Test	Train	Validation	dev	Shot NO.	Seçmeli
ARC	400	812	—	—	25	Evet
HellaSwag	—	891	641	—	10	Hayır
TruthfulQA	—	—	635	—	0	Evet
MMLU	662	—	—	30	5	Evet

IV. KARŞILAŞTIRMA ÖLÇÜTLERI

Bu çalışmada, Türkçe dil modellerinin performanslarının kapsamlı bir şekilde değerlendirilmesi amacıyla üç farklı ölçüt kullanılmıştır. Bu ölçütler, modellerin Türkçe dilini ne kadar iyi anladıklarını ve dili kullanma becerilerini ölçmek