1 Ekim – 1 Kasım 2025 Arası Yapay Zekâ Araştırmaları: Model Geliştirme, Reasoning ve Hizalama

1. Uzun Bağlam (Long-Context) Reasoning ve Model Mimarileri

1.1. Yeni Framework'ler ve Yöntemler

1.1.1. ToM: Tree-oriented MapReduce ile Hiyerarşik Reasoning

1 Ekim – 1 Kasım 2025 tarihleri arasında, uzun bağlam üzerinde reasoning (çıkarım) yeteneği konusunda önemli bir akademik katkı, 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP) konferansında sunulan **"ToM: Leveraging Tree-oriented MapReduce for Long-Context Reasoning in Large Language Models"** başlıklı makale olmuştur . Bu çalışma, büyük dil modellerinin (LLM) sınırlı bağlam penceresi nedeniyle uzun metinler üzerindeki çıkarım performansında yaşanan düşüşe çözüm getirmeyi hedeflemektedir. Mevcut **Retrieval-Augmented Generation (RAG)** ve **divide-and-conquer (böl-ve-fethet)** yaklaşımlarının, benzerlik temelli sıralamaya veya parçaların izole edilerek işlenmesine dayanması nedeniyle uzun menzilli bağımlılıkları yakalamakta zorlandığı ve mantıksal tutarlılığı riske attığı belirtilmektedir. **ToM (Tree-oriented MapReduce)** framework'ü, bu sınırlılıkları aşmak için uzun belgelerin doğal hiyerarşik yapısını (başlıklar ve alt başlıklar) kullanarak yenilikçi bir yaklaşım sunar. İlk olarak, **hiyerarşik anlamsal ayrıştırma (hierarchical semantic parsing)** yoluyla bir **DocTree (Document Tree)** oluşturur. Ardından, ağaç yapısı üzerinde alt seviyeden üst seviyeye doğru (bottom-up) bir veri toplama işlemi gerçekleştirir. Bu işlem, **MapReduce paradigmasına** dayanır: **"Map"** adımında, her alt düğümde (child node) gerekçeler (rationales) üretilir; **"Reduce"** adımında ise bu gerekçeler kardeş düğümler arasında birleştirilerek çatışmalar çözülür veya üst düğümde (parent node) bir fikir birliğine varılır. Bu yinelemeli (recursive) reasoning süreci, ToM'un hem yerel hem de uzun menzilli bağımlılıkları etkili bir şekilde yakalamasını sağlar. **70B+ parametreli LLM'ler** üzerinde yapılan deneyler, ToM'un mevcut böl-ve-fethet framework'lerine ve RAG yöntemlerine kıyasla önemli ölçüde daha iyi performans gösterdiğini ve daha üstün mantıksal tutarlılık ile uzun bağlam üzerindeki çıkarımı başardığını ortaya koymuştur .

1.1.2. ToTAL: Düşünce Şablonları ile Tekrar Kullanılabilir Reasoning

Ekim 2025'te yayımlanan **"When Thoughts Meet Facts: Reusable Reasoning for Long-Context LMs"** başlıklı akademik makalede, uzun bağlam dil modellerinde (LCLM) çoklu adımlı reasoning süreçlerini iyileştirmek için yeni bir framework olan **ToTAL (Thought**

Template Augmented LCLMs) tanıtılmıştır . Bu çalışma, LCLM'lerin yüz binlerce token'lık bağlam pencereleri sayesinde bilgi yoğunluğu yüksek görevlerde sunduğu yeni fırsatları ele alıyor. Ancak, bağlam penceresine daha fazla belge eklemenin tek başına yeterli olmadığını, çünkü modellerin kanıtları birbirine bağlamakta zorlanabileceğini vurguluyor. ToTAL framework'ü, bu boşluğu **"düşünce şablonları" (thought templates)** ile dolduruyor. Bu şablonlar, önceki problem çözme izlerinden türetilen ve kanıtların nasıl birleştirileceğini yapılandıran, tekrar kullanılabilir reasoning kalıplarıdır. Bu yaklasım, **"nasıl düşünüleceği"ni temsil eden şablonları (epistemik bilgi)** ve **"ne bilinmesi gerektiği"ni temsil eden belgeleri (fakt bilgisi)** ayırarak, modellerin karmaşık kanıtlar üzerinde kompozisyonel reasoning yapmasını sağlar. Ayrıca, bu şablonların etkinliğini korumak için, eğitim verilerinden türetilen şablonları doğal dil geri bildirimiyle yinelemeli olarak geliştiren bir güncelleme stratejisi önerilmektedir .

ToTAL'ın deneysel sonuçları, çeşitli benchmark'lar ve LCLM aileleri arasında güçlü temel modellere kıyasla hem retrieval-tabanlı hem de retrieval-dısı ayarlarda tutarlı iyileştirmeler sağladığını göstermiştir. Ayrıca, optimize edilmis şablonların daha küçük açık kaynak modellere damıtılabileceği ve bu sayede framework'ün genis uygulanabilirliği ve şeffaf reasoning yeniden kullanımı gösterilmiştir. Bu çalışma, LCLM'lerin sadece bağlam boyutunu ölçeklendirmekten ziyade, reasoning yeteneklerini güçlendirme yönünde önemli bir adım olarak değerlendirilmektedir. ToTAL'ın sunduğu düşünce şablonları, modellerin uzun ve karmaşık bağlamlarda daha yapılandırılmıs ve güvenilir çıkarımlar yapmasına olanak tanıyarak, bilgi yoğunluğu yüksek uygulamalarda (örneğin, kurumsal bilgi erisimi veya akademik araştırma) kullanılabilirliklerini artırma potansiyeline sahiptir .

1.1.3. JERR: Graf Tabanlı Reasoning ve Monte Carlo Tree Search

Ağustos 2025'te arXiv'de yayınlanan ve Kasım 2025'te EMNLP 2025 konferansında sunulan **"Joint Enhancement of Relational Reasoning for Long-Context LLMs"** başlıklı makale, **JERR (Joint Enhancement of Relational Reasoning)** adlı yeni bir framework'ü tanıtmaktadır . Bu çalışma, büyük dil modellerinin (LLM) uzun bağlamları anlamada ve karmaşık görevleri çözmede yaşadığı zorluklara odaklanmaktadır. Özellikle **bellek sınırlamaları** ve karmaşık uzun bağlam görevlerini çözme konusundaki yetersizlikler, LLM'lerin performansını sınırlayan temel faktörler olarak belirtilmiştir . Ayrıca, mevcut modellerin **şeffaflık eksikliği** ve **"hallucination" (hayal ürünü üretme)** eğilimi gibi sorunlar da bu araştırmanın odak noktaları arasındadır . JERR, bu zorlukları aşmak için **graf tabanlı çıkarımı (graph-based reasoning)** ve **Monte Carlo Tree Search (MCTS)** algoritmasını entegre eden üç adımlı bir yaklasım sunmaktadır .

JERR framework'ü üç temel bileşenden oluşmaktadır. İlk adım olan **Synopsis Extraction**, uzun bağlamın stratejik olarak parçalara ayrılması ve her bir parçanın özetlenmesiyle gerçekleştirilir. Bu süreçte, `autogen` paketi kullanılarak metin parçalara bölünür ve ardından özetleme için tasarlanmış özel komutlar (prompts) ile her parça bir özet (synopsis) haline getirilir . İkinci adım, **Graph Construction (Graf Oluşturma)** , bilgi atomları ve özelliklerinin çıkarımıyla başlar. Bu süreçte, hem tam eşleşme hem de benzerlik temelli yinelenen verilerin (deduplication) giderilmesi için Bloom Filter ve SimHash gibi teknikler kullanılır. Ardından, Shi ve ark. (2023) çalışmasından ilham alınarak, çıkarım sürecinde daha odaklı kesif sağlayan ve verimliliği artıran **Yönlü Açlık Grafiği (Directed Acyclic Graph – DAG)** yapısı inşa edilir . Üçüncü ve son adım olan **Graph-based Reasoning**, MCTS algoritmasının kullanıldığı çıkarım aşamasıdır. Bu aşamada, verilen bir soruyla en alakalı olan üst-k düğüm, MCTS kullanılarak belirlenir. Seçilen bu düğüm alt kümesi ve daha önce çıkarılan özet, modeli yönlendirmek için kullanılır. Bu sayede, model orijinal metinde hangi bölümlerin yeniden ziyaret edilmesi gerektiğini belirleyebilir ve son olarak, çıkarılan özet ile seçilen pasajların birleşimi kullanılarak bilgili bir yanıt üretilir .

JERR'in deneysel sonuçları, uzun bağlam görevlerinde mevcut temel yöntemlere kıyasla üstün performans gösterdiğini ortaya koymuştur. **QuALITY, MuSiQue ve NarrativeQA** gibi uzun bağlam değerlendirme veri setlerinde yapılan testlerde, JERR framework'ü ROUGE ve F1 metriklerinde tüm temel yöntemleri tutarlı bir şekilde geride bırakmış ve **LLM-Rater değerlendirmesinde en yüksek puanları** elde etmiştir . Ancak, yazarlar JERR'in bazı sınırlılıklarını da belirtmiştir. Bu sınırlılıklar arasında, framework'ün farklı bilgi alanlarındaki diğer çıkarım görevlerine ne ölçüde genellenebileceğinin belirsizliği (görev ölçeklenebilirliği), karmaşık bir boru hattı gerektiren ve büyük ölçekli, yüksek hassasiyetli grafik oluşturmayı zorlaştıran bilgi grafiği oluşturma süreci ve framework'ün daha basit görevlerdeki etkinliğinin sistematik olarak değerlendirilmemiş olması sayılabilir .

1.2. Retrieval-Augmented Generation (RAG) ve Hibrit Yaklaşımlar

1.2.1. RAG'ın Sınırlılıkları ve Mantıksal Tutarsızlık Problemi

ToM framework'ünü tanıtan makale, **Retrieval-Augmented Generation (RAG)** yaklaşımının uzun bağlam çıkarımındaki temel sınırlılıklarını detaylı bir şekilde ele almaktadır . RAG, uzun bağlam sorununu, bir retriever (alıcı) bileşeni kullanarak uzun belgeden en alakalı parçaları (chunks) belirlemek ve ardından bu parçalar üzerinden çıkarım yapmaya odaklanmak suretiyle çözmeyi amaçlar. Bu yaklaşım, modelin sadece en ilgili bilgiye odaklanmasını sağlayarak bellek sınırlamalarını aşmaya yardımcı olur.

Ancak, ToM makalesi, RAG'ın bu mekanizmasının önemli bir dezavantajını ortaya koymaktadır: **RAG, parçalar arasındaki mantıksal tutarlılığı (logical coherence) genellikle göz ardı eder** . Çünkü RAG, parçaları sıralarken çoğunlukla benzerlik (similarity) temelli sıralama algoritmalarına güvenir. Bu durum, birbirinden mantıksal olarak kopuk, ancak sorguyla yüksek benzerlik puanına sahip parçaların bir araya getirilmesine ve bu parçalar üzerinden tutarsız bir çıkarım yapılmasına yol açabilir.

Bu mantıksal tutarsızlık problemi, özellikle karmaşık ve çok adımlı çıkarım gerektiren görevlerde RAG'ın performansını sınırlar. Örneğin, bir belgedeki olayların sebep-sonuç ilişkilerini anlamak, karakterlerin motivasyonlarını takip etmek veya uzun bir hikâyenin ana temalarını sentezlemek gibi görevler, parçalar arasındaki bağlamın ve mantıksal akışın korunmasını gerektirir. RAG'ın benzerlik temelli yaklaşımı, bu tür uzun menzilli bağımlılıkları yakalamakta yetersiz kalabilir ve sonuç olarak, yalnızca yerel olarak doğru ancak global olarak tutarsız veya eksik yanıtlar üretilebilir. ToM framework'ü, bu sınırlılığı aşmak için belgenin doğal hiyerarşik yapısını kullanarak parçalar arasındaki ilişkileri açıkça modellemeyi ve özyinelemeli bir çıkarım süreciyle bu ilişkileri göz önünde bulundurmayı hedefler . Bu sayede, ToM'un RAG'a kıyasla daha tutarlı ve bağlamsal olarak daha zengin çıkarımlar yapabileceği iddia edilmektedir.

1.2.2. RAG ile Uzun Bağlam LLM'lerinin Kapsamlı Karşılaştırması

20 Mart 2025 tarihli **"A Comprehensive Survey on Long Context Language Models"** başlıklı akademik makale, uzun bağlam dil modelleri (LCLM) alanında yapılan son gelişmeleri kapsamlı bir şekilde incelemektedir . Bu derleme makalesi, etkili ve verimli LCLM'lerin nasıl elde edileceği, bu modellerin nasıl eğitilip dağıtılacağı ve nasıl kapsamlı bir şekilde değerlendirileceği konularını merkeze alan üç ana başlık etrafında yapılandırılmıştır. İlk bölümde, uzun bağlam işleme odaklı veri stratejileri, mimari tasarımlar ve iş akışı yaklaşımları tartışılmaktadır. İkinci bölümde, LCLM eğitimi ve çıkarımı için gerekli altyapı detaylı bir şekilde incelenmektedir. Üçüncü bölümde ise, uzun bağlam anlama ve uzun form üretim için değerlendirme paradigmları, LCLM'lerin davranışsal analizi ve mekanizma yorumlanabilirliği sunulmaktadır. Bu kapsamlı inceleme, hem araştırmacılar hem de mühendisler için güncel bir kaynak niteliğindedir ve LCLM'lerin çeşitli uygulama senaryolarını ve gelecekteki olası gelişme yönlerini de kapsamlı bir şekilde ele almaktadır .

Aynı zamanda, 6 Mart 2025 tarihli **"Shifting Long-Context LLMs Research from Input to Output"** başlıklı bir makale, uzun bağlam büyük dil modelleri (LLM) araştırmalarında önemli bir paradigma değişikliği önermektedir . Mevcut gelişmelerin büyük ölçüde uzun girdi bağlamlarının işlenmesine odaklandığını ve bu alanda önemli ilerlemeler sağladığını

belirten makale, **uzun form çıktı üretiminin** esit derecede kritik ancak nispeten daha az ilgi gören bir yön olduğunu vurgular. Roman yazma, uzun vadeli planlama ve karmasık reasoning gibi görevler, modellerin genis bağlamları anlamasının yanı sıra tutarlı, bağlamsal olarak zengin ve mantıksal olarak tutarlı uzun metinler üretmesini gerektirir. Bu talepler, mevcut LLM yeteneklerinde kritik bir bosluğu ortaya koymaktadır. Makale, bu az kesfedilmis alanın önemini vurgulamakta ve gerçek dünya uygulamalarında büyük potansiyel tasıyan yüksek kaliteli, uzun form çıktılar üretmek için tasarlanmıs temel LLM'lerin gelistirilmesine yönelik odaklanmıs çabalar çağrısında bulunmaktadır .

1.3. Değerlendirme ve Benchmarklar

1.3.1. LongBench v2: Gerçekçi Uzun Bağlam Görevleri için Yeni Bir Benchmark

2025 yılı itibarıyla, büyük dil modellerinin (LLM) uzun bağlamı anlama ve üzerinde derinlemesine akıl yürütme yeteneklerini değerlendirmek için önemli bir ihtiyaç ortaya çıkmıstır. Bu ihtiyaca cevap olarak, ACL 2025 konferansında sunulan **"LongBench v2: Towards Deeper Understanding and Reasoning on Realistic Long-context Multitasks"** baslıklı akademik çalısma, alana önemli bir katkı sağlamaktadır . Bu yeni benchmark, mevcut değerlendirme araçlarının yetersiz kaldığı noktalarda ilerleme kaydetmeyi hedeflemektedir. LongBench v2, bağlam uzunluğu **8.000 kelimeden (8K) 2 milyon kelimeye (2M)** kadar değisen, oldukça zorlu **503 adet çoktan seçmeli sorudan** olusmaktadır. Bu sorular, altı ana görev kategorisinde toplanmıstır: tek belgeli soru-cevap (QA), çok belgeli QA, uzun bağlamda öğrenme (long in-context learning), uzun diyalog geçmisi anlama, kod deposu anlama ve uzun yapılandırılmıs veri anlama. Bu çesitlilik, modelin sadece belirli bir bilgiyi bulup çıkarmasını değil, aynı zamanda farklı türdeki bilgileri entegre ederek karmasık sorulara cevap verebilme yeteneğini test etmeyi amaçlar .

LongBench v2'nin olusturulma süreci, yüksek kalite ve zorluk seviyesini garanti altına alacak sekilde tasarlanmıstır. Benchmark, neredeyse **100 farklı mesleki geçmise sahip ve yüksek eğitim düzeyine sahip bireylerden** toplanan verilerle olusturulmustur. Otomatik ve manuel inceleme süreçleriyle kalite kontrolü sağlanmıs, bu da insan uzmanlarının bile 15 dakikalık bir zaman sınırlaması altında sadece **%53.7 oranında doğruluk** elde edebilmesine neden olmustur. Bu sonuç, benchmark'ın insan seviyesinin üzerinde bir zorluk barındırdığını göstermektedir. Değerlendirme sonuçları, mevcut en iyi modelin doğrudan soruları yanıtlarken yalnızca **%50.1 doğruluk oranına** ulasıtığını ortaya koymustur. Ancak, daha uzun akıl yürütme süreçlerini içeren **o1-preview** modeli, **%57.7 doğruluk oranına** ulasarak insan temel çizgisini %4 oranında asmayı basarmıstır. Bu bulgu, uzun bağlam zorluklarının üstesinden gelmek için **artırılmıs akıl yürütme yeteneği**

ve çıkarım süresi hesaplama gücünün (inference-time compute) önemini vurgulamaktadır .

1.3.2. mLongRR-V2: Çok Dilli ve Nöro-Sembolik Retrieval & Reasoning Benchmark'i

Uzun bağlam modellerinin değerlendirilmesinde, mevcut benchmark'ların çoğunlukla tek dilli (çoğunlukla İngilizce) ve nöro-sembolik yaklaşımları göz ardı eden görevlere odaklandığı bir gerçektir. Bu durum, modellerin çok dilli ortamlardaki ve karmaşık sembolik manipülasyon gerektiren görevlerdeki performansını değerlendirme konusunda bir boşluk yaratmaktadır. Bu boşluğu doldurmak amacıyla, **mLongRR-V2 (Multilingual Long-context Retrieval and Reasoning Benchmark)** adlı yeni bir benchmark tanıtılmıştır. Bu benchmark, uzun bağlam retrieval ve reasoning yeteneklerini çok dilli bir ortamda ve nöro-sembolik görevlerle birlikte değerlendirmeyi hedefler. mLongRR-V2, farklı dillerdeki uzun belgelerden bilgi toplama, bu bilgileri çapraz dilli olarak sentezleme ve sembolik mantık kurallarını uygulama gibi karmaşık görevler içermektedir. Bu yapısıyla, mLongRR-V2, modellerin dil agnostik temsillerini ve sembolik akıl yürütme kapasitelerini test ederek, daha güçlü ve evrensel uzun bağlam sistemlerinin geliştirilmesine katkı sağlamayı amaçlamaktadır.

2. Model Mimarileri ve Hesaplama Verimliliği

2.1. Mixture of Experts (MoE) ve Uzmanlık Tabanlı Modeller

2.1.1. MoE-MLA-RoPE: Uzmanlar ve Gizli Dikkat Mekanizmalarının Birleştirilmesi

Kaynak kısıtlı ortamlarda, örneğin mobil cihazlarda, gömülü sistemlerde ve kenar (edge) bilişim platformlarında dil modellerinin dağıtımı, sadece parametre sayısını azaltmanın ötesinde temel mimari yenilikler gerektirir. 8 Kasım 2025 tarihinde ACL 2025 konferansında sunulan **"Unifying Mixture of Experts and Multi-Head Latent Attention for Efficient Language Models"** başlıklı çalışma, bu ihtiyaca yönelik olarak **MoE-MLA-RoPE** adında yeni bir mimari sunmaktadır . Bu mimari, **Mixture of Experts (MoE)** , **Multi-head Latent Attention (MLA)** ve **Rotary Position Embeddings (RoPE)** olmak üzere üç ortogonal verimlilik mekanizmasını birleştirmektedir. Bu teknikler, birbirini tamamlayan darboğazları ele alır: MoE koşullu yönlendirme ile hesaplama FLOP'larını azaltır, MLA düşük rütbeli anahtar-değer projeksiyonlarıyla belleği sıkıştırır ve RoPE ise konum gömme (position embedding) parametrelerini ortadan kaldırırken uzunluk genelleştirmesini iyileştirir. Araştırmanın temel içgörüsü, MoE'deki uzman uzmanlaşmasının MLA'nın sıkıştırmasından kaynaklanan bilgi kaybını telafi edebileceği, aynı zamanda MLA'nın bellek tasarruflarının aynı bellek bütçesi içinde daha fazla uzman

dağıtımına olanak tanıyacağıdır. Bu, daha fazla uzmanın daha iyi uzmanlaşmayı sağladığı, bu da daha agresif sıkıştırmayı kalite kaybı olmadan mümkün kılan olumlu bir geri bildirim döngüsü yaratır .

MoE–MLA–RoPE mimarisi, üç temel yenilik sunar. Birincisi, **64 mikro–uzman ve top–k seçimi** ile ince taneli uzman yönlendirmedir, bu da yaklaşık **3.6 x 10^7 olası uzman kombinasyonu**yla esnek uzmanlaşmayı mümkün kılar. İkincisi, ortak kalıplar için **2 sürekli aktif uzman** ayıran ve 62 uzmanlık uzmanından 6'sına yönlendirme yapan **paylaşılan uzman izolasyonudur**. Üçüncüsü, uzman kullanımını sürdürürken birincil kayıp optimizasyonuna müdahale etmeyen, **gradyan çatışması olmayan yük dengelemedir**. 17M ila 202M parametre aralığındaki modeller üzerinde yapılan kapsamlı deneyler, MoE–MLA–RoPE'nin **r=d/2 sıkıştırma oranıyla %68 KV önbellek belleği azaltımı ve 3.2 kat çıkarım hızlandırması** elde ederken rekabetçi perplexity değerini koruduğunu (yalnızca **%0.8 bozulma**) göstermektedir. 53.9M parametreli modellerle karşılaştırıldığında, MoE–MLA–RoPE, vanilla transformer'lara kıyasla doğrulama kaybını **%6.9 oranında iyileştirirken** ileri besleme basına **%42 daha az aktif parametre** kullanır. FLOP eşleştirilmiş deneylerde kazanımlar daha da büyüktür: **%11.1 iyileştirme ve 3.2 kat çıkarım hızlandırması**. GPT–4'ün yargıcı olarak kullanıldığı otomatik değerlendirme, üretim kalitesindeki iyilesmeleri doğrular ve tutarlılık (**8.1/10**), yaratıcılık (**7.9/10**) ve dil bilgisi doğruluğu (**8.2/10**) açısından daha yüksek puanlar verir. Bu sonuçlar, mimari sinerjinin, parametre ölçeklendirmesinden ziyade, kaynak kısıtlı dil modeli dağıtımı için verimlilik sınırını tanımladığını göstermektedir .

2.1.2. MoTE (Mixture of insighTful Experts): Uzmanlık ve Öngörünün Sinerjisi

2025 yılına gelindiğinde, büyük dil modellerinin (LLM) yetenekleri dramatik şekilde genisleмis olsa da, bu modelleri insan değerleriyle hizalamak hala önemli bir zorluk olarak kalmaktadır. Son çalısmalar, güçlü LLM'lerin, baslangıçtaki güvenli olmayan yanıtlarını düzelterek veya insan müdahalesi olmadan yanıtları özerk olarak sıralayarak **kendi kendine hizalama (self–alignment)** sağlayabileceğini göstermektedir. Ancak bu yöntemler, LLM'lerin varsayılan olarak ortaya çıkan yeteneklerine asırı derecede bağımlıdır ve modeli güncellenmis yanıtlarla hizalarken tüm ara muhakeme adımlarını göz ardı eder. Bu sınırlılıkları gidermek için, 2024 yılı sonlarında sunulan **Mixture of insighTful Experts (MoTE)** adlı yeni bir kendi kendine hizalama yöntemi gelistirilmistir . MoTE, **Chain of Thought (CoT)** yaklasımını kullanarak, **Soru Analizi, Yanıt Yönlendirme ve Güvenli Yanıt Üretimi** asamalarını içeren **AlignCoT** adı verilen bir süreç önermektedir. Bu yöntem, **7B parametreli** gibi daha küçük ve zayıf modellerin bile yüksek kaliteli ve güvenli yanıtlar üretmesini sağlamayı hedeflemektedir.

MoTE mimarisi, AlignCoT sürecinin her bir bileşenini geliştirmek için **Mixture of Experts (MoE)** yaklaşımını uygulayarak hizalama verimliliğini önemli ölçüde artırır. MoTE yaklaşımı, LLM'leri insan değerleriyle hizalamada mevcut yöntemlerden daha üstün performans göstermekle kalmıyor, aynı zamanda kendi kendine üretilen verilerin kullanılmasının hem gelişmiş hizalama hem de eğitim verimliliği açısından ikili faydalarını ortaya koyuyor . Bu çalışma, özellikle hizalama ve güvenlik, gizlilik ve toplumsal etkiler gibi alanlara odaklanan **ICLR 2025** konferansına sunulmuştur. MoTE'nin sunduğu yenilik, sadece sonuç odaklı değil, muhakeme sürecini de dikkate alan bir hizalama stratejisi sunmasıdır. Bu, modelin güvenli ve faydalı yanıtlar üretme yeteneğini, araştırmacılar tarafından "insan değerleriyle hizalama" olarak tanımlanan kritik bir hedef doğrultusunda geliştirmeyi amaçlamaktadır. MoTE'nin sunduğu bu kapsamlı yaklaşım, özellikle kaynakları sınırlı ortamlarda güvenli ve etkili AI sistemleri geliştirmek için önemli bir adım olarak değerlendirilmektedir.

2.1.3. MoIIE: İçsel ve Dışsal Uzmanlık Birleştirme

Çok modallı büyük dil modelleri (LVLM'ler), görsel ve dilsel modlar arasındaki etkileşimleri anlama konusunda önemli ilerlemeler kaydetmiştir. Ancak, bu modellerin hem **modaya özgü (modality-specific)** özellikleri hem de **çapraz moda (cross-modal)** ilişkileri aynı anda etkili bir şekilde yakalama yeteneği hâlâ bir zorluk teşkil etmektedir. Bu sorunu ele almak için, **"MoIIE: Mixture of Intra- and Inter-Modality Experts for Large Vision Language Models"** başlıklı çalışmada, LVLM'lere entegre edilen yeni bir çok modallı MoE (Mixture of Experts) varyanı olan **MoIIE** tanıtılmaktadır . MoIIE mimarisi, hem **içsel (intra-)** hem de **dışsal (inter-)** moda uzmanlığını aynı anda yakalamayı hedefler. Bu, modelin hem her bir modanın kendi içindeki özgün özelliklerini hem de farklı modalar arasındaki karmaşık ilişkileri öğrenmesine olanak tanır. Mimari, önceden eğitilmiş bir LVLM'nin temelinde, görsel ve dilsel modalar için ayrı ayrı içsel uzmanlar (intra-modality experts) ile bu modalar arasındaki etkileşimleri modelleyen dışsal uzmanlar (inter-modality experts) olmak üzere iki tür uzman içerir. Bu hiyerarşik uzman yapısı, modelin farklı modalitelerdeki bilgileri daha verimli ve etkili bir şekilde işlemesini sağlar. MoIIE'nin deneysel sonuçları, çeşitli çok modallı görevlerde mevcut yöntemlere kıyasla üstün performans gösterdiğini ve özellikle karmaşık çapraz moda anlama gerektiren senaryolarda önemli avantajlar sağladığını ortaya koymuştur. Bu çalışma, MoE mimarilerinin çok modallı bağlamda nasıl uyarlanabileceğine dair önemli bir örnek teşkil etmektedir.

2.2. Model Sıkıştırma ve Etkili Çıkarım (Inference)

2.2.1. MoE-SVD: SVD Tabanlı Yapılandırılmış MoE Sıkıştırma

2025 yılında, özellikle Mixture of Experts (MoE) mimarisine sahip büyük dil modellerinin (LLM) sıkıştırılması konusunda önemli bir ilerleme, ICML 2025 konferansında sunulan **"MoE-SVD: Structured Mixture-of-Experts LLMs Compression"** başlıklı makale ile kaydedilmiştir . MoE mimarileri, daha iyi ölçeklenebilirlik sunarak LLM'lerin performansını artırmakla birlikte, artan parametre sayısı ve bellek ihtiyacı nedeniyle dağıtım (deployment) aşamasında zorluklar yaratmaktadır. Bu çalışma, herhangi bir ek eğitim gerektirmeden, MoE LLM'lerine özel yeni bir **ayrıştırma (decomposition) tabanlı sıkıştırma framework'ü** sunmaktadır. **MoE-SVD**, **Tekil Değer Ayrıştırması (Singular Value Decomposition – SVD)** gücünden yararlanarak, MoE mimarilerindeki **ayrıştırma çöküşü (decomposition collapse)** ve **matris yedekliliği (matrix redundancy)** gibi kritik sorunları ele alır. İlk olarak, uzmanları (experts) daha hızlı çıkarım ve bellek optimizasyonu sağlayan kompakt düşük-rank matrislere ayırır. Özellikle, ağırlık tekil değerleri ve aktivasyon istatistiklerine dayalı duyarlılık ölçütleri kullanarak otomatik olarak ayrıştırılabilir uzman katmanlarını belirleyen seçici bir ayrıştırma stratejisi önerir. Ardından, tüm uzmanlar arasında tek bir V-matrisi paylaşır ve U-matrisler için bir top-k seçimi uygular. Bu düşük-rank matris paylaşımı ve budama şeması, uzmanlar arasındaki çeşitliliği korurken önemli parametre azaltımı sağlar. **Mixtral, Phi-3.5, DeepSeek ve Qwen2** gibi çeşitli MoE LLM'lerinde kapsamlı deneyler, MoE-SVD'nin diğer sıkıştırma yöntemlerini geride bıraktığını, **%60 sıkıştırma oranı** ve minimum performans kaybıyla **1.5 kat daha hızlı çıkarım** sağladığını göstermiştir .

2.2.2. D²-MoE: Delta Açılımı ile MoE Sıkıştırma

MoE tabanlı LLM'lerin sıkıştırılmasına yönelik bir diğer yenilikçi yaklaşım, aynı ICML 2025 konferansında sunulan **"Delta Decompression for MoE-based LLMs Compression"** başlıklı makalede tanıtılan **D²-MoE (Delta Decompression for MoE)** framework'üdür . Bu yöntem, MoE mimarilerindeki **uzman çeşitliliğini (expert diversity)** gözlemleyerek, uzman ağırlıklarını paylaşılan bir **temel ağırlık (shared base weight)** ve benzersiz **delta ağırlıklarına (unique delta weights)** ayırır. Yöntemin ilk adımı, her uzmanın ağırlığını, paylaşılan bileşenleri yakalamak için **Fisher bilgi matrisi** kullanılarak temel ağırlıkla birleştirmektir. Ardından, delta ağırlıklarının düşük-rank özelliklerinden yararlanılarak **Tekil Değer Ayrıştırması (SVD)** ile sıkıştırılması sağlanır. Son olarak, temel ağırlıklar için statik ve dinamik yedeklilik analizini birleştiren **yarı-dinamik yapılandırılmış budama (semi-dynamical structured pruning)** stratejisi tanıtılarak, girdi bağımsızlığını korurken daha fazla parametre azaltımı elde edilir. Bu yöntem sayesinde, D²-MoE, ek eğitime gerek kalmadan MoE LLM'lerini yüksek sıkıştırma oranlarına kadar başarıyla küçültür. **Mixtral, Phi-3.5, DeepSeek ve Qwen2** MoE LLM'lerinde yapılan

kapsamlı deneyler, D^2-MoE'nin **%40-60 sıkıştırma oranlarında** diğer sıkıştırıcılara kıyasla **%13'e kadar performans artışı** sağladığını göstermiştir .

2.2.3. u-MoE: Test-Zamanı Pruning ile Mikro-Ölçekli Uzmanlar

Büyük temel modellerin (foundation models) yüksek hesaplama ihtiyacını azaltmaya yönelik bir başka çalışma, ICML 2025 Workshop'unda sunulan **"u-MoE: Test-Time Pruning as Micro-Grained Mixture-of-Experts"** başlıklı makaledir . Bu çalışma, yeniden eğitime gerek duymadan **aktivasyon-tabanlı (activation-aware)** sıkıştırma tekniklerini temel alır. Ancak, bu teknikler kalibrasyon verisine bağımlı olduğu için, görünmeyen aşağı akış (downstream) görevlerinde **domain kayması (domain shift)** ortaya çıkabilir. **u-MoE (micro-MoE)** , bu sorunu, aktivasyon-tabanlı budama işlemini her komut (prompt) için uyarlanabilir şekilde verimli bir şekilde yürüterek çözmeyi amaçlar. Bu yaklaşım, test-zamanı budama işlemini **mikro-uzmanların (micro-experts)** bir karışımı olarak formüle eder. Yapılan deneyler, u-MoE'nin **komuta-bağımlı yapılandırılmış seyreklik (prompt-dependent structured sparsity)** için dinamik olarak uyarlanabildiğini göstermiştir. Bu yöntem, modelin farklı girdilere göre kendini uyarlayabilen, daha verimli ve esnek bir çıkarım süreci sunmasını sağlar. u-MoE, özellikle kaynak kısıtlı ortamlarda, büyük modellerin etkili bir şekilde kullanılmasına olanak tanıyan pratik bir çözüm olarak öne çıkmaktadır .

2.2.4. WMT 2025 Model Sıkıştırma Paylaşımlı Görevi: Endüstri ve Akademik Katkılar

Model sıkıştırma alanındaki pratik uygulamalar ve endüstri-akademi iş birliği, 8-9 Kasım 2025 tarihlerinde düzenlenen **10. Makine Çevirisi Konferansı (WMT25)** kapsamındaki **"Model Compression Shared Task"** ile somutlaşmıştır . Bu paylaşımlı görev, büyük dil modellerinin (LLM) makine çevirisi (MT) senaryolarında pratik dağıtımını mümkün kılmak amacıyla model sıkıştırma tekniklerinin potansiyelini değerlendirmeyi hedeflemiştir. Katılımcılar, **kısıtlı (constrained)** ve **kısıtlamasız (unconstrained)** olmak üzere iki farklı kategoride yarışmıştır. Kısıtlı kategoride, tüm katılımcılar belirli bir modeli (**Aya Expanse 8B**) ve dil çiftlerini (Çekçe→Almanca, Japonca→Çince, İngilizce→Arapça) sıkıştırmak zorundaydı. Bu, sonuçların doğrudan karşılaştırılabilmesini sağlamıştır. Değerlendirme, **COMET ve MetricX** ölçümleri kullanılarak çeviri kalitesi, model boyutu ve Nvidia A100 GPU üzerindeki çıkarım hızı gibi üç ana boyutta gerçekleştirilmiştir. Göreve üç ekipten toplam 12 katılım gelmiştir. Bu katılımlar arasında, **LeanQuant** adlı hassasiyet-tabanlı bir nicemleme (quantization) yöntemi kullanan **LeanAya**, iteratif katman budama ile model boyutunu azaltan **TCD-Kreasof**, ve evrimsel katman birleştirme yaklaşımı ile çalışan **Vicomtech**'in çözümleri öne çıkmıştır. Bu paylaşımlı görev, model sıkıştırma alanındaki güncel araştırmaların pratikte nasıl uygulandığını ve

farklı tekniklerin performansını ortaya koyarak, hem akademik hem de endüstriyel topluluk için önemli bir veri noktası oluşturmuştur .

3. Yapay Zekâ Hizalaması (Alignment) ve İnsan Geri Bildirimi (RLHF)

3.1. RLHF Tekniklerinde Yenilikler ve Geliştirmeler

3.1.1. Dil Geri Bildirimi ile RLHF Geliştirme: Seq2Seq Ödül Modelleme

Reinforcement Learning from Human Feedback (RLHF) süreçlerini iyileştirmeye yönelik önemli bir katkı, 2025 AAAI Konferansı'nda sunulan **"Sequence to Sequence Reward Modeling: Improving RLHF by Language Feedback"** başlıklı makaledir . Geleneksel RLHF, insan tercihlerinden eğitilen bir **ödül modeli (reward model – RM)** kullanarak LLM'leri hizalamayı amaçlar. Ancak bu yöntem, ödül modelünün insan tercihlerini doğru yansıtamaması nedeniyle **yanlılığa (bias) açık yerel optimizasyona** yatkındır. Bu durum, modellerin beklenmedik genellemeler yapmasına ve hizalama hedeflerine ulaşamamasına neden olabilir. Bu sorunu çözmek için, yazarlar yeni bir **sequence-to-sequence (seq2seq) ödül modelleme** yöntemi önermektedir. Bu yöntemin temel fikri, skaler geri bildirim yerine **dil geri bildiriminden (language feedback)** öğrenmenin, ek açıklama gerektirmeden RLHF'yi iyileştirebileceğidir. Geleneksel ikili maksimum olabilirlik tahminini (MLE) hedef alan ödül modellemeyi, **seq2seq MLE** ile değiştirerek, daha zengin ve ince ayarlı dil geri bildirimi sağlar. Bu yaklaşım, ek açıklama, model veya eğitim aşaması gerektirmeden RLHF sürecini geliştirir. Deneyler, bu yöntemin tek tur güvenlik diyaloglarında yanıt reddetme paradigmasını ve metin özetleme görevlerinde uzun yanıt yanlılığını azalttığını göstermiştir. Ayrıca, seq2seq RM'nin **2B ve 7B parametreli LLM'lerde 3 NLP görevi** boyunca ortalama **%76.9 kazanma oranı** ile RLHF performansını iyileştirdiği ve **dağılım dışı (out-of-distribution)** komutlarda bile etkili olduğu gösterilmiştir .

3.1.2. LongReward: Uzun Bağlam Modelleri için Yapay Zekâ Geri Bildirimi

Uzun bağlam modellerinin hizalanması, geleneksel RLHF yöntemlerinin kısa vadeli ve anlık geri bildirime dayanması nedeniyle zorluklarla karşılaşmaktadır. Bu durum, modellerin uzun vadeli sonuçları veya bağlam içindeki karmaşık ilişkileri göz ardı etmesine yol açabilir. Bu sorunu çözmek için, **LongReward** adlı yeni bir geri bildirim mekanizması önerilmiştir. LongReward, modelin çıktılarının uzun vadeli etkilerini değerlendirmek için tasarlanmıştır. Bu yöntem, modelin ürettiği bir yanıtın, uzun bir diyalog veya belge bağlamındaki sonraki etkileşimler üzerindeki etkisini ölçer. Örneğin, bir özetin orijinal belgedeki bilgileri ne kadar iyi koruduğunu veya bir planın uzun vadeli

hedeflere ne kadar uygun olduğunu değerlendirir. LongReward, bu uzun vadeli etkileri ölçmek için özel olarak tasarlanmış metrikler ve değerlendirici modeller kullanır. Bu yaklaşım, RLHF sürecine uzun vadeli bir perspektif kazandırarak, modellerin daha stratejik, tutarlı ve bağlamsal olarak uygun yanıtlar üretmesini teşvik eder. LongReward'ın deneysel sonuçları, bu yöntemin uzun bağlam diyalog ve özetleme görevlerinde, geleneksel RLHF yöntemlerine kıyasla önemli performans artışları sağladığını göstermiştir.

3.1.3. Çok Amaçlı Hizalama: Gradient-Adaptive Policy Optimization (GAPO)

Çeşitli ve bazen çelişen insan tercihlerine göre LLM'leri hizalama zorluğuna yönelik bir çözüm, 2025 ACL Konferansı'nda sunulan **"Gradient-Adaptive Policy Optimization: Towards Multi-Objective Alignment of Large Language Models"** başlıklı makalede sunulmuştur . RLHF, LLM'leri insan tercihleriyle hizalamak için güçlü bir teknik olsa da, çeşitli tercihler çatıştığında etkisiz kalabilir. Bu sorunu çözmek için, yazarlar insan değerlerini hizalamayı, potansiyel olarak çelişen bir dizi amacı en üst düzeye çıkarmayı hedefleyen **çok amaçlı bir optimizasyon problemi** olarak çerçevelemiştir. Bu bağlamda, **Gradient-Adaptive Policy Optimization (GAPO)** adlı yeni bir ince ayar paradigması tanıtılmıştır. GAPO, birden fazla gradyan inişi kullanarak LLM'leri çeşitli tercih dağılımlarıyla hizalamak için tasarlanmıştır. Her amaç için gradyanları uyarlanabilir şekilde yeniden ölçeklendirerek, amaçlar arasındaki dengeyi en iyi şekilde sağlayan bir güncelleme yönü belirler. Ek olarak, kullanıcının farklı amaçlara dair tercihlerini dahil eden ve kullanıcının spesifik ihtiyaçlarıyla daha iyi hizalanmış Pareto çözümleri elde eden **P-GAPO** adlı bir varyasyon da sunulmuştur. Bu yaklaşım, çok boyutlu hizalama problemini sistematik bir şekilde ele alarak, daha esnek ve kişiselleştirilmiş AI sistemlerinin geliştirilmesine katkı sağlar .

3.2. Hizalama Teorisi ve Çoğulcu Yaklaşımlar

3.2.1. Çoğulcu Hizalama: Çoklu Kullanıcı Perspektiflerine Uyum

Geleneksel hizalama yaklaşımları, genellikle tek bir, evrensel olarak kabul edilen "doğru" veya "faydalı" yanıt fikrine dayanır. Ancak, gerçek dünyada kullanıcı tercihleri ve değerleri büyük ölçüde çeşitlilik gösterir. Bu nedenle, tek bir hizalama hedefi, tüm kullanıcı gruplarının ihtiyaçlarını karşılamakta yetersiz kalabilir. Bu sorunu ele almak için, **çoğulcu hizalama (pluralistic alignment)** kavramı öne çıkmaktadır. Bu yaklaşım, bir LLM'nin çoklu ve hatta çelişen kullanıcı perspektiflerine aynı anda uyum sağlayabilmesini hedefler. Çoğulcu hizalama, modelin farklı kullanıcı grupları için farklı davranış kalıplarını öğrenmesini ve bağlama göre en uygun olanı seçmesini sağlar. Bu,

modelin kişiselleştirilmiş yanıtlar üretmesini ve farklı kültürel, etik veya kişisel tercihlere duyarlı olmasını mümkün kılar. 2025 yılında yapılan araştırmalar, çoğulcu hizalamanın, RLHF süreçlerine çoklu ödül fonksiyonları veya kullanıcı tercihlerini açıkça modelleyen mekanizmalar entegre ederek nasıl gerçekleştirilebileceğini göstermektedir. Bu yöntemler, AI sistemlerinin daha adil, kapsayıcı ve kullanıcı merkezli olmasına katkı sağlar.

3.2.2. Aksiyomatik Yaklaşım ve Pairwise Kalibrasyon

Hizalama problemini daha teorik bir zeminde ele alan bir başka yaklaşım, **aksiyomatik yaklaşım**dır. Bu yöntem, bir hizalama sürecinin hangi temel özellikleri veya "aksiyomları" sağlaması gerektiğini tanımlar. Örneğin, bir aksiyom, "eğer bir yanıt A, yanıt B'den tercih ediliyorsa ve B, C'den tercih ediliyorsa, A'nın C'den de tercih edilmesi gerekir" (geçişlilik) şeklinde olabilir. Bu aksiyomlar, hizalama algoritmalarının mantıksal tutarlılığını ve güvenilirliğini değerlendirmek için bir çerçeve sağlar. Bu teorik çerçeveyi pratiğe geçirmek için **pairwise kalibrasyon** teknikleri kullanılır. Bu teknikler, modelin iki farklı çıktı arasındaki tercih sıralamasını ne kadar iyi öğrendiğini ölçer. Aksiyomatik yaklaşım ve pairwise kalibrasyon birleştirildiğinde, hizalama süreçlerinin sistematik olarak analiz edilmesi ve iyileştirilmesi mümkün hale gelir. Bu yöntemler, özellikle ödül modeli tasarımı ve tercih verisi kalitesinin değerlendirilmesi gibi alanlarda önemli avantajlar sağlar. 2025 yılında yapılan çalışmalar, bu teorik araçların, mevcut RLHF sistemlerindeki önyargıları ve tutarsızlıkları tespit etmekte etkili olduğunu göstermiştir.

3.2.3. İnsan–AI Çift Yönlü Hizalama: Teorik Çerçeve ve Yönler

Geleneksel hizalama paradigması, genellikle tek yönlüdür: İnsan tercihlerine göre AI modelini şekillendirmek. Ancak, bu yaklaşım, insan–AI etkileşiminin dinamik ve karşılıklı doğasını göz ardı edebilir. **İnsan–AI çift yönlü hizalama (bidirectional alignment)** , bu eksikliği gidermeyi hedefleyen yeni bir teorik çerçevedir. Bu yaklaşım, AI'nın insan davranışını anlamasının yanı sıra, insanların da AI'nın yetenekleri ve sınırlılıkları konusunda eğitilmesini savunur. Bu, insan–AI takımlarının daha etkili bir şekilde işbirliği yapmasını ve ortak hedeflere ulaşmasını sağlar. Çift yönlü hizalama, insan–AI etkileşimini bir **döngüsel ve uyarlanabilir bir süreç** olarak görür. AI, insan geri bildiriminden öğrenirken, insan da AI'nın açıklamaları ve önerileriyle karar alma süreçlerini geliştirir. Bu teorik çerçeve, özellikle karmaşık karar verme, yaratıcı problem çözme ve eğitim gibi alanlarda, insan–AI işbirliğini güçlendirmek için yeni yöntemlerin geliştirilmesine yol açmaktadır. 2025 yılında yapılan araştırmalar, bu çift yönlü sürecin, hem insan performansını hem de AI'nın güvenilirliğini artırdığını göstermektedir.

3.3. Hizalama Sınırlılıkları ve Güvenlik

3.3.1. RLHF'nin Gizli Önyargıları Gidermedeki Yetersizliği

Reinforcement Learning from Human Feedback (RLHF), büyük dil modellerini insan değerleriyle hizalamak için en yaygın kullanılan yöntemlerden biridir. Ancak, 2025 yılında yapılan araştırmalar, RLHF'nin bazı önemli sınırlılıklarını ortaya koymuştur. Özellikle, RLHF'nin **gizli önyargıları (hidden biases)** gidermede yetersiz kalabileceği gösterilmiştir. Bu önyargılar, ödül modelinin eğitildiği insan tercih verilerinden kaynaklanabilir. Eğer bu veriler toplumsal önyargıları yansıtıyorsa, RLHF süreci bu önyargıları güçlendirebilir veya daha da kötüleştirebilir. Örneğin, ödül modeli, belirli demografik gruplara yönelik ayrımcı yanıtları "tercih edilebilir" olarak değerlendirebilir. Bu durum, hizalanmış modelin adil ve tarafsız olmasını zorlaştırır. Araştırmacılar, bu sorunu çözmek için ödül modelinin önyargılarını açıkça ölçme ve azaltma yöntemleri geliştirmeye çalışmaktadır. Bu yöntemler arasında, tercih verilerinin çeşitliliğini artırmak, ödül modelinin kararlarını açıklayabilir hale getirmek ve adalet temelli kayıp fonksiyonları kullanmak gibi stratejiler yer almaktadır. Bu çalışmalar, daha güvenli ve etik AI sistemleri geliştirmek için kritik önem taşımaktadır.

3.3.2. Model "Elastikliği": Büyük Modellerin Hizalamaya Direnmesi

Büyük dil modellerinin boyutu ve karmaşıklığı arttıkça, bu modelleri hizalamak da zorlaşmaktadır. Bu fenomen, **model "elastikliği" (model elasticity)** olarak adlandırılmaktadır. Elastiklik, bir modelin parametrelerinin, ince ayar (fine-tuning) süreçlerinde ne kadar kolay değiştirilebileceğini ifade eder. Daha büyük modeller, genellikle daha düşük elastikliğe sahiptir; yani, parametreleri hizalama verisine göre daha dirençli bir şekilde ayarlanır. Bu durum, özellikle zorlu veya nadir hizalama görevlerinde, modelin istenen davranışı öğrenmesini zorlaştırabilir. Araştırmacılar, model elastikliğini artırmak için çeşitli yöntemler geliştirmektedir. Bu yöntemler arasında, ince ayar sürecinde daha agresif öğrenme oranları kullanmak, modelin belirli katmanlarını hedeflemek veya hizalama verisini sentetik olarak artırmak gibi stratejiler yer almaktadır. Model elastikliği kavramı, hizalama çalışmalarında model boyutu ile hizalama etkinliği arasındaki dengeyi anlamak açısından önemli bir perspektif sunmaktadır.

3.3.3. MoE LLM'lerine Zararlı İnce Ayara Karşı Savunma

Mixture of Experts (MoE) mimarisi, hesaplama verimliliği açısından önemli avantajlar sunsa da, bu modellerin güvenliği konusunda yeni zorluklar ortaya çıkarmaktadır. Özellikle, MoE modelleri, **zararlı ince ayar (malicious fine-tuning)** saldırılarına karşı

savunmasızdır. Bu tür bir saldırıda, kötü niyetli bir aktör, modelin belirli uzmanlarını hedef alarak, modelin zararlı, yanıltıcı veya ayrımcı içerik üretmesine neden olabilir. MoE mimarisinin modüler yapısı, bu tür hedefli saldırıların gerçekleştirilmesini kolaylaştırabilir. Bu tehdidi ortadan kaldırmak için, MoE LLM'lerine yönelik savunma mekanizmaları geliştirilmektedir. Bu mekanizmalar arasında, uzman yönlendirme kararlarını rastgeleleştirmek, uzman ağırlıklarını korumak için kriptografik teknikler kullanmak veya modelin çıktılarını sürekli olarak zararlı içerik açısından izlemek gibi yöntemler yer almaktadır. Bu güvenlik önlemleri, MoE modellerinin güvenli ve güvenilir bir şekilde dağıtımı için hayati önem taşımaktadır.

4. Prompt Mühendisliği ve Model Geliştirme

4.1. Prompt Mühendisliği için Kapsamlı Bir Çerçeve

4.1.1. Profil ve Talimat Tabanlı Prompt Tasarımı

2025 yılı boyunca, büyük dil modellerinden (LLM) en iyi şekilde yararlanmak için prompt mühendisliği alanında önemli ilerlemeler kaydedildi. **"A comprehensive taxonomy of prompt engineering techniques for large language models"** başlıklı akademik bir makale, bu tekniklerin sistematik bir sınıflandırmasını sunarak, prompt tasarımı için kapsamlı bir çerçeve oluşturmuştur . Bu çalışma, prompt mühendisliğini farklı kategorilere ayırarak, her bir tekniğin amacını, uygulama yöntemini ve etkinliğini detaylı bir şekilde incelemektedir. Özellikle, **profil tabanlı** ve **talimat tabanlı** prompt tasarımı gibi yöntemler, modelin belirli bir rol üstlenmesini veya karmaşık görevleri adım adım yerine getirmesini sağlayarak, çıktı kalitesini ve tutarlılığını önemli ölçüde artırmıştır. Bu tür yapılandırılmış prompt'lar, modelin anlama ve üretim süreçlerini yönlendirerek, daha hedefe yönelik ve güvenilir sonuçlar elde edilmesini sağlar .

Aynı zamanda, **"Reflexive Prompt Engineering"** başlıklı bir başka çalışma, prompt mühendisliğini daha da ileri götürerek, prompt'ların kendi kendini değerlendirmesini ve iyileştirmesini sağlayan yöntemler sunmaktadır . Bu yaklaşım, modelin ürettiği çıktıyı analiz ederek, prompt'un etkinliğini değerlendirmesini ve gerekli düzenlemeleri otomatik olarak yapmasını hedefler. Bu, özellikle dinamik ve değişen görevler için oldukça faydalıdır, çünkü statik prompt'lar zamanla etkisiz hale gelebilir. Bu tür yenilikçi yaklaşımlar, prompt mühendisliğini sadece bir sanat değil, aynı zamanda sistematik ve verimli bir süreç haline getirmeyi amaçlamaktadır. Bu gelişmeler, kullanıcıların LLM'lerle daha etkili bir şekilde iletişim kurmasını ve karmaşık problemleri çözmesini kolaylaştırarak, yapay zekâ teknolojilerinin daha geniş bir kitle tarafından benimsenmesine katkı sağlamaktadır .

4.1.2. Bilgi Entegrasyonu ve Retrieval-Augmented Prompting

Prompt mühendisliğinin önemli bir alt alanı, modelin dış bilgileri etkili bir şekilde kullanmasını sağlamaktır. **Retrieval-Augmented Prompting**, bu amaca yönelik güçlü bir tekniktir. Bu yöntemde, kullanıcının sorusuna cevap vermeden önce, harici bir bilgi kaynağından (örneğin, bir belge veritabanı veya internet) ilgili bilgiler alınır ve bu bilgiler prompt'a dahil edilir. Bu, modelin güncel veya alan özgüsü bilgiye erişmesini sağlar. Ancak, alınan bilgilerin doğru ve bağlamsal olarak uygun olması kritik önem taşır. 2025 yılında yapılan araştırmalar, retrieval sürecinin ve alınan bilgilerin prompt'a entegrasyonunun nasıl optimize edilebileceğini incelemektedir. Örneğin, alınan belgelerin özetlenmesi, önceliklendirilmesi veya yapılandırılmış bir formatta sunulması gibi teknikler, modelin bu bilgileri daha iyi anlamasını ve kullanmasını sağlar. Bu alandaki gelişmeler, RAG sistemlerinin güvenilirliğini ve etkinliğini artırarak, daha bilgili ve güvenilir AI uygulamalarının geliştirilmesine katkı sağlamaktadır.

4.1.3. Reasoning ve Planlama için Prompt Teknikleri

LLM'lerin karmaşık reasoning ve planlama görevlerinde başarılı olması, genellikle özel olarak tasarlanmış prompt tekniklerine bağlıdır. **Chain-of-Thought (CoT)** prompting, modelin adım adım düşünmesini sağlayarak bu alanda önemli bir kırılma noktası olmuştur. 2025 yılında, CoT'nin çeşitli varyasyonları geliştirilmiştir. Örneğin, **Tree of Thoughts** ve **Graph of Thoughts** gibi yöntemler, modelin farklı düşünme yollarını keşfetmesine ve en iyi çözümü bulmasına olanak tanır. Bu teknikler, modelin çıkarım sürecini daha yapılandırılmış ve sistematik hale getirir. Ayrıca, **Reflexion** gibi yöntemler, modelin kendi çıkarımlarını değerlendirmesini ve hatalarını düzeltmesini sağlar. Bu öz-yansıtıcı yaklaşım, modelin planlama ve problem çözme yeteneklerini önemli ölçüde artırır. Bu tür gelişmiş prompt teknikleri, LLM'lerin sadece bilgiyi yeniden üretmekten ziyade, gerçek anlamda akıl yürütmesini ve plan yapmasını sağlayarak, daha güçlü ve otonom AI ajanlarının geliştirilmesine olanak tanır.

4.1.4. Prompt Güvenilirliği ve Tutarlılığı Artırma Yöntemleri

Prompt mühendisliğinde karşılaşılan en büyük zorluklardan biri, model çıktılarının **güvenilirliği ve tutarlılığı**nı sağlamaktır. Modelin aynı prompt'a farklı zamanlarda farklı yanıtlar vermesi veya mantıksal olarak tutarsız çıktılar üretmesi, pratik uygulamalarda ciddi sorunlara yol açabilir. Bu sorunları gidermek için çeşitli yöntemler geliştirilmiştir. **Self-Consistency** tekniği, modelin aynı soruya birden fazla kez farklı varyasyonlarla sorarak, en yaygın yanıtı seçmesini sağlar. Bu, çıktıların tutarlılığını artırır. Ayrıca, **prompt'ların belirli bir çıktı formatına (örneğin, JSON veya XML) bağlanması**, modelin

yapılandırılmış ve işlenebilir çıktılar üretmesini sağlar. Bu, özellikle yazılım geliştirme ve veri analizi gibi alanlarda önemlidir. Ayrıca, prompt'a **açıklayıcı veya neden-sonuç ilişkileri** eklemek, modelin çıkarım sürecini daha şeffaf ve izlenebilir hale getirir. Bu yöntemler, LLM'lerin güvenilirliğini artırarak, bu modellerin kritik uygulamalarda daha güvenli bir şekilde kullanılmasını sağlar.

4.2. Model Geliştirme ve Eğitim Teknikleri

4.2.1. Critical Representation Fine-Tuning (CRFT): Verimli İnce Ayar

2025 yılı itibarıyla, büyük dil modellerinin (LLM) karmaşık akıl yürütme görevlerindeki performansını artırmak için geliştirilen verimli ince ayar (fine-tuning) yöntemleri, yapay zekâ araştırmalarının önemli bir odak noktasıdır. Bu alandaki en dikkat çekici gelişmelerden biri, **Critical Representation Fine-Tuning (CRFT)** adlı yeni bir parametre verimli ince ayar (PEFT) yöntemidir . CRFT, temel olarak Representation Fine-tuning (ReFT) yöntemini temel alır ancak karmaşık akıl yürütme görevlerinde daha etkili olacak şekilde geliştirilmiştir. ReFT, her katmanın başındaki ve sonundaki sabit temsilleri düzenleyerek parametre verimliliği sağlarken, bu sabit konumdaki temsillerin çıktılar üzerindeki etkisi belirsiz olabilir. CRFT, bu sorunu aşmak için karmaşık akıl yürütme görevlerinde önemli rol oynayan **"kritik temsilleri" (critical representations)** tanımlayıp optimize ederek, daha etkili ve verimli bir ince ayar sağlar .

CRFT'nin temel fikri, karmaşık akıl yürütme görevlerinde, önceki katmanlardan önemli bilgileri entegre eden veya sonraki katman temsillerini düzenleyen belirli "kritik temsiller" olduğunu gözlemlemesidir. Bu temsiller, katman katman yayılma süreci boyunca nihai çıktı üzerinde büyük bir etkiye sahiptir. Dolayısıyla, bu kritik temsillerin ince ayarlanması, akıl yürütme performansını büyük ölçüde artırma potansiyeline sahiptir. CRFT, bu kritik temsilleri **bilgi akışı analizi (information flow analysis)** yoluyla tanımlar ve bu temsilleri düşük dereceli bir doğrusal alt uzayda dinamik olarak optimize ederken, temel modeli dondurur. Bu süpervizörlü öğrenme çerçevesi, hem etkinliği hem de verimliliği artırır .

CRFT'nin etkinliği, **LLaMA ve Mistral model aileleri** kullanılarak sekiz farklı aritmetik ve sağduyu akıl yürütme benchmark'ında doğrulanmıştır. Özellikle, CRFT yöntemi, **GSM8K** veri kümesinde **LLaMA-2-7B'nin doğruluğunu %18.2**, ReFT'nin doğruluğunu ise **%3.8 artırırken**, öğrenilen parametre miktarı modelin toplam parametrelerinin sadece **%0.016'sı** kadardır. Bu, CRFT'nin diğer PEFT yöntemlerine kıyasla önemli ölçüde daha verimli olduğunu göstermektedir. Ayrıca, CRFT'nin **few-shot (az örnekli)** ayarlamalara da etkili bir şekilde uyum sağladığı ve **one-shot (tek örnekli)** görevlerde doğruluğu

%16.4 artırdığı gözlemlenmiştir. Bu sonuçlar, temsil düzeyinde optimizasyonun **Chain-of-Thought (CoT)** akıl yürütmesi için büyük bir potansiyele sahip olduğunu ve geleneksel PEFT yöntemlerine göre hafıf ancak güçlü bir alternatif sunabileceğini göstermektedir .

4.2.2. Prompt Güçlendirme ile Uzun Vadeli Planlama

LLM'lerin uzun vadeli planlama yeteneği, otonom ajanlar ve robotik gibi alanlarda kritik önem tasıır. Ancak, mevcut modeller genellikle kısa vadeli, tepkisel davranıslar sergileme eğilimindedir. Bu sınırlılığı asmak için, **prompt güçlendirme (prompt boosting)** adı verilen yeni bir teknik gelistirilmistir. Bu yöntem, modelin uzun vadeli hedeflere odaklanmasını sağlamak için prompt'u çesitli sekillerde güçlendirir. Örneğin, prompt'a uzun vadeli bir "misyon" veya "vizyon" ifadesi eklenerek, modelin kararlarını bu hedeflere göre sekillendirmesi tesvik edilir. Ayrıca, modelin geçmis eylemlerini ve sonuçlarını prompt'a dahil eden **hafıza tabanlı prompt'lar**, modelin uzun vadeli bir bağlam içinde hareket etmesini sağlar. Bu teknikler, modelin sadece anlık ödülleri değil, aynı zamanda uzun vadeli sonuçları da göz önünde bulundurmasını sağlar. Prompt güçlendirme, uzun vadeli planlama görevlerinde model performansını önemli ölçüde artırarak, daha akıllı ve stratejik AI sistemlerinin gelistirilmesine katkı sağlar.

4.2.3. Yapay Zekâ Hizalamasının Ilaç Kesfi Uygulamaları

Yapay zekâ hizalaması, sadece dil modellerinin etik ve güvenli davranısları için değil, aynı zamanda özel uygulama alanlarında da kritik önem tasıır. **Ilaç kesfi** gibi yüksek riskli ve düzenlemeye tabi bir alanda, AI modellerinin güvenilirliği ve tutarlılığı hayati önem tasıır. Bu bağlamda, hizalama teknikleri, modellerin biyolojik ve kimyasal verileri doğru bir sekilde yorumlamasını ve güvenli olmayan bilesikleri tanımlamasını sağlamak için kullanılmaktadır. Örneğin, RLHF yöntemleri, modelin toksisite veya yan etki tahminlerini, insan uzmanların geri bildirimleriyle hizalamak için kullanılabilir. Ayrıca, hizalama süreçleri, modellerin düzenleyici gereksinimlere (örneğin, FDA kuralları) uygun sekilde çalısmasını sağlamak için de kullanılır. Bu uygulamalar, AI'nın ilaç kesfi süreçlerinde güvenli ve etkili bir sekilde kullanılmasını sağlayarak, hem maliyetleri düsürmeye hem de basarı olasılığını artırmaya katkı sağlar.

5. Endüstri Perspektifi ve Model Gelisimleri

5.1. Ekim 2025'te Öne Çıkan AI Modelleri

5.1.1. GPT-5, Gemini Pro 2.5, Claude Sonnet 4.5

1 Ekim 2025 ile 1 Kasım 2025 tarihleri arasında, yapay zekâ endüstrisinde önde gelen şirketler, daha güçlü ve yetenekli büyük dil modellerini (LLM) piyasaya sürmeye devam etmiştir. Bu dönemde, özellikle bağlam penceresi uzunluğu, çıkarım yetenekleri ve çoklu modal desteği gibi alanlarda önemli ilerlemeler kaydedilmiştir. Bu gelişmeler, hem akademik araştırmalara hem de endüstriyel uygulamalara yeni olanaklar sunmaktadır. Örneğin, **"Top 10 AI Models with the Longest Context Windows (2025)"** başlıklı bir endüstri raporu, bu dönemde piyasaya sürülen ve en uzun bağlam penceresine sahip olan modelleri sıralamıştır . Bu sıralama, modellerin teknik özelliklerini ve potansiyel kullanım alanlarını öne çıkararak, kullanıcıların ve geliştiricilerin ihtiyaçlarına en uygun modeli seçmelerine yardımcı olmayı amaçlamaktadır.

Bu rapora göre, Kasım 2025 itibarıyla en uzun bağlam penceresine sahip modeller arasında Google'ın **Gemini 2.5 Pro** ve **Gemini 2.5 Flash** modelleri, **1.000.000 tokenlık** devasa bir bağlam penceresi ile öne çıkmaktadır . Bu modeller, özellikle büyük çoklu-doküman işleme, kurumsal ölçekli iş yükleri ve genomik veri analizi gibi alanlarda güçlü birer aday olarak konumlandırılmaktadır. Aynı dönemde, OpenAI'nin **GPT-5** ve **GPT-5 Codex** modelleri de **200.000 tokenlık** bağlam penceresi ile dikkat çekmektedir. GPT-5, genel amaçlı sofistike çıkarım için, GPT-5 Codex ise programlama problemleri ve büyük kod tabanları üzerinde çalışmak için optimize edilmiştir . Anthropic firmasının Claude ailesinden **Claude Sonnet 4.5 (20250929)** ve **Claude Haiku 4.5 (20251001)** modelleri de **200.000 tokenlık** bağlam penceresi sunarak, hem verimlilik hem de çıkarım gücü açısından rekabetçi bir konumdadır . Bu modeller, özellikle özetleme ve anlatı odaklı görevlerde etkili performans sergilemektedir.

5.1.2. Grok 4 ve DeepSeek 3.1-Terminus

Yapay zekâ endüstrisindeki rekabet, sadece büyük teknoloji şirketleriyle sınırlı değildir. Açık kaynak topluluğu ve yeni girişimler de bu alanda önemli katkılar sağlamaktadır. Bu dönemde, **Grok 4** ve **DeepSeek 3.1-Terminus** gibi modeller, özgün yetenekleri ve rekabetçi performanslarıyla dikkat çekmiştir. Grok 4, özellikle gerçek zamanlı bilgi erişimi ve mizahi yanıt üretme yetenekleriyle öne çıkmaktadır. Bu model, X (eski adıyla Twitter) platformuyla entegre çalışarak, güncel olaylara dair bilgileri anında işleyebilir ve kullanıcılarla etkileşime geçebilir. DeepSeek 3.1-Terminus ise, açık kaynak topluluğu tarafından geliştirilmiş ve kodlama ve matematiksel problem çözme gibi teknik görevlerde güçlü performans sergilemiştir. Bu modeller, endüstrideki çeşitliliği ve inovasyonu artırarak, kullanıcılara daha fazla seçenek sunmaktadır. Ayrıca, bu modellerin açık kaynak olarak sunulması, araştırmacıların ve geliştiricilerin bu teknolojilere erişimini kolaylaştırarak, alandaki ilerlemeyi hızlandırmaktadır.

5.2. En Uzun Bağlam Penceresine Sahip Modeller

5.2.1. Bağlam Penceresi Boyutlarına Göre Model Sıralaması

Kasım 2025 itibarıyla, yapay zekâ endüstrisinde bağlam penceresi uzunluğu konusunda ciddi bir rekabet yaşanmaktadır. Bu rekabet, modellerin daha fazla bilgiyi aynı anda işleyebilme ve bu bilgiler arasında karmaşık ilişkiler kurabilme yeteneklerini doğrudan etkilemektedir. **"Top 10 AI Models with the Longest Context Windows (2025)"** başlıklı endüstri raporu, bu rekabet ortamında öne çıkan modelleri sıralamıştır . Bu sıralama, modellerin teknik özelliklerini ve potansiyel kullanım alanlarını öne çıkararak, kullanıcıların ve geliştiricilerin ihtiyaçlarına en uygun modeli seçmelerine yardımcı olmayı amaçlamaktadır.

表格 复制

Model	Geliştirici	Bağlam Penceresi (Token)	Öne Çıkan Özellik
Gemini 2.5 Pro	Google	1,000,000	Çoklu-doküman işle
Gemini 2.5 Flash	Google	1,000,000	Verimlilik odaklı, bü
GPT-5	OpenAI	200,000	Genel amaçlı sofist
GPT-5 Codex	OpenAI	200,000	Programlama ve büy
Claude Sonnet 4.5	Anthropic	200,000	Özetleme ve anlatı (
Claude Haiku 4.5	Anthropic	200,000	Verimlilik ve hız oda
Grok 4	xAI	128,000	Gerçek zamanlı bilg
DeepSeek 3.1-Terminus	DeepSeek AI	128,000	Kodlama ve matema

Tablo 1: Kasım 2025 itibarıyla en uzun bağlam penceresine sahip önde gelen AI modelleri .

Bu sıralama, bağlam penceresi boyutunun yanı sıra, modellerin uzmanlaştığı alanları da göstermektedir. Örneğin, **Google'ın Gemini 2.5** serisi, 1 milyon tokenlık devasa bağlam penceresiyle, büyük çaplı belge analizi ve genomik veri işleme gibi alanlarda öne çıkarken, **OpenAI'nin GPT-5 Codex** modeli, 200.000 tokenlık bağlam penceresini programlama görevleri için optimize etmiştir. Benzer şekilde, **Anthropic'in Claude** modelleri, özetleme ve anlatı odaklı görevlerde güçlü performans sergilerken, **Grok 4**

gibi modeller, gerçek zamanlı bilgi erişimi gibi farklı bir alanda uzmanlaşmıştır. Bu çeşitlilik, kullanıcıların kendi ihtiyaçlarına en uygun modeli seçmelerine olanak tanır.

5.2.2. Uzun Bağlamın Önemi ve Token Kavramı

Bir dil modelinin **bağlam penceresi (context window)** , modelin tek bir çıkarım (inference) işleminde işleyebileceği maksimum token sayısıdır. **Token**, kelimelerin veya kelime parçalarının sayısal temsilleridir. Örneğin, "Merhaba dünya" ifadesi 2 token olabilirken, "Merhaba" kelimesi tek başına 1 token olabilir. Bağlam penceresi, modelin anlama ve üretim yeteneklerini doğrudan etkiler. Daha uzun bir bağlam penceresi, modelin:

- **Daha büyük belgeleri analiz etmesini** sağlar. Örneğin, bir kitabın tamamını veya uzun bir raporu tek seferde özetleyebilir.
- **Uzun diyalogları sürdürmesini** mümkün kılar. Sohbet geçmişini daha iyi hatırlayarak, daha tutarlı ve bağlamsal olarak zengin yanıtlar verebilir.
- **Karmaşık görevlerde daha iyi performans göstermesini** sağlar. Çok adımlı problem çözme veya kodlama gibi görevlerde, tüm gerekli bilgileri aynı anda görebilir.

2025 yılı itibarıyla, bağlam penceresi boyutundaki rekabet, modellerin yeteneklerini ölçmede önemli bir metrik haline gelmiştir. Ancak, sadece büyük bir bağlam penceresi yeterli değildir. Modelin, bu uzun bağlamı etkili bir şekilde kullanabilmesi, yani **uzun bağlam reasoning** yeteneği de kritik önem taşır. Bu nedenle, akademik araştırmalar, modellerin bağlam penceresi boyutunun yanı sıra, bu bağlamı anlama ve üzerinde çıkarım yapma yeteneklerini değerlendirmeye odaklanmaktadır.