

Tıbbi Akıl Yürütme için Gemma 3 Modellerinin CoT Destekli SFT ve GRPO ile İncelenmesi

İsmail İŞERİ^{1,a}, Alper YILDIRIM^{2,b}, Alihan ÖZTORUN^{2,c}, Tuğba TUNA^{3,d},
Arda TURAN^{3,e}

¹Ondokuzmayıs Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Samsun, Türkiye

²Düzce Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Düzce, Türkiye

³Arena Bilgisayar A.Ş., İstanbul, Türkiye

^aORCID: 0000-0002-0442-1406; ^bORCID: 0009-0004-8029-1772; ^cORCID: 0009-0006-2931-4945;

^dORCID: 0009-0002-0352-593X; ^eORCID: 0009-0007-1145-2888

Makale Bilgileri

Geliş : 01.08.2025

Kabul : 09.09.2025

DOI: 10.21605/cukurovaumfd.1755684

Sorumlu Yazar

İsmail İŞERİ

ismail.iseri@omu.edu.tr

Anahtar Kelimeler

Büyük dil modelleri

Tıbbi akıl yürütme

Düşünce zinciri

Grup görelisi politika optimizasyonu

Özet: İŞERİ, İ., YILDIRIM, A., ÖZTORUN, A., TUNA, T., TURAN, A., (2025). Tıbbi Akıl Yürütme için Gemma 3 Modellerinin CoT Destekli SFT ve GRPO ile İncelenmesi. Çukurova Üniversitesi, Mühendislik Fakültesi Dergisi, 40(3), 593-606.

ÖZ

Bu çalışmada, Gemma 3 1B ve Gemma 4B büyük dil modellerinin tıbbi alandaki karmaşık akıl yürütme yeteneklerinin geliştirilmesi ve değerlendirilmesi amaçlanmıştır. Bu kapsamda, SFT (Gözetimli İnce Ayar) ve GRPO (Grup Görelisi Politika Optimizasyonu) gibi eğitim stratejilerinin Gemma 3 1B ve Gemma 4B modelleri üzerindeki performansı incelenmiştir. Temel modellerin değerlendirilmesi sürecinde, SFT ile Düşünce Zinciri (CoT) formatının öğretilmesi ve GRPO ile akıl yürütmenin rafine edilmesi adımlarını içeren çok aşamalı bir yaklaşım izlenmiştir. GPT-4.1 hakem modeliyle yapılan değerlendirmeler sonucunda, model performansının belirgin şekilde artırıldığı gösterilmiştir. SFT ve GRPO eğitimleriyle modelin mantıksal olarak tutarlı bir akıl yürütme süreci oluşturma becerisinin (Mantık Yürütme Doğruluğu'nun %26'dan %31'e yükseltilmesiyle) başarıyla geliştirildiği ortaya konmuştur. Bu durum, modelin cevapları ezberlemek yerine nasıl düşüneceğinin öğrenildiğini kanıtlamaktadır.

Investigation of Low-Parameter Gemma 3 Models for Medical Reasoning Using CoT-Supported SFT and GRPO

Article Info

Received : 01.08.2025

Accepted : 09.09.2025

DOI: 10.21605/cukurovaumfd.1755684

Corresponding Author

İsmail İŞERİ

ismail.iseri@omu.edu.tr

Keywords

Large language models

Medical reasoning

Chain-of-thought (CoT)

Group relative policy optimization (GRPO)

How to cite: İŞERİ, İ., YILDIRIM, A., ÖZTORUN, A., TUNA, T., TURAN, A., (2025). Investigation of Low-Parameter Gemma 3 Models for Medical Reasoning Using CoT-Supported SFT and GRPO. Çukurova University, Journal of the Faculty of Engineering, 40(3), 593-606.

ABSTRACT

This study aimed to develop and evaluate the complex reasoning capabilities of the Gemma 3 1B and Gemma 4B large language models within the medical domain. In this context, the performance of training strategies such as SFT (Supervised Fine-Tuning) and GRPO (Group Relative Policy Optimization) on the Gemma 3 1B and Gemma 4B models was investigated. A multi-stage approach was followed, starting with the evaluation of the base models, then teaching the Chain-of-Thought (CoT) format via SFT, and finally refining the reasoning process with GRPO. Evaluations conducted using the GPT-4.1 as a judge model demonstrated a significant improvement in model performance. It was shown that SFT and GRPO training successfully enhanced the model's ability to generate a logically consistent reasoning process, evidenced by an increase in Reasoning Accuracy from 26% to 31%. This outcome proves that the model learned how to think rather than merely memorizing answers.

1. GİRİŞ

Büyük dil modelleri (Large Language Model - LLM), klinik karar destek sistemlerinden hasta eğitime kadar uzanan geniş bir yelpazede sundukları potansiyel ile sağlık hizmetleri alanında devrim yaratma potansiyeline sahip olmaktadır. Literatürdeki çalışmalarda, çeşitli GPT tabanlı modellerin; tıbbi bilgi çıkarımı, teşhis desteği ve hastalık tahmini gibi spesifik görevlerdeki başarısı ortaya konmuştur. Benzer şekilde, küçük dil modelleri (Small Language Model- SLM) ise maliyet etkinliği ve veri gizliliği avantajlarıyla öne çıkarılmıştır. Akıl yürütme yeteneklerinin geliştirilmesi amacıyla çalışmalarda takviyeli öğrenme gibi gelişmiş yöntemlerin kullanıldığı ve modellerin doğruluğunu artırmaya odaklanıldığı görülmektedir. LLM'lerin tıbbi alandaki yeteneklerini artırma literatürü içindeki özgün katkımız, HuatuoGPT [1] ve MedReason [2] gibi yaklaşımların ürettiği akıl yürütme verilerinin güvenilirliğini ölçmek ve doğrulamak için temel bir mekanizma ve metodoloji sunarak bu "kendini iyileştirme döngüsü"nün kalitesini yükseltmeye odaklanmaktadır.

Ancak mevcut çalışmaların büyük bir kısmında, model performansı yalnızca nihai "Cevap Doğruluğu (Answer Accuracy)" metriği üzerinden ölçülmüştür. Bu yaklaşım, karmaşık ve açık uçlu tıbbi problemlerin değerlendirilmesinde modelin gerçek yeteneğini ölçmekte yetersiz kalabilmektedir. Bunun nedeni, bu tür problemlerin birden fazla klinik olarak geçerli yolun veya cevabın bulunmasına olanak sağlamasıdır. Ayrıca, bir modele doğru cevaba şans eseri veya mantıksal olmayan yollarla ulaşma imkânının verilmesi, klinik güvenilirlik açısından ciddi bir risk olarak değerlendirilmektedir. Mevcut çalışmalar, Bilgi Grafikleri bütünleştiren MedReason [2] gibi yaklaşımlarla sentetik veri kalitesini sürekli artıran bir "kendini iyileştirme döngüsü" üzerine kuruludur. Ancak, bu döngünün en kritik sorunu, üretilen akıl yürütmenin kalitesini ölçme problemidir.

Bu çalışmada, literatürdeki bu boşluğun doldurulması hedeflenmiştir. Araştırmamızda, birçok çalışmadan farklı olarak yalnızca modelin doğru sonuca ulaşıp ulaşmadığına değil; aynı zamanda bu sonuca ulaşılan Düşünce Zinciri (Chain-of-Thought - CoT) sürecinin mantıksal tutarlılığına ve tıbbi geçerliliğine başarıyla ölçütü olarak odaklanılmıştır. Çalışmamız, yalnızca nihai sonuca değil, sürecin tıbbi sağlamlığına dayanan "Akıl Yürütme Doğruluğu" adlı yeni bir metrik önererek, davranışsal eğitimin modelin problem çözme güvenilirliğini nasıl artırdığını kantitatif olarak göstermektedir.

Bu amaçla, Gemma 3 ailesinden farklı ölçekteki (1B ve 4B) modellerin performansı; temel (baseline) durumdan başlanarak, Gözetimli İnce Ayar (Supervised Fine-Tuning - SFT) ile davranışsal formatın kazandırılması ve akıl yürütmelerinin rafine edilmesi aşamalarından geçirilerek sistematik biçimde analiz edilmektedir.

Çalışmamızda ortaya konan temel hipotez, SFT ve GRPO gibi eğitimlerin, cevap doğruluğunda bir artış sağlanmasa bile modelin nasıl düşüneceğinin öğrenilmesini mümkün kılarak Mantık Yürütme Doğruluğunun (Reasoning Accuracy) geliştirilebileceğidir. Bu sayede, tıbbi uygulamalarda daha güvenilir ve yorumlanabilir modellerin geliştirilmesinin önemi vurgulanmıştır.

1.1. İlgili Çalışmalar

Büyük dil modellerinin (LLM) sağlık ve biyomedikal alanlardaki en kritik gelişim alanlarından biri, karmaşık akıl yürütme yeteneklerinin artırılmasıdır. Bu amaçla araştırmacılar, takviyeli öğrenme ve özel eğitim gibi çeşitli teknikler geliştirmektedir. Örneğin, Junying Chen ve arkadaşları [1], doğrulayıcı destekli eğitimle HuatuoGPT-O1 modelini geliştirerek bu alandaki etkinliği kanıtlamıştır. Benzer bir yaklaşımla Yuxiang Lai ve arkadaşları [3], Med-R1 adlı pekiştirmeli öğrenme yöntemiyle GPT-4 dahil birçok modeli geride bırakan ve uzman hekimlerce daha çok tercih edilen başarılı sonuçlar elde etmiştir. Bu alandaki bir diğer önemli gelişme ise, Juncheng Wu ve arkadaşlarının [2] çalışması olan büyük dil modellerinin tıbbi akıl yürütme becerisini geliştirmek için tıbbi bilgi graflerinden (knowledge graph) yararlanarak olgusal düşünce zincirleri (Chain-of-Thought) üreten MedReason metodolojisidir. Bu yaklaşımla eğitilen modeller, performanslarını önemli ölçüde artırarak kendi sınıflarında en son teknoloji (state-of-the-art) sonuçlar elde etmiştir. Akıl yürütme performansını artıran diğer önemli çalışmalar arasında DeepSeek-R1 [4] modeli, Fei Ding ve arkadaşlarının [5] geliştirdiği iki katmanlı MGRPO yöntemi ve Robert Tinn ile ekibinin [6] ince ayar (fine-tuning) sürecindeki kararsızlığı azaltan teknikleri yer almaktadır.

Modellerin akıl yürütme yeteneklerini objektif bir şekilde ölçmek ve değerlendirmek de bu sürecin ayrılmaz bir parçasıdır. Bu doğrultuda, Wei Dai ve arkadaşları [7], uzman hekimlerce hazırlanan 1.719 soruluk yüksek kaliteli QoQ-Med benchmark setini sunmuştur. Bu testler, GPT-4 ve Claude-3 gibi gelişmiş modellerin başarısını teyit ederken, bazı açık kaynaklı modellerin güvenilirlik açısından eksik kaldığını göstermiştir. Benzer şekilde, Mianxin Liu ve takım arkadaşları [8], Çin tıbbına özgü multimodal LLM'leri değerlendirmek amacıyla MedBench'i geliştirmiş ve bu alanda da ChatGPT gibi modellerin yüksek performansını ortaya koymuştur.

Mevcut modellerin performansını karşılaştıran çalışmalar, LLM'lerin potansiyelini ve sınırlarını gözler önüne sermektedir. Cagatay Umut Ogdu ve arkadaşları [9], MedMCQA ve PubMedQA veri setleri üzerinde yaptıkları testlerde, özellikle GPT-4'ün ROUGE skorları ve Semantik Benzerlik ölçümlerinde üstün olduğunu belirtmiştir. Ancak, Hazrat Ali ve arkadaşları [10] ile Jin Li ve ekibinin [11] de vurguladığı gibi, bu modellerin tıp sınavlarındaki başarı potansiyeline rağmen hata ve yanlışlık riskleri bulunmaktadır. Bu nedenle, GPT-4 ve Gemini gibi en güçlü modellerin dahi klinik kullanıma sunulmadan önce kapsamlı bir doğrulamadan geçmesi kritik bir zorunluluktur.

Büyük dil modelleri (LLM'ler), sağlık sektöründe klinik dil anlama, bilgiye erişim ve karar destek sistemleri gibi alanlarda devrim yaratma potansiyeli taşımaktadır. Bu alanda, Xi Yang ve arkadaşlarının [12] geliştirdiği GatorTron gibi modeller, karmaşık klinik doğal dil işleme görevlerinde yüksek performans göstermektedir.

LLM'lerin çok yönlü uygulamaları, birçok çalışma ile kanıtlanmıştır. Örneğin, Rajendra Kumar ve ekibi [13] LLM ve NLP entegrasyonu ile klinik karar desteği sunarken, Sirui Ding ve arkadaşları [14] CKLE yöntemiyle çok modlu EHR verilerinden sağlık olaylarını tahmin etmiştir. Cai Wang ve ekibinin [15] KEMedGPT'si tıbbi danışmanlıkta, Cheng Peng ve arkadaşlarının [16] GatorTronGPT'si biyomedikal ilişki çıkarımında, Laila Rasmy ve çalışma arkadaşlarının [17] Med-BERT'i ise hastalık tahmininde önemli başarılar elde etmiştir. Ayrıca, M. Arif Rahman ve ekibi [18] Acil Sağlık Hizmetleri'ne özgü EMS-BERT'i, Wenqiang Li ve arkadaşları [19] Çince tıbbi soru-cevaplar için DoctorGPT'yi ve Akilesh S. ile ekibi [20] kullanıcı dostu bir tıbbi sohbet botu geliştirmiştir. Georgios Peikos ve çalışma arkadaşları [21] ise küçük, açık kaynaklı LLM'lerin klinik deney eşleştirmesinde bile etkili olabildiğini göstermiştir.

Kişiselleştirilmiş tıp alanında Dr. Kumar P. ve ekibinin [22] Medify-AI sistemi ile Dr. Rakesh Kumar M ve ekibinin [23] MedHub projesi gibi kullanıcı dostu sistemler geliştirilmiştir. Shadi AlZu'bi ve arkadaşları [24], LLM'leri Makine Öğrenmesi (Machine Learning) ve Derin Öğrenme (Deep Learning) ile entegre ederek tanı doğruluğunu ve tedavi önerilerini iyileştirmiştir. Son olarak, Fernando S.P.A.A. ve Dilani Wickramaarachchi [25], sağlık hizmetine erişimi kısıtlı bölgeler için bir ön tıbbi danışmanlık platformu oluşturmuştur.

Bununla birlikte, Zahir Al Nazi ve Wei Peng [26] ile Shubo Tian ve arkadaşlarının [27] da vurguladığı gibi, bu büyük potansiyele rağmen veri gizliliği, önyargıların azaltılması ve etik gibi zorluklar halen çözüm beklemektedir.

Küçük dil modelleri (SLM'ler), büyük dil modellerinin (LLM) getirdiği yüksek maliyet ve yoğun kaynak ihtiyacı gibi zorluklara karşı sağlık alanında pratik ve erişilebilir bir çözüm olarak öne çıkmaktadır. Bu modeller, ince ayar (fine-tuning) gibi özel eğitim yöntemleriyle belirli tıbbi görevler için optimize edilebilir. Böylece hem veri gizliliği ve güvenlik avantajları sunar hem de daha verimli bir performans sergilerler.

Bu verimli uyarlamayı sağlayan yöntemlerin başında Edward Hu ve arkadaşlarının [28] yapmış olduğu çalışma olan LoRA (Low-Rank Adaptation) gelmektedir. Edward Hu ve arkadaşları tarafından önerilen bu yöntemde, büyük dil modellerini uyarlamak için ana model ağırlıkları dondurulup yalnızca küçük, "düşük rütbeli" (low-rank) matrisler eğitilir. Bu yaklaşım, eğitilebilir parametre sayısını ve bellek kullanımını büyük ölçüde azaltırken, tam yeniden eğitime (full fine-tuning) ile aynı veya daha iyi performansı, çıkarım sırasında ek gecikme olmadan sağlar.

Bu alandaki çalışmalar, SLM'lerin potansiyelini somutlaştırmaktadır. Örneğin, Aditya Karnam Gururaj Rao ve arkadaşları [29], LLama 3.2 3B gibi küçük modellerin ince ayar ile sağlık sohbet botlarında ne kadar etkili olabileceğini göstermiştir. Benzer şekilde, Hyunjae Kim ve çalışma arkadaşlarının [30] geliştirdiği Meerkat-7B ve Meerkat-8B modelleri, tıbbi akıl yürütme konusunda GPT-4'e yakın bir performans sergileyerek güvenli ve erişilebilir alternatifler sunmaktadır.

Gizlilik konusundaki potansiyeli ise Matteo Magnini ve arkadaşları [31], yerel olarak çalıştırılan SLM'lerin tıbbi sohbet botları için güvenli senaryolar oluşturabileceğini belirterek vurgulamıştır. SLM'lerin bir diğer önemli kullanım alanı ise mobil sağlıktır. Xin Wang ve arkadaşları [32], bu modellerin mobil sağlık takip uygulamalarında düşük gecikme ve az kaynak tüketimiyle ne kadar verimli çalıştığını kanıtlamıştır.

1.2. Çalışmamızın Literatürdeki Yeri ve Katkısı

Yukarıda özetlenen çalışmalar, tıbbi alanda LLM'lerin artan yeteneklerini ve çeşitliliğini ortaya koymaktadır. Bizim çalışmamız, bu geniş literatür içinde, özellikle akıl yürütme kalitesinin üretimi, eğitimi ve en önemlisi ölçümü konularına odaklanarak özgün bir katkı sunmaktadır.

Mevcut literatür, daha güçlü temel modelleri bir "jüri" olarak kullanarak sürekli daha kaliteli sentetik veri üreten bir "kendini iyileştirme döngüsü" üzerine kurulmuştur. Bu döngüde, HuatuoGPT-o1 [1] gibi çalışmalar, davranışsal bir süreçle akıl yürütme verisi üretirken; MedReason [2] gibi daha yeni yaklaşımlar, Bilgi Grafikleri (Knowledge Graph) entegre ederek üretilen verinin faktörel doğruluğunu artırmış ve döngünün kalitesini bir üst seviyeye taşımıştır. Bu veri zenginleştirme yaklaşımları, modelin bilgi tabanını genişletirken, üretilen akıl yürütme sürecinin içsel kalitesini ve güvenilirliğini ölçme sorununu da beraberinde getirmektedir.

Bizim çalışmamız, bu kendini iyileştirme döngüsünün en kritik noktasına, yani kalite kontrol ve davranışsal etki'nin ölçümüne odaklanmaktadır.

1. Metrik Olarak Akıl Yürütme Kalitesine Odaklanma: Diğer çalışmalardan farklı olarak, biz başarının birincil ölçütü olarak, bu sentetik veri üretim hatlarının çıktılarını sadece nihai sonuca göre değil, aynı zamanda sürecin kendisinin tıbbi olarak ne kadar sağlam olduğuna göre değerlendiren yeni bir standart olan Akıl Yürütme Doğruluğu metriğini öneriyor ve uyguluyoruz. Bu metrik, "şans eseri doğru" cevapları, "faydalı ama teknik olarak yanlış" olan klinik olarak değerli çıktılardan ayırt etmemizi sağlayarak, bir modelin gerçek güvenilirliğini ölçmek için daha hassas bir yöntem sunmaktadır.

2. Davranışsal Eğitimin Etkisini Ölçme: SFT ve GRPO gibi eğitimlerin modele yeni tıbbi bilgi "ezberletmekten" ziyade, onun problem çözme "davranışını" değiştirdiğini ve bu davranışsal değişimin, modelin güvenilirliğini doğrudan artırdığını kantitatif olarak kanıtıyoruz.

Özetle, çalışmamız mevcut literatüre, HuatuoGPT [1] ve MedReason [2] gibi veri zenginleştirme yaklaşımlarının ürettiği akıl yürütmenin kalitesini ölçmek ve doğrulamak için yeni bir metrik ve metodoloji sunarak katkıda bulunmaktadır. Bu yaklaşım, söz konusu "kendini iyileştirme döngüsü"nün sadece daha fazla veri üretmesini değil, aynı zamanda kanıtlanabilir şekilde daha güvenilir veri üretmesini sağlayacak temel bir mekanizmadır.

2. MATERYAL VE YÖNTEM

Bu çalışmada, büyük dil modellerinin (LLM) tıbbi alandaki karmaşık akıl yürütme yeteneklerinin geliştirilmesi ve değerlendirilmesi amacıyla çok aşamalı bir eğitim ve analiz metodolojisi izlenmiştir. Araştırmanın temel amacı doğrultusunda, model ölçeğinin (1B ve 4B parametre), farklı eğitim stratejilerinin (SFT ve GRPO) ve görev türlerinin (RAG ve doğrudan sorgu) model performansı üzerindeki etkisi sistematik olarak ortaya konulmuştur. Nihai hedef olarak, yalnızca doğru cevabın bulunması değil, aynı zamanda bu cevaba giden Düşünce Zinciri (Chain-of-Thought - CoT) sürecinin mantıksal ve tutarlı bir şekilde oluşturulabilme becerisinin ölçülmesi amaçlanmıştır.

2.1. Modeller ve Veri Setleri

2.2.1. Temel Modeller ve Eğitim Tekniği

Çalışmada, model ölçeğinin etkisini karşılaştırmak amacıyla Gemma 3 ailesinden 1B ve 4B parametrelili modeller temel olarak kullanılmıştır. Tüm eğitim süreçlerinde, hesaplama verimliliği sağlamak amacıyla LoRA (Low-Rank Adaptation) tekniği, Unsloth kütüphanesi aracılığıyla uygulanmıştır. Bu sayede modellerin ana ağırlıkları dondurulmuş ve yalnızca küçük adaptör katmanları eğitilmiştir.

3.2.2. Eğitim ve Değerlendirme Veri Setleri

Çalışmada, her biri farklı bir amaca hizmet eden dört temel veri seti veya bunlardan türetilen veri setleri kullanılmıştır [33-37].

Tablo 1. Çalışmada kullanılan eğitim ve değerlendirme veri setleri

Veri seti adı	Kaynak	Seçilen örnek sayısı	Kullanım amacı
SFT eğitim seti	FreedomIntelligence/ medical-o1-reasoning-SFT	~19,700	Gözetimli ince ayar (SFT)
GRPO eğitim seti	Yujivus/mmlu_health_grpo_gsm8k_format	~1,200	Pekiştirmeli öğrenme (GRPO)
Değerlendirme seti 1	PubMedQA	100	Değerlendirme (RAG)
Değerlendirme seti 2	FreedomIntelligence/ medical-o1-verifiable-problem	100	Değerlendirme (Doğrudan sorgu)

Eğitim Veri Setleri

Model eğitiminde iki aşamalı bir yaklaşım benimsenmiştir:

Gözetimli İnce Ayar (SFT) Aşaması: Modelin temel davranışsal formatını, yani bir soruya cevap vermeden önce adım adım akıl yürütme (Chain-of-Thought) sürecini sergilemesini öğretmek amacıyla, FreedomIntelligence/medical-o1-reasoning-SFT veri setindeki yaklaşık 19,700 örnek üzerinde eğitim yapılmıştır.

Pekiştirmeli Öğrenme (GRPO) Aşaması: SFT ile temel düşünme becerisini kazanan modelin, özellikle tıbbi alandaki akıl yürütme kalitesini daha da rafine etmek için pekiştirmeli öğrenme uygulanmıştır. Bu aşamada, MMLU veri setinin tıp, biyoloji ve sağlıkla ilgili alt kümelerinden özel olarak hazırladığımız ve Yujivus/mmlu_health_grpo_gsm8k_format adıyla topluluğun kullanımına sunduğumuz, yaklaşık 1,200 soruluk bir veri seti kullanılmıştır.

Değerlendirme Veri Setleri

Modelin performansını farklı senaryolarda ölçmek için iki ana değerlendirme veri seti kullanılmıştır:

RAG Görevi (PubMedQA): Modelin, kendisine sunulan bir metin (bağlam) içerisinde akıl yürüterek cevap çıkarma, yani Geri-Getirme Destekli Üretim (Retrieval-Augmented Generation - RAG) yeteneğini ölçmek amacıyla standart PubMedQA benchmark'ının test seti kullanılmıştır.

Doğrudan Sorgu Görevi (Doğrulanabilir Problemler veya FreedomIntelligence/ medical-o1-verifiable-problem): Çalışmamızın ana odağı olan, modele herhangi bir ek bağlam verilmeden, doğrudan kendi içsel bilgisi ve muhakeme yeteneğiyle karmaşık tıbbi problemleri çözme becerisini ölçmek için FreedomIntelligence/medical-o1-verifiable-problem veri seti kullanılmıştır.

Veri Kirliliğinin Önlenmesi (Data Decontamination): Modelin genelleme yeteneğini adil bir şekilde ölçmek hayati önem taşıdığından, "Doğrudan Sorgu Görevi" için kullanılan bu nihai test setinin, SFT ve GRPO

eğitim aşamalarında kullanılan verilerle hiçbir kesişime sahip olmadığı garanti edilmiştir. Test seti, medical-o1-verifiable-problem setinden, SFT setindeki tüm soruların çıkarılmasıyla oluşturulmuştur. Bu titiz ayırım, modelin öğrendiği becerilerin transfer edilebilirliğinin ölçülmesi için güvenilir bir zemin sağlamaktadır.

Doğrulanabilir Problemler Veri Setinde Kullanılan Soru Örneği: Based on the chest radiograph and abdominal CT scan of a middle-aged male complaining of nagging abdominal pain for the past 2 weeks, what is the probable diagnosis that should be considered?

Cevabı: Hydatid Cyst.

3.3. Eğitim Aşamaları ve Metodolojisi

Aşama 1 - Temel Model Performansının Belirlenmesi: Eğitimden önce, 1B ve 4B parametrelili Gemma 3 temel modelleri, her iki değerlendirme veri seti üzerinde test edilerek "sıfırıncı gün" performansları kaydedilmiştir. Bu işlemle, sonraki aşamalarda gerçekleştirilecek gelişimin ölçülmesi için kritik bir başlangıç noktası sağlanmıştır.

Aşama 2 - SFT ile Davranışsal Öğrenme: Modelin temel davranışsal formatını, yani bir soruya cevap vermeden önce adım adım akıl yürütme (Chain-of-Thought - CoT) sürecini sergilemesini öğretmek amacıyla, FreedomIntelligence/medical-o1-reasoning-SFT veri seti kullanılarak gözetimli ince ayara tabi tutulmuştur. Bu aşamanın birincil amacı modele yeni tıbbi bilgiler öğretmek değil, karmaşık problemleri yapılandırılmış bir CoT formatında sunma becerisini kazandırmaktır.

Aşama 3 - GRPO ile Akıl Yürütmenin Rafine Edilmesi: SFT ile temel düşünme becerisini kazanan modelin akıl yürütme kalitesini daha da rafine etmek için Pekiştirmeli Öğrenme (GRPO) uygulanmıştır. Bu aşamada, Yujivus/mmlu_health_grpo_gsm8k_format veri seti kullanılmıştır. Eğitim sırasında, modelin hem yapısal formatı korumasını hem de doğru cevabı vermesini teşvik eden çok bileşenli bir ödül sistemi tasarlanmıştır. Bu ödül fonksiyonu üç ana kısımdan oluşmaktadır:

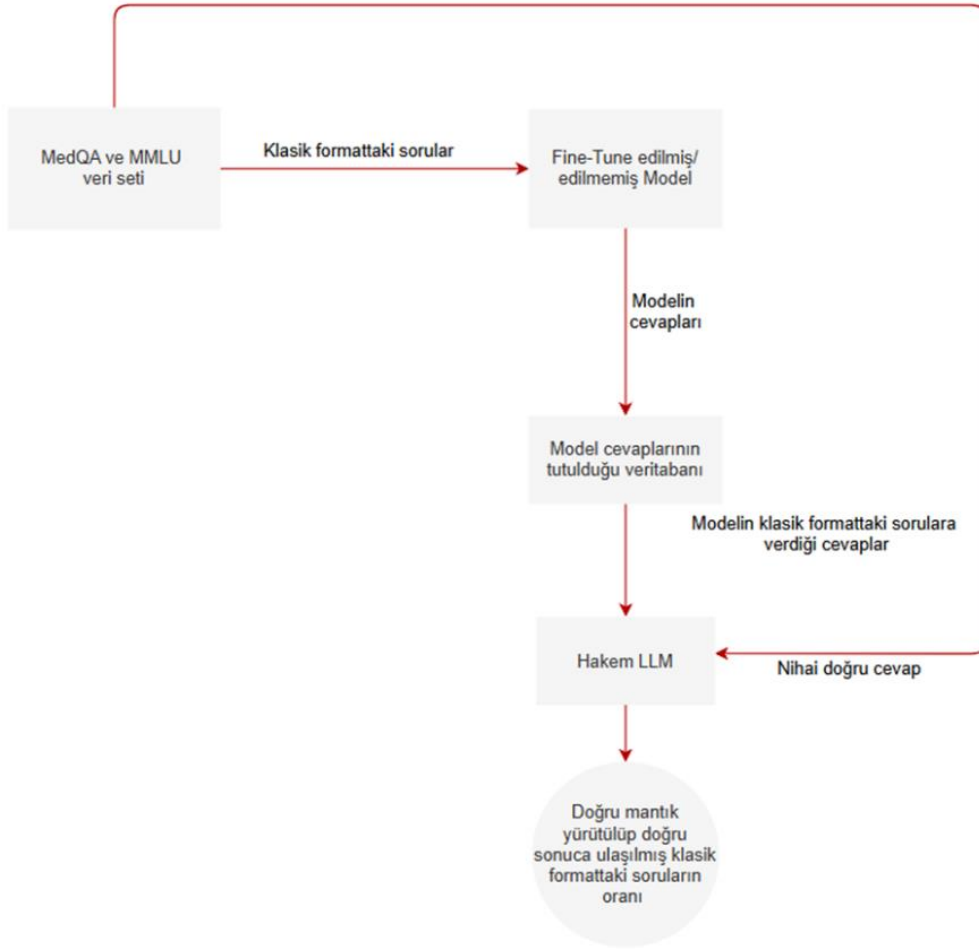
Tam Format Uyumu (match_format_exactly): Modelin çıktısının, <start_working_out>, <SOLUTION> gibi etiketleri içeren tam ve doğru yapısal formata sahip olması durumunda yüksek bir pozitif ödül (+3.0) verilir.

Yaklaşık Format Uyumu (match_format_approximately): Bu kritik etiketlerin varlığını veya yokluğunu kontrol ederek kısmi bir ödül veya ceza uygular.

Doğru ve Tekil Cevap (check_for_unique_pair_final_batched): Modelin <SOLUTION> etiketleri arasına, sorudaki çoktan seçmeli şıklardan yalnızca bir tanesini ve doğru olanı eklemesi durumunda büyük bir ödül (+5.0) verilirken, birden fazla şık sunarak belirsizlik yaratması negatif bir ödül (-2.0) ile cezalandırılır.

3.4. Değerlendirme Metodolojisi

Model performansının bütünsel bir resminin elde edilmesi amacıyla, basit doğruluk metriklerinin ötesine geçen, hakem LLM tabanlı bir değerlendirme süreci benimsenmiştir. Tüm değerlendirmelerde, endüstri standardı olarak kabul edilen GPT-4.1 modeli, "hakem" olarak kullanılmıştır. Bu sayede, özellikle açık uçlu cevapların ve karmaşık mantık zincirlerinin nüanslı bir şekilde değerlendirilmesi sağlanmıştır. Her bir model çıktısı, hakem LLM tarafından iki ana eksenle değerlendirilmiştir: Cevap Doğruluğu, model tarafından üretilen nihai sonucun, veri setindeki "gerçek değer (ground-truth)" cevap ile eşleşip eşleşmediği (eş anlamlı ya da klinik olarak çok yakın terimlere karşı tolerans gösterilmiştir) ve Mantık Yürütme Doğruluğu, cevaba ulaşırken sunulan akıl yürütme sürecinin (varsa) tıbbi olarak geçerli ve mantıksal olarak tutarlı olup olmadığının değerlendirilmesidir. Doğru cevaba bariz mantık hatalarıyla ulaşılan durumlar, "şans eseri doğru" olarak kabul edilmiştir.



Şekil 1. Sistem mimarisi

Bu mimari, model performansını hem Cevap Doğruluğu hem de Akıl Yürütme Doğruluğu eksenlerinde ölçmek için izlenen veri akışını göstermektedir. (1) FreedomIntelligence/medical-o1-verifiable-problem veri setinden alınan açık uçlu tıbbi problemler, temel (eğitilmemiş) veya ince ayar yapılmış (SFT/GRPO) Gemma modellerine girdi olarak sunulur. (2) Modelin soruya karşılık ürettiği tam cevap (akıl yürütme süreci ve nihai çözüm dahil) bir veritabanında saklanır. (3) Ardından, Hakem LLM (GPT-4.1), üç temel bilgiyi girdi olarak alır: orijinal problem, modelin ürettiği tam cevap ve veri setindeki nihai doğru cevap (ground truth). (4) Hakem LLM, bu girdileri karşılaştırarak modelin mantık zincirinin kalitesini ve nihai cevabının doğruluğunu değerlendirir. Bu süreç, çalışmamızda kullanılan Akıl Yürütme Doğruluğu metriğinin temelini oluşturur.

3. DENEYSEL BULGULAR

Tablo 2’de, Gemma 3 serisinin 1B ve 4B parametrelili temel (base) modellerinin, standart bir tıbbi soru-cevap veri seti olan PubMedQA üzerindeki başlangıç (baseline) performansı ortaya konmuştur. Bu değerlendirme, sonraki eğitim aşamalarının etkisinin ölçülmesi için bir referans noktası olarak kullanılmıştır. Sonuçlar aracılığıyla, model ölçeğinin performansa doğrudan ve pozitif bir etkisinin bulunduğu açık bir şekilde gösterilmiştir.

4B modelinin, %71 Cevap Doğruluğu ve %85 Akıl Yürütme Doğruluğu ile 1B modelinin (%54 ve %50) oldukça ilerisinde olduğu ortaya koyulmuştur. Bu durum, daha büyük modele doğal olarak daha gelişmiş bir bilgi tabanı ve muhakeme kapasitesinin kazandırıldığını teyit etmektedir. 4B modelinde gözlemlenen yüksek Akıl Yürütme Doğruluğu sayesinde, hakem LLM tarafından modelin kısa cevabında dahi örtük bir mantıksal tutarlılığın tespit edildiği veya cevap doğruluğunun daha yüksek bir güvenle değerlendirildiği ima edilmiştir.

Tablo 2. Temel Gemma modellerinin PubMedQA veri seti ile değerlendirmesi

Test edilen model	Eğitim durumu	Test veri seti (Benchmark)	Cevap doğruluğu (%)	Akıl yürütme doğruluğu (%)
Gemma 3 1B	Temel (Base)	PubMedQA	54%	50%
Gemma 3 4B	Temel (Base)	PubMedQA	71%	85%

Tablo 3’de, modellerin temel halleri üzerinden FreedomIntelligence/medical-o1-reasoning-SFT veri seti ile Gözetimli İnce Ayar (Supervised Fine-Tuning - SFT) işleminden geçirildikten sonraki PubMedQA performansı analiz edilmiştir. Metodolojide açıkça belirtildiği üzere, bu aşamanın birincil amacı modele yeni tıbbi bilgi kazandırmak değil, karmaşık problemleri yapılandırılmış bir Düşünce Zinciri (CoT) formatında ifade etme davranışının modele kazandırılmasıdır. Bu bağlamda, PubMedQA üzerindeki performans değişimlerinin dikkatle yorumlanması gerekmektedir. 4B modelinde, Cevap Doğruluğu sabit tutulurken (%71), Akıl Yürütme Doğruluğu %85’ten %81’e hafifçe gerilemiştir. Bu durum, model tarafından genel bir görev formatından (doğrudan cevap) daha spesifik ve yapısal bir formata (CoT) geçişin bir yan etkisi olarak yorumlanmaktadır. Daha küçük olan 1B modelde ise hem Cevap Doğruluğunda (%54’ten %53’e) hem de Akıl Yürütme Doğruluğunda (%50’dan %38’e) daha belirgin bir düşüşün gözlemlendiği raporlanmaktadır. Bu da, daha sınırlı kapasiteye sahip 1B model tarafından, yeni bir davranışsal format öğrenilmeye çalışılırken genel muhakeme yeteneğinde daha fazla "hizalama vergisi" (alignment tax) ödendiğini düşündürmektedir.

Tablo 3. SFT ile eğitilmiş modellerin PubMedQA veri seti ile değerlendirmesi

Test edilen model	Eğitim durumu	Test veri seti (Benchmark)	Cevap doğruluğu (%)	Akıl yürütme doğruluğu (%)
Gemma 3 1B	SFT (Huatu)	PubMedQA	53%	38%
Gemma 3 4B	SFT (Huatu)	PubMedQA	71%	81%

Tablo 4’te, temel modellerin doğrudan sorgu yoluyla karmaşık teşhis yeteneğini ölçen "Doğrulanabilir Problemler" veri seti üzerindeki ham yetenekleri gösterilmiştir. Bu veri setinin zorluk seviyesinin, PubMedQA’ya kıyasla çok daha yüksek olduğu belirtilmektedir. Elde edilen sonuçlarla, her iki temel modelin de bu karmaşık görevde özel bir eğitim almaksızın oldukça düşük bir başarı oranına sahip olduğu ortaya konmuştur. 4B modelinin (%29 Cevap Doğruluğu, %26 Akıl Yürütme Doğruluğu), 1B modeline (%8 Cevap Doğruluğu, %4 Akıl Yürütme Doğruluğu) kıyasla daha iyi bir performans sergilediği gösterilse de, bu skorlarla temel modellerin yapılandırılmış ve çok adımlı tıbbi akıl yürütme konusunda ciddi sınırlamalara sahip olduğu kanıtlanmıştır. Elde edilen bu bulgularla, SFT gibi hedefe yönelik eğitim stratejilerinin neden gerekli olduğu açık bir şekilde ortaya konmuştur.

Tablo 4. Temel Gemma modellerinin doğrulanabilir problemler ile değerlendirilmesi

Test edilen model	Eğitim durumu	Test veri seti (Benchmark)	Cevap doğruluğu (%)	Akıl yürütme doğruluğu (%)
Gemma 3 1B	Temel (Base)	Doğrulanabilir problemler (Verifiable problems)	%8	%4
Gemma 3 4B	Temel (Base)	Doğrulanabilir problemler	29%	26%

Bu çalışmanın en dikkat çekici bulgularından biri Tablo 5’te sunulmuş ve SFT eğitiminin "Doğrulanabilir Problemler" görevindeki etkisi detaylı şekilde incelenmiştir. Özellikle Gemma 3 4B modelinde gözlemlenen değişimle, SFT aşamasının başarısı gözler önüne serilmiştir: Modelin Cevap Doğruluğu %29’dan %27’ye marjinal bir düşüş göstermiş; buna karşın, Akıl Yürütme Doğruluğu %26’dan %29’a yükselmiştir.

Bu sonuçla birlikte, SFT’nin amacına ulaştığı güçlü bir şekilde ortaya konmuştur. Eğitim süreci boyunca, modele doğrudan doğru cevabın ezberletilmesinden ziyade, teşhise giden yolda mantıksal olarak daha tutarlı ve tıbben geçerli bir Düşünce Zinciri (CoT) oluşturma becerisinin kazandırıldığı görülmektedir. Nihai teşhis formüle edilirken küçük hatalar yapılmış olsa bile, sorunun analiz edilmesi ve mantıksal adımların sunulması konularında modelin daha yetkin hale geldiği anlaşılmaktadır.

Bu gelişim, modelin pratik tıbbi uygulamalardaki denetlenebilirliği ve güvenilirliği açısından, salt cevap doğruluğundan çok daha değerli bir kazanım olarak değerlendirilmektedir. Sonuç olarak, SFT ile modele hedeflenen davranışsal yeteneğin başarıyla kazandırıldığı Tablo 5'te net bir şekilde gösterilmiştir.

Tablo 5. SFT ile eğitilmiş Gemma modellerinin doğrulanabilir problemler ile değerlendirilmesi

Test edilen model	Eğitim durumu	Test veri seti (Benchmark)	Cevap doğruluğu (%)	Akıl yürütme doğruluğu (%)
Gemma 3 1B	SFT (Huatuo)	Doğrulanabilir problemler	%8	%4
Gemma 3 4B	SFT (Huatuo)	Doğrulanabilir problemler	27%	29%

GRPO ile Akıl Yürütmenin Rafine Edilmesi ve Metriklerin Sınırları

SFT aşamasının sonuçlarıyla, 1B modelinin bu karmaşık davranışsal eğitimi özümsemeye kapasite sınırlamaları sergilediği ortaya konmuştur. Özellikle PubMedQA'daki akıl yürütme performansında gözlemlenen belirgin düşüş ile "Doğrulanabilir Problemler" veri setinde herhangi bir ilerlemenin kaydedilememesi sonucunda, bu modelin daha ileri eğitim adımları için uygun bir aday olmadığı belirlenmiştir. Bu nedenle, çalışmanın bir sonraki aşaması olan GRPO ile pekiştirmeli öğrenmenin yalnızca daha yüksek kapasiteli ve SFT'ye daha olumlu yanıt verdiği gözlemlenen 4B model üzerinde uygulanmasına karar verilmiştir.

GRPO eğitiminin ardından elde edilen sonuçlarla, çalışmanın en ilgi çekici bulguları sunulmuştur: Modelin Cevap Doğruluğu %27'den %26'ya düşürülmüş; buna karşın, Akıl Yürütme Doğruluğu %29'dan %31'e yükseltilmiştir. Bu durumla birlikte, SFT ve GRPO eğitimlerinin birleşik etkisiyle modelin muhakeme kalitesinin nihai sonuçtan daha fazla önceliklendirildiği görülmektedir.

Paradoksal olarak değerlendirilebilecek bu eğilim, değerlendirme metodolojisinin doğası ve tıbbi teşhis sürecinin karmaşıklığıyla açıklanmaktadır. Klinik pratikte, sunulan bulgularla uyumlu birden fazla geçerli tanının (diferansiyel tanı) bulunabildiği; ancak veri setinde yer alan "ground truth"un bu olasılıklardan yalnızca birini temsil ettiği belirtilmektedir. Modelin akıl yürütme yeteneği geliştirildikçe, veri setindeki tekil doğru cevaba kilitlenmek yerine, klinik olarak geçerli fakat farklı bir sonuca ulaştıran bir mantık zinciri üretebildiği gözlemlenmektedir.

Hakem LLM tarafından bu mantık zinciri tutarlı ve doğru olarak değerlendirildiğinde, Akıl Yürütme puanı artırılmakta; ancak nihai cevabın "ground truth" ile eşleşmemesi durumunda, Cevap Doğruluğu puanı düşürülmektedir. Bu doğrultuda, GRPO'nun modeli daha sofistike ve tıbben nüanslı bir akıl yürütmeye teşvik ettiği; fakat bu gelişimin, yalnızca tek bir doğru cevaba dayalı metriklerle tam olarak yakalanamadığı sonucuna varılmaktadır.

4. TARTIŞMA VE SONUÇ

Bu çalışmada, farklı ölçeklerdeki (1B ve 4B) Gemma 3 dil modellerinin tıbbi alandaki akıl yürütme yetenekleri, çeşitli eğitim aşamaları (temel, SFT, GRPO) ve görevler (RAG ve doğrudan sorgu) üzerinden kapsamlı bir şekilde analiz edilmiştir. Elde edilen bulgularla, model ölçeğinin, özellikle bağlam sunulduğunda (PubMedQA RAG görevi), akıl yürütme performansı üzerinde kritik bir etkiye sahip olduğu net bir şekilde ortaya konmuştur. 4B temel model tarafından, %85 Mantık Yürütme Doğruluğu elde edilerek bu alanda doğal bir yetkinliğe sahip olduğu gösterilmiş; buna karşılık, 1B modelin yetersiz kaldığı ve SFT süreciyle bu durumun iyileştirilemediği gözlemlenmiştir.

Çalışmanın asıl odak noktasını oluşturan, bağlam sunulmayan ve açık uçlu cevap gerektiren "Doğrulanabilir Tıbbi Problemler" test setinde ise daha farklı ve katmanlı bir sonuca ulaşılmıştır. 4B temel modelle %26 gibi düşük bir Mantık Yürütme Doğruluğu elde edilerek modelin ham tıbbi bilgi hatırlama becerisi açısından zayıf olduğu teyit edilmiştir. Huatuo verisiyle gerçekleştirilen SFT eğitimi sonrasında, modelin Cevap Doğruluğu (%27) önemli ölçüde değişmemiş; ancak Mantık Yürütme Doğruluğu %29'a yükseltilmiştir. Bu sonuçla, SFT süreciyle modele yeni tıbbi gerçeklerin öğretilmesinden çok, düşüncelerini adım adım yapılandırma (Düşünce Zinciri – CoT) davranışının başarıyla kazandırıldığı gösterilmiştir.

Son aşama olan MMLU medikal verisiyle gerçekleştirilen GRPO eğitimiyle ise bu eğilim daha da ileri taşınmıştır. Cevap Doğruluğu %26'da sabit kalırken, Mantık Yürütme Doğruluğu %31'e çıkarılarak tüm aşamalar arasında en yüksek seviyeye ulaşılmıştır.

Bu sonuçlarla, çalışmanın temel hipotezinin doğrulandığı görülmektedir: Açık uçlu tıbbi problemlerde, tek bir "doğru cevap" etiketine sıkı sıkıya bağlı kalınmasının, modelin gerçek yeteneğini ölçmek açısından yetersiz kalabileceği ortaya konmuştur. Birden fazla klinik olarak geçerli cevap bulunabileceğinden, asıl başarı metriğinin, modelin mantıksal olarak tutarlı ve tıbben sağlam bir akıl yürütme süreci sergileyip sergileyemediği olması gerektiği vurgulanmıştır. GRPO modeliyle ulaşılan %31'lik Mantık Yürütme Doğruluğu ile, modelin cevapları ezberlemek yerine nasıl düşüneceğini öğrendiği açıkça ortaya konmuştur. Sonuç olarak, bu çalışmayla az parametrelili büyük dil modellerinin tıbbi alanda uzmanlaştırılmasında, davranışsal ve yapısal akıl yürütme becerilerinin kazandırılmasının önemi vurgulanmıştır.

Ek 1: GRPO ile Eğitilmiş Modelin Akıl Yürütme Örneği

Aşağıdaki örnek, SFT eğitimi üzerine GRPO ile eğitilmiş Gemma 4B modelinin, doğrudan sorgu görevindeki performansı ve Hakem LLM (GPT-4.1) tarafından yapılan değerlendirmenin bir kesitini sunmaktadır. Bu örnek, çalışmanın temel hipotezi olan, modelin mantık yürütme kalitesinin geliştirilmesinin nihai cevap doğruluğundan daha değerli bir kazanım olduğunu açıkça göstermektedir.

Soru: An 88-year-old woman with osteoarthritis is experiencing mild epigastric discomfort and has vomited material resembling coffee grounds multiple times. Considering her use of naproxen, what is the most likely cause of her gastrointestinal blood loss?

Modelin Yanıtı:

model <start_working_out>Alright, let's think this through. We've got an 88-year-old woman who's dealing with some pretty common issues like osteoarthritis. She's also been having this mild epigastric discomfort and has been vomiting stuff that looks like coffee grounds. That's a pretty classic sign of something going on with her stomach lining. Now, she's taking naproxen. Naproxen is a nonsteroidal anti-inflammatory drug, or NSAID for short. These are known to be pretty tough on the stomach lining. They can cause a lot of damage because they inhibit the production of prostaglandins, which are important for protecting the stomach. So, with her taking naproxen, it's not surprising that she's experiencing this kind of gastrointestinal bleeding. The coffee ground-like vomit is a big clue here. It suggests that there's blood coming from somewhere in the upper part of her digestive tract. Given her age and the fact that she's on naproxen, the most likely culprit for her bleeding is probably a peptic ulcer. These ulcers can form in the stomach or the duodenum, and they can bleed. The symptoms she's showing, like the epigastric discomfort and the coffee ground vomit, fit well with this diagnosis. It's pretty common for NSAIDs to cause these symptoms, especially in older adults. So, putting it all together, it seems pretty clear that the naproxen is likely causing her peptic ulcer, which is leading to the bleeding. That's why she's experiencing these symptoms. Yeah, that makes sense. The connection between her medication, her symptoms, and the likely cause of her bleeding is pretty solid.<end_working_out> <SOLUTION>The most likely cause of the gastrointestinal blood loss in this 88-year-old woman is a peptic ulcer. Naproxen, a nonsteroidal anti-inflammatory drug (NSAID), is known to cause damage to the stomach lining by inhibiting the production of prostaglandins, which are protective for the stomach. This can lead to the formation of peptic ulcers, particularly in older adults. The symptoms of epigastric discomfort and coffee ground-like vomit are classic indicators of upper gastrointestinal bleeding, often associated with peptic ulcers. Therefore, the use of naproxen is likely contributing to her peptic ulcer, resulting in the observed gastrointestinal bleeding.</SOLUTION>

Hakem LLM Değerlendirmesi:

Doğruluk Puanı (correctness): false

Akıl Yürütme Doğruluğu (reasoning_correctness): true

Gerekçe (justification): The model's final answer is 'peptic ulcer,' which is a broader term that includes both gastric and duodenal ulcers. The ground truth specifies 'gastric ulcer,' which is more precise and most likely given the NSAID use and the patient's symptoms. Therefore, the final answer is not a synonymous match and is too general. However, the reasoning is medically sound: the model correctly links NSAID use

to prostaglandin inhibition, mucosal injury, and ulcer formation, and accurately interprets the symptoms as upper GI bleeding. The reasoning is logical and without significant errors.

Yorum: Bu örnekte görüldüğü gibi, modelin nihai cevabı (peptic ulcer), veri setindeki kesin doğru cevap (gastric ulcer) ile tam olarak eşleşmediği için **Cevap Doğruluğu** puanı düşürülmüştür. Ancak, modelin yürüttüğü akıl yürütme süreci, hastanın semptomlarını (kahve telvesi benzeri kusma) ve ilaç kullanımını (naproxen, bir NSAID) doğru bir şekilde ilişkilendirmiş ve bu durumun peptik ülserlere yol açabileceği yönünde tıbben geçerli bir mantık zinciri oluşturmuştur. Hakem LLM bu akıl yürütmeyi mantıksal ve hatasız bulduğu için **Akıll Yürütme Doğruluğu** puanı yükselmiştir. Bu durum, GRPO eğitiminin modeli, sadece ezberlenmiş bir cevaba değil, klinik olarak sağlam bir mantık sürecine yönlendirdiğini göstermektedir. Bu örnek, Cevap Doğruluğu ve Mantık Yürütme Doğruluğu metrikleri arasındaki farkı ve çalışmanın ana bulgularının pratik önemini vurgulamaktadır.

Bu durum, kullanılan eğitim ve test verilerinin format farklılıklarıyla da açıklanmaktadır. SFT eğitiminde kullanılan veriler genellikle açık uçlu ve yapılandırılmış bir Düşünce Zinciri (CoT) formatına odaklanırken, GRPO eğitimi, MMLU veri setinin çoktan seçmeli yapısı üzerinde gerçekleştirilmiştir. MMLU sorularında model, doğru cevaba ulaşmak için birden fazla aday arasından en uygun olanı seçmeye teşvik edilmektedir. Bu nedenle, GRPO ile pekiştirilen modelin, nihai açık uçlu teşhis sorularında, doğru cevabı doğrudan vermek yerine, çoktan seçmeli sorular için yeterli olabilecek üst başlıkları veya genel terimleri (örneğin, "gastric ulcer" yerine "peptic ulcer") öne çıkarmaya meyletmesi beklenen bir sonuçtur. Model, çoktan seçmeli formatta öğrendiği "doğru cevabı seçme" becerisini, açık uçlu formatta "genel olarak doğru ve güvenli bir terim kullanma" şeklinde uyguluyor olabilir. Bu durum, modelin mantıksal muhakeme yeteneğinin geliştiğini, ancak farklı bir görev formatına (çoktan seçmeli) adapte olmasının, açık uçlu sorular için daha geniş bir dil üretimine neden olduğunu göstermektedir. Bu, modelin "ezberlemek yerine düşünme" hipotezini desteklerken, aynı zamanda eğitim verisi formatının model davranışını nasıl şekillendirdiğini de ortaya koymaktadır.

Akıll Yürütme Doğruluğunun Klinik Fayda İçin Bir Gösterge Olarak Metodolojik Doğrulanması

Akıll Yürütme Doğruluğu metriğimizin pratik sonuçlarını titizlikle doğrulamak ve geleneksel doğruluk skorlarının ötesine geçmek amacıyla, değerlendirme sonuçlarımızın en paradoksal alt kümeleri üzerinde detaylı bir niteliksel analiz gerçekleştirdik. Analizde iki kritik gruba odaklandık: modelin akıl yürütmesi doğru olmasına rağmen nihai cevabının yanlış olduğu "faydalı ama teknik olarak yanlış" vakalar ve akıl yürütmesi kusurlu olmasına rağmen cevabının doğru olduğu "şans eseri doğru" vakalar. Bu kritik görevin gerektirdiği incelikli değerlendirmeyi en yüksek standartta yapabilmek için, mevcut en gelişmiş dil modellerinden biri olan Gemini 2.5 Pro, uzman bir klinik değerlendirici olarak kullanılmıştır.

Her bir vaka, pratik faydanın üç temel yönünü ölçen katı bir değerlendirme çerçevesine göre analiz edilmiştir: **modelin hekimin karar verme sürecini doğru bir eylem planına (doğru tanısal adım, tedavi veya uzmanlık alanı sevk) yönlendirme başarısı**, hasta için kısa vadeli potansiyel risk ve çıktının genel klinik değeri. Analiz, çarpıcı ve kesin bir sonuç ortaya koymuştur: Akıll yürütmesi sağlam olan vakaların %100'ünde, modelin çıktısının, nihai cevabın ground truth etiketiyle anlamsal eşleşmesine bakılmaksızın, **hekim için güvenilir bir karar desteği sağladığı ve onu doğru sonraki adımlara yönlendirmede etkili olduğu** tespit edilmiştir.

Bu doğrudan korelasyon, bir modelin hekime sunduğu pratik faydayı ölçen ve "**Hekim Karar Desteği Doğruluğu**" olarak adlandırdığımız yeni ve daha anlamlı bir metrik tanımlamamıza olanak sağlamıştır. Bu yeni metriğin, Akıll Yürütme Doğruluğu ile birebir örtüşmesi, modelin gerçek dünyadaki pratik faydasının en net göstergesinin, mantık zincirinin kalitesi olduğunu kanıtlamaktadır. Bu metrikler arasındaki ilişki, aşağıdaki tabloda özetlenmiştir.

Tablo 6. Farklı başarı metriklerinin base model ve GRPO ile eğitilmiş model üzerindeki karşılaştırması

Değerlendirme metriği	Temel (base) model (%)	GRPO modeli (%)
Cevap doğruluğu	29	26
Akıll yürütme doğruluğu	26	31
Hekim karar desteği doğruluğu	26	31

Tablo 6'nın ortaya koyduğu en önemli bulgu, modellerin güvenilirlik profilindeki dönüşümdür. Temel modelin sonuçları incelendiğinde, akıl yürütmesi kusurlu olan "şans eseri doğru" vakaların sayısının (4 vaka), akıl yürütmesi sağlam olan "faydalı ama yanlış" vakaların sayısından (1 vaka) dört kat daha fazla olduğu görülmektedir. Bu durum, eğitilmemiş modelin, güvenilir ve faydalı bir çıktı üretmekten ziyade, hatalı bir mantıkla tesadüfen doğru sonuca ulaşmaya daha yatkın olduğunu ve klinik kullanım için güvenilirmez olduğunu göstermektedir.

GRPO ile gerçekleştirdiğimiz eğitim süreci ise bu oranı tamamen tersine çevirmiştir. Eğitilmiş modelimizde, "faydalı ama yanlış" vakaların sayısı (7 vaka), artık "şans eseri doğru" vakaların sayısını (2 vaka) 3.5 kat aşmaktadır. Bu dönüşüm, eğitim sürecimizin modelin sadece performansını artırmakla kalmayıp, aynı zamanda onu tutarlı bir şekilde sağlam ve güvenilir muhakeme yapmaya yönlendirdiğinin en net kanıtıdır. Bu nedenle, 'Akıl Yürütme Doğruluğu'nun, bir modelin gerçek klinik faydasını ve güvenliğini ölçmede vazgeçilmez bir standart olduğunu öne sürmekteyiz.

Analizimizde ortaya koyduğumuz 'şans eseri doğru' vakalar, güvenilir tıbbi yapay zeka geliştirilmesinin önündeki en temel zorluklardan birini gözler önüne sermektedir: akıl yürütmesi kusurlu bir teşhis, tesadüfen doğru sonuca ulaşsa bile, sistemin geneli hakkında bir güvenilmezlik sinyali verir. Bu bulgu, özellikle hastayı teşhis edip otonom olarak doğru kliniğe yönlendirmesi hedeflenen gelecekteki 'Agentic' sistemler için hayati bir önem taşımaktadır. Böyle bir sistemin güvenliği, yönlendirme eyleminin dayandığı teşhisin kalitesine mutlak bir şekilde bağlıdır ve hatalı bir mantık zinciri, hastanın yanlış bir uzmanlık alanına yönlendirilmesi gibi kritik hatalara yol açabilir.

Çalışmamızın en temel çıktısı, bir yapay zeka sisteminin güvenle doğru eylemde bulunabilmesi için, önce kanıtlanabilir şekilde doğru düşünmesi gerektiği ilkesidir. Bu ilke, güvenli tıbbi yapay zekanın geleceği için bir yol haritası sunmakta ve bu alanda geliştirilecek otonom sistemler için temel bir yapı taşı oluşturmaktadır.

5. TEŞEKKÜR

Bu çalışma, Türkiye Bilimsel ve Teknolojik Araştırma Kurumu (TÜBİTAK) tarafından TEYDEB 1505 – Üniversite-Sanayi İşbirliği Destekleme Programı kapsamında desteklenen 5240094 numaralı proje kapsamında gerçekleştirilmiştir. Sağladığı katkılar nedeniyle TÜBİTAK'a teşekkür ederiz. Bu yayında yer alan görüş, öneri ve sonuçlar yalnızca yazarların sorumluluğundadır ve TÜBİTAK'ın görüşlerini yansıtmamaktadır.

6. KAYNAKLAR

1. Chen, J., Cai, Z., Ji, K., Wang, X., Liu, W., Wang, R., Hou, J. & Wang, B. (2024). *FreedomIntelligence/HuatuoGPT-o1-8B. Hugging Face*, <https://huggingface.co/FreedomIntelligence/HuatuoGPT-o1-8B>, Erişim tarihi: 14.07.2025.
2. Wu, J., Deng, W., Li, X., Liu, S., Mi, T., Peng, Y., Xu, Z., Liu, Y., Cho, H., Choi, C.I. & Cao, Y. (2025). Medreason: Eliciting factual medical reasoning steps in LLMs via knowledge graphs. *arXiv preprint*, arXiv:2504.00993, Erişim tarihi: 25.08.2025.
3. Lai, Y., Zhong, J., Li, M., Zhao, S. & Yang, X., (2025). Med-r1: Reinforcement learning for generalizable medical reasoning in vision-language models. *arXiv preprint*, arXiv:2503.13939, Erişim tarihi: 23.07.2025.
4. DeepSeek, (2024). *GitHub - deepseek-ai/DeepSeek-R1*. GitHub, <https://github.com/deepseek-ai/DeepSeek-R1>, Erişim tarihi: 14.07.2025.
5. Ding, F., Wang, B., Zeng, Z. & Wang, Y. (2025). Multi-layer GRPO: Enhancing reasoning and self-correction in large language models. *arXiv preprint*, arXiv:2506.04746, Erişim tarihi: 23.07.2025.
6. Tinn, R., Cheng, H., Gu, Y., Usuyama, N., Liu, X., Naumann, T., Gao, J. & Poon, H. (2023). Fine-tuning large neural language models for biomedical natural language processing. *Patterns*, 4(4), 1-11.
7. Dai, W., Chen, P., Ekbote, C. & Liang, P.P. (2025). QoQ-Med: Building multimodal clinical foundation models with domain-aware GRPO training. *arXiv preprint*, arXiv: 2506.00711, Erişim tarihi: 23.07.2025.
8. Liu, M., Hu, W., Ding, J., Xu, J., Li, X., Zhu, L., Bai, Z., Shi, X., Wang, B., Song, H. & Liu, P. (2024). Medbench: A comprehensive, standardized, and reliable benchmarking system for evaluating Chinese medical large language models. *Big Data Mining and Analytics*, 7(4), 1116-1128.

9. Ogdu, C.U., Gurbuz, S., Karakose, M. & Hanoglu, E. (2025). Medical implications of LLM based clinical decision support systems in healthcare. *29th International Conference on Information Technology (IT 2025)*, Zabljak, Montenegro, 1-4.
10. Ali, H., Qadir, J., Alam, T., Househ, M. & Shah, Z. (2023). ChatGPT and large language models in healthcare: Opportunities and risks. *IEEE International Conference on Artificial Intelligence, Blockchain, and Internet of Things (AIBThings 2023)*, Mount Pleasant, MI, USA, 1-4.
11. Li, J., Deng, Y., Sun, Q., Zhu, J., Tian, Y., Li, J. & Zhu, T. (2024). Benchmarking large language models in evidence-based medicine. *IEEE Journal of Biomedical and Health Informatics*, 1-14.
12. Yang, X., Chen, A., PourNejatian, N., Shin, H.C., Smith, K.E., Parisien, C., Compas, C., Martin, C., Costa, A.B., Flores, M.G. & Zhang, Y. (2022). A large language model for electronic health records. *NPJ Digital Medicine*, 5(1), 194.
13. Kumar, R., Lomchavakarn, P., Angasinha, C., Intaratat, K., Boonsawad, P. & Sridee, S. (2025). Large language model based system for clinical decision support. *International Conference on Cognitive Computing in Engineering, Communications, Sciences and Biomedical Health Informatics (IC3ECSBHI 2025)*, Greater Noida, India, 389-394.
14. Ding, S., Ye, J., Hu, X. & Zou, N. (2024). Distilling the knowledge from a large-language model for health event prediction. *Scientific Reports*, 14(1), 30675.
15. Wang, C., Chen, Q., Shao, W. & He, X. (2024). KEMedGPT: Intelligent medical pre consultation with knowledge enhanced large language model. *Proceedings – 2024 IEEE International Conference on Medical Artificial Intelligence (MedAI 2024)*, Chongqing, China, 386-391.
16. Peng, C., Yang, X., Chen, A., Smith, K.E., PourNejatian, N., Costa, A.B., Martin, C., Flores, M.G., Zhang, Y., Magoc, T. & Lipori, G. (2023). A study of generative large language model for medical research and healthcare. *NPJ Digital Medicine*, 6(1), 210.
17. Rasmy, L., Xiang, Y., Xie, Z., Tao, C. & Zhi, D. (2021). Med-BERT: Pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ Digital Medicine*, 4(1), 86.
18. Rahman, M.A., Preum, S.M., Williams, R.D., Alemzadeh, H. & Stankovic, J. (2023). EMS BERT: A pretrained language representation model for the emergency medical services (EMS) domain. *Proceedings – 2023 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE 2023)*, Orlando, FL, USA, 34-43.
19. Li, W., Yu, L., Wu, M., Liu, J., Hao, M. & Li, Y. (2023). DoctorGPT: A large language model with Chinese medical question answering capabilities. *2023 International Conference on High Performance Big Data and Intelligent Systems (HDIS 2023)*, Macau, China, 186-193.
20. Akilesh, S., Sheik Abdullah, A., Abinaya, R., Dhanushkodi, S. & Sekar, R. (2023). A novel AI based chatbot application for personalized medical diagnosis and review using large language models. *2023 IEEE International Conference on Research Methodologies in Knowledge Management, Artificial Intelligence and Telecommunication Engineering (RMKMATE 2023)*, R.M.K. Engineering College, Chennai, India.
21. Peikos, G., Kasela, P. & Pasi, G. (2024). Leveraging large language models for medical information extraction and query generation. *2024 IEEE/WIC International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, Bangkok, Tayland, 367-372.
22. Kumar, P., Shreenidhi, G.L. & Rakesh Kumar, M. (2025). Medify-AI based LLM based healthcare system. *International Conference on Frontier Technologies and Solutions (ICFTS 2025)*, Chennai, India, 1-9.
23. Kumar, R., Shreenidhi, G.L. & Sowmiya, S. (2025). MedHub – LLM-based healthcare system. *3rd International Conference on Augmented Intelligence and Sustainable Systems (ICAISS 2025)*, Trichy, India, 1650-1657.
24. Aizu'bi, S., Kanan, T. & Almiani, M. (2024). Large language models for knowledge discovery in healthcare. *International Conference on Intelligent Computing, Communication, Networking and Services (ICCNS 2024)*, Dubrovnik, Croatia, 183-190.
25. Fernando, S.P.A.A. & Wickramaarachchi, D. (2025). Large language model (LLM) support for preliminary consultation in healthcare. *5th International Conference on Advanced Research in Computing (ICARC 2025)*, Belihuloya, Sri Lanka, 1-6.
26. Nazi, Z.A. & Peng, W. (2024). Large language models in healthcare and medical domain: A review. *Informatics*, 11(3), 57.
27. Tian, S., Jin, Q., Yeganova, L., Lai, P.T., Zhu, Q., Chen, X., Yang, Y., Chen, Q., Kim, W., Comeau, D.C. & Islamaj, R. (2023). Opportunities and challenges for ChatGPT and large language models in biomedicine and health. *Briefings in Bioinformatics*, 25(1), bbad493.

28. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L. & Chen, W. (2022). LoRA: Low-rank adaptation of large language models. *ICLR*, 1(2), 3.
29. Rao, A.K.G., Jaggi, A. & Naidu, S. (2025). MEDFIT LLM: Medical enhancements through domain focused fine tuning of small language models. *2nd International Conference on Research Methodologies in Knowledge Management, Artificial Intelligence and Telecommunication Engineering (RMKMATE 2025)*, Chennai, India, 1-5.
30. Kim, H., Hwang, H., Lee, J., Park, S., Kim, D., Lee, T., Yoon, C., Sohn, J., Park, J., Reykhart, O. & Fetherston, T. (2025). Small language models learn enhanced reasoning skills from medical textbooks. *NPJ Digital Medicine*, 8(1), 240.
31. Magnini, M., Aguzzi, G. & Montagna, S. (2025). Open-source small language models for personal medical assistant chatbots. *Intelligence-Based Medicine*, 11, 100197.
32. Wang, X., Dang, T., Kostakos, V. & Jia, H. (2024). Efficient and personalized mobile health event prediction via small language models. *ACM MobiCom 2024 – 30th International Conference on Mobile Computing and Networking*, Washington, D.C., USA, 2353-2358.
33. Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D. & Steinhardt, J. (2020). Measuring massive multitask language understanding. *arXiv preprint*, arXiv: 2009.03300, Erişim tarihi: 23.07.2025.
34. Hendrycks, D., Burns, C., Basart, S., Critch, A., Li, J., Song, D. & Steinhardt, J. (2020). Aligning AI with shared human values. *arXiv preprint*, arXiv: 2008.02275, Erişim tarihi: 23.07.2025.
35. Jin, D., Pan, E., Oufattole, N., Weng, W.H., Fang, H. & Szolovits, P. (2021). What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14), 6421.
36. Pal, A., Umapathi, L.K. & Sankarasubbu, M. (2022). MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering. *Conference on Health, Inference, and Learning*, ABD.
37. Hugging Face, (2025). *cais/mmlu. Hugging face datasets*, <https://huggingface.co/datasets/cais/mmlu>, Erişim tarihi: 25.07.2025.
38. Fırat, H. ve Üzen, H. (2024). MR görüntülerinden alzheimer hastalığının sınıflandırılması için inception ve sıkma-uyarma ağı tabanlı derin öğrenme modeli. *Çukurova Üniversitesi Mühendislik Fakültesi Dergisi*, 39(2), 555-567.
39. Abut, S. (2024). AI-based model design for prediction of COPD grade from chest X-ray images: Bir model önerisi (COPD-GradeNet). *Çukurova Üniversitesi Mühendislik Fakültesi Dergisi*, 39(2), 325-338.