# İnce Ayar ile Geliştirilmiş Gemma 3 Tabanlı Müşteri Temsilcisi Değerlendirme Sistemi

**Murat TAŞKIRAN[1], Emre ASLAN[2], Nihan KAHRAMAN[3]**

[1]E-mail: mrttskrn@yildiz.edu.tr; Yıldız Teknik Üniversitesi, Elektrik Elektronik Fakültesi, Elektronik ve Haberleşme Bölümü, İstanbul / Türkiye.
[2]E-mail: emre.aslan@mplusgroup.com.tr: MPlus Group Türkiye, İstanbul/ Türkiye.
[3]E-mail: nicoskun@yildiz.edu.tr; Yıldız Teknik Üniversitesi, Elektrik Elektronik Fakültesi, Elektronik ve Haberleşme Bölümü, İstanbul / Türkiye.

***Özet***
Büyük Dil Modelleri doğal dil işleme alanında dil yapısı, anlam bilgisi ve akıl yürütme gibi birçok görevde yüksek performans elde eden bir kavram olarak karşımıza çıkmaktadır. Büyük Dil Modelleri üzerine çalışan araştırmacıların ilgilendiği konulardan biri de diyalog kalitesi değerlendirmedir ve bu başlık birçok sektör için hayati öneme sahiptir. Bu çalışma kapsamında, kalite değerlendirme birimleri tarafından gerçekleştirilen Müşteri temsilcilerine ait konuşmaların incelenmesi işlemini, hızlı ve yüksek başarımla gerçekleştirebilecek özelleşmiş bir model oluşturmaya odaklanılmıştır. Son dönemde önerilmiş yüksek performanslı dil modelleri ile gerçekleştirilen kıyaslama çalışmaları sonucunda ince ayar yapılmış modelin %95'lik başarım elde ettiği gözlemlenmiştir. Bu sonuç, otonom müşteri temsilcisi değerlendirme modelleri yaratabilecek gelecek çalışmalar için umut vadetmektedir.
***Anahtar Kelimeler:*** Büyük Dil Modelleri, Diyalog Kalitesi Değerlendirme, İnce Ayar, Müşteri Temsilcisi Analitiği

***Abstract***
Large Language Models (LLMs) are emerging as a concept in natural language processing that achieves high performance in many tasks, such as language structure, semantics, and reasoning. One of the areas of interest for researchers working on LLMs is diaogue quality evaluation, a topic of vital importance for many industries. This study focuses on developing a specialized model capable of analyzing customer agent conversations, a task typically performed by quality assessment units, quickly and with high performance. Benchmarking studies with recently proposed high-performance language models have shown that the fine-tuned model achieves 95% accuracy. This result is promising for future studies that could create autonomous customer agent evaluation models.
***Keywords:*** Large Language Models (LLMs), Dialogue Quality Evaluation, Fine-Tuning, Customer Agent Analytics

## GİRİŞ

Büyük dil modelleri (LLM'ler), doğal dil işlemede temel bir paradigma olarak ortaya çıkmış ve büyük ölçekli verilerden dilsel yapı, anlambilim ve akıl yürütme öğrenme konusunda olağanüstü bir kapasite göstermiştir. Ölçekleme yasaları üzerine yapılan deneysel çalışmalar, tutulan verilerdeki çapraz entropi kaybıyla ölçülen performansın, model boyutu (parametre sayısı), eğitim verisi hacmi (token) ve işlem bütçesi açısından öngörülebilir kuvvet yasası iyileştirmelerini izlediğini ortaya koymuştur (Kaplan vd., 2020, Mohammadabadi vd., 2025). Bu bulgular, daha büyük modellerin örnek-verimliliğinin daha iyi olduğunu ve sınırlı ince ayar (fine-tuning) ile bile üstün dil anlama ve üretim yeteneklerine ulaşabildiğini göstermektedir. Sonuç olarak, modern LLM'ler özetleme, çeviri, soru-cevap ve açık uçlu metin üretimi gibi çok çeşitli görevler için esnek motorlar haline gelmiştir.

Son zamanda araştırmacılar LLM mimarileri geliştirmenin yanında, bu tür genel amaçlı LLM'leri ince ayar, talimat ayarlama, anında mühendislik ve insan odaklı geri bildirim yoluyla alan-özgü görevlere uyarlama stratejilerini araştırmıştır. Bu araştırmalar sonuncunda da, talimat ayarlı veya az sayıda adımlı öğrenme

yöntemleri ile, nispeten daha alt kapsama sahip görev-özgü verilerin kullanıldığı modellerin genel LLM modellerine kıyasla çok daha başarılı şekilde çalıştığını göstermişlerdir (Kojima vd. 2022; Gao, Fisch & Chen, 2020). Örneğin, eğitim veya klinik diyaloglar gibi bağlamlarda, özelleştirilmiş LLM tabanlı analizler, insan açıklama yükünü azaltmada umut verici sonuçlar göstermiştir: örneğin, bir LLM, pedagojik kalite ve etkileşim açısından sınıf içi diyalogları analiz etmek için kullanılmıştır (Long, Luo ve Zhang, 2024) ve başka bir çalışmada ise, LLM'lerin klinik diyalog özetleme performansını, hesaplamalı metrikleri uzman insan değerlendirmesiyle birleştirerek değerlendirmiştir (Fraile Navarro vd., 2025). Bu çalışmalarda da görüldüğü üzere, alanla ilgili veriler ve değerlendirme kriterleri dikkatlice tanımlandığı takdirde, LLM'lerin bağlama duyarlı görevlerde yüksek performanslar gösterebildiği açıkça gözlemlenmiştir.

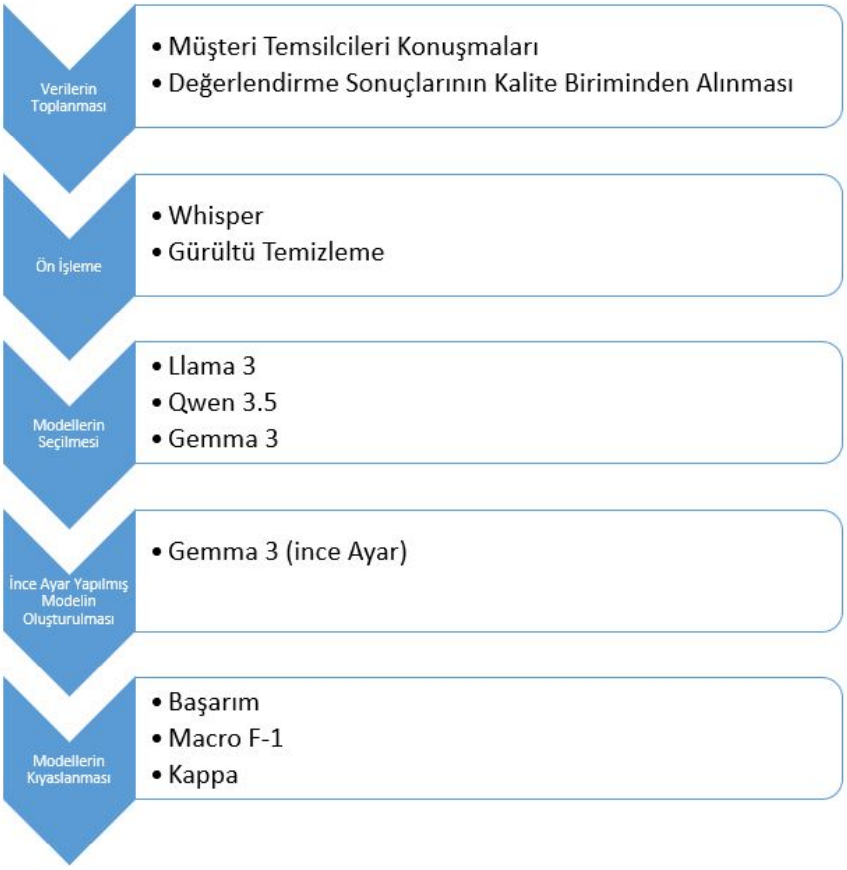
Bu gelişmelere dayanarak, son yıllardaki çalışmalarda LLM mimarileri yalnızca dil üretmek için değil, aynı zamanda müşteri hizmetleri, çağrı merkezleri veya temsilci performans değerlendirmesi için kritik öneme sahip olan diyalog kalitesini değerlendirmek için de kullanılmaya başlanmıştır. Örneğin, 2024 yılında Jia ve arkadaşları tarafından gerçekleştirilen çalışmada, düşünce zinciri yönlendirmesi ve özenle seçilmiş bağlam içi örnekler kullanılarak büyük ve ince ayarlı LLM modellerinin, konuşma kalitesine ilişkin performansının insan yargılarına yaklaşabileceği gösterilmiştir (Jia vd., 2024). Diyalog değerlendirme yöntemlerine dair kapsamlı bir inceleme sunan bir derleme çalışmasının literatürde yer alıyor olması (Braggaar vd., 2023), görev odaklı ve müşteri hizmetleri diyalog sistemlerinde LLM tabanlı değerlendirmeden yararlanma eğiliminin giderek arttığını daha da vurgulamaktadır. Bu çalışmalar, toplu olarak, doğru şekilde ayarlanmış ve alan farkındalığı yüksek LLM'lerin, temsilci performans değerlendirme iş akışlarında insan değerlendiriciler için etkili birer aracı olarak hizmet edebileceğini hatta ilerleyen süreçlerde ise insan değerlendiricilerin yerine geçebileceklerini göstermektedir.

Literatür çalışmalarında son yıllarda gerçekleşen bu gelişmeler, diyalog temelli değerlendirme sistemlerinin hızla geliştiğini açıkça göstermiştir. Bu gelişmeler ışığında, bu çalışma kapsamında Müşteri Temsilcilerinin Müşteriler ile gerçekleştirdiği görüşmeleri değerlendirmek için görev odaklı ince ayar yapılmış bir LLM oluşturulmasına odaklanılmıştır. Oluşturulan bu LLM ile, her firmada yer alan Müşteri temsilcilerinin görüşmelerinin değerlendirilmesinde görevli Kalite birimi görevlilerine yardımcı olunabilecek, ilerleyen süreçte ise uzun süre alan bu işlemlerin tamamen özelleşmiş bir LLM ile kısa sürede tamamlanabilmesi hedeflenmiştir.

İlerleyen başlıklarda, öncelikle özelleşmiş LLM'in oluşturulmasında hangi adımların kullanıldığının detaylı bir anlatımı "YÖNTEM" başlığı altında verilecek, daha sonra bu LLM modeli ile elde edilen test sonuçları paylaşılacak ve son olarak "SONUÇ ve TARTIŞMA" başlığı altında hem elde edilen sonuçlar yorumlanacak hem de gelecek çalışmalarda neler yapılabilir bu konu hakkında bilgiler verilecektir.

## YÖNTEM

Bu çalışma kapsamında, Müşteri Temsilcilerinin görüşmelerinin değerlendirilmesi için görev-temelli özelleşmiş bir LLM modeli oluşturulmaya çalışılmıştır. Önerilen yöntem için kullanılan adımları gösteren blok diyagram Şekil 1'de verilmiştir.




Şekil 1. Müşteri Temsilcisi Değerlendirme için Özelleştirilmiş LLM modeli için Blok Diyagram

Müşteri Temsilcisi Değerlendirme için Özelleştirilmiş LLM modeli oluşturmak için gerçekleştirilen ilk adım Müşteri Temsilcilerinin müşteriler ile gerçekleştirdiği konuşmaların toplanmasıdır. Bu amaç doğrultusunda, Kalite Değerlendirme Birimi tarafından daha önce değerlendirilmiş olan en az 300 adet müşteri temsilcisi görüşmesi seçilmiş ve bu görüşmelerin insan değerlendirici puanları referans etiket olarak kullanılmıştır. İnsan değerlendiriciler tarafından yapılan kalite değerlendirmeleri aşağıdaki kriterlere dayanmaktadır:

1. Müşteri ihtiyacını doğru anlama
2. Sunulan çözümün doğruluğu ve yeterliliği
3. İletişim tonu, nezaket ve empati
4. Kurumsal süreç ve prosedürlere uyum

Her kriter 5'li Likert ölçeği üzerinden puanlanmış, ardından bu puanlar birleştirilerek her görüşme için nihai kalite etiketi oluşturulmuştur.

Bu çalışma kapsamında Kalite Değerlendirme Birimi tarafından daha önce manuel olarak değerlendirilmiş toplam 300 adet müşteri temsilcisi görüşmesi kullanılmıştır. Veri seti %70 eğitim (210), %15 doğrulama (45) ve %15 test (45) olacak şekilde bölünmüştür. Ardından, tüm konuşma verileri anonimleştirilmiş, temizlenmiş ve model giriş formatına uygun şekilde yapılandırılmıştır. Bu ön işleme adımlarının

uygulanabilmesi içinde ilk adımda bu verilerin sesten yazıya dönüştürülmesi OpenAI tarafından geliştirilen "Whisper" ile gerçekleştirilmiştir (Radford vd. 2023). Bir sonraki adımda ise, literatürde hali hazırda kullanılmakta olan ve yüksek performans gösteren LLM modellerinin performansını kıyaslamak amacıyla, çalışmanın ilk adımında oluşturulan veri seti kullanılarak ve testler gerçekleştirilmiş ve en yüksek performansı gösteren model ince ayar işlemi için seçilmiştir.

Bu çalışma kapsamında kullanılmasına karar verilen üç büyük dil modeli, Llama3, Qwen 3.5 ve Gemma 3 mimarileridir. Llama 3, geliştirilmiş muhakeme, kodlama ve çok dilli yeteneklere sahip, kod çözücü dönüştürücü mimarisine sahip Llama ailesinin en son önerdiği mimarisidir; model çok uzun bağlam pencerelerini (128K'ya kadar belirteç) destekler ve uzun biçimli metin oluşturma ve anlama görevlerinde sağlam performans sağladığı literatürde gösterilmiştir (Dubey vd. 2024). Qwen 3, Alibaba tarafından geliştirilen, hem yoğun hem de Uzman Karışımı (MoE) varyantlarını birleştiren ve iş yükü taleplerine bağlı olarak derin muhakeme veya daha hızlı yanıtlar için optimize edilmiş modlar sunan "hibrit muhakeme" yapısına sahip diğer bir büyük dil modelidir. Bu çalışmada kullanılan Qwen 3.5 modeli, Qwen2 mimarisi temel alınarak geliştirilen ve kurumsal diyalog analizi senaryoları için optimize edilmiş deneysel bir varyanttır (Bai et al., 2023; Yang et al., 2024). Google tarafından geliştirilen Gemma 3, metin, görüntü ve potansiyel olarak diğer biçimleri destekleyen çok modlu ve çok dilli modellerden oluşan bir diğer LLM modelidir; tasarımı, verimliliği, geniş dil kapsamını (140'tan fazla dil) ve özetleme, muhakeme ve çok büyük modellere kıyasla nispeten orta düzeyde kaynak gereksinimleriyle diğer modellerden daha ön plana çıkabilmektedir (Team G. vd. 2025).

Bu modellerin bu çalışma kapsamında seçilme nedeni, her birinin farklı avantajları bulunmasıdır. Llama 3'ün uzun bağlamlı muhakemesi, Qwen 3'ün hibrit muhakeme yeteneği ve parametre ölçeklemesi, Gemma 3'ün ise çok dilli ve çok modlu esnekliği bu üç modeli hem literatürdeki diğer dil modellerine hem de kendi aralarındaki bir kıyaslamada daha avantajlı hale getirmektedir. "BULGULAR" alt başlığında detayları verilecek olan kıyaslama testleri sonucunda Gemma 3'modeli en yüksek performansı göstermiş ve bu çalışma kapsamında ince ayar gerçekleştirilecek model olarak seçilmiştir. Daha sonrasında hem üç büyük dil modeli hem de ince ayar yapılmış olan Gemma 3 modelinin performansı üç değerlendirme metriği kullanılarak kıyaslanmıştır. Bu metriklerden ilki başarım metriğidir. Başarım metriğini Model tahminlerinin insan etiketi ile tamamen örtüştüğü örneklerin oranı tanımlayabiliriz. İkinci metrik olan Macro-F1 ise Tüm sınıfların F1 skorlarının eşit ağırlıklı ortalaması; özellikle sınıf dengesizliği olan yapılarda önemlidir. Son değerlendirme metriği olan Cohen's Kappa ise, İnsan–model arasındaki gerçek anlaşma düzeyini, tesadüfî anlaşmayı düzelterek ölçen istatistiksel bir metriktir (Cohen, 1960). İnsan değerlendiriciler; müşteri ihtiyacını anlama, çözüm doğruluğu, iletişim tonu ve süreç uyumu kriterlerine göre 5'li Likert ölçeği kullanarak puanlama yapmıştır. Model çıktıları bu nihai etiketlerle birebir karşılaştırılmıştır.

## BULGULAR

Müşteri Temsilcisi Değerlendirme için özelleştirilmiş bir LLM yapısı oluşturma kapsamında yapılan çalışmalar sonucunda 3 adet büyük dil modeli, önceden hazırlanmış olan veri seti ile test edilmiş ve üç değerlendirme metriği için sonuçlar elde edilmiş ardından da en iyi performansı gösteren Gemma 3 modeline ince ayar yapılarak LLM'in başarımının artırılması hedeflenmiştir. Bu amaç doğrultusunda gerçekleştirilen test sonuçları Tablo 1'de verilmiştir. Bu tablonun amacı nihai modelin insan değerlendiricilerle olan uyumunu göstermektedir.

**Tablo 1.** *LLM Modellerinin Performans Kıyaslaması*

| Model | Başarım (%) | Macro-F1 (%) | Cohen's Kappa |
|---|---|---|---|
| Llama 3 | 85 | 84 | 0.78 |
| Qwen 3.5 | 90 | 88 | 0.82 |
| Gemma 3 | 93 | 92 | 0.87 |
| **Gemma 3 (İnce Ayar)** | **95** | **94** | **0.90** |

Tabloda da görüldüğü üzere çalışma kapsamında gerçekleştirdiğimiz, önce büyük dil modellerinin herhangi bir işleme tabi tutmadan elde bulunan Müşteri temsilcisi görüşmelerini nasıl değerlendirdiğinin test edilmesi ve ardından en iyi modelin amaca yönelik olarak ince ayara tabi tutulması sonucunda elde edilen yeni model, yüksek performans göstererek görüşmeleri Kalite birimindeki çalışanlara yakın bir şekilde değerlendirmeyi başarabilmiştir.

## SONUÇ VE TARTIŞMA

Bu çalışma kapsamında son yıllarda yaygın olarak kullanılan Büyük Dil Modellerinden faydalanılarak, son dönemlerde araştırmacılarında üzerine yoğunlaştığı Diyalog Kalitesi Analizi'nin bir alt sınıfı olan Müşteri Temsilcisi görüşmelerinin değerlendirilmesi için ince ayar gerçekleştirilen bir LLM modeli oluşturulmuştur. Oluşturulan model, çalışma kapsamında literatürde yüksek performansa sahip üç adet Büyük Dil Modeli kıyaslanmış ve sonuçlar ince ayar gerçekleştirilen modelin diğer modellere kıyasla yüksek performans gösterdiğini üç farklı değerlendirme metriğini de kullanarak göstermiştir. Elde edilen %95'lik başarım ve 0.90 Cohen's Kappa değeri modelin insan değerlendiricilerine yakın bir performans gösterdiğini açıkça belirtmiştir. Bu sonuçlar, görev-odaklı ve alan-farkındalığı yüksek LLM'lerin müşteri temsilcisi değerlendirme süreçlerinde yalnızca destekleyici değil, uzun vadede **otonom değerlendirici** olarak da kullanılabileceğini ortaya koymaktadır.

Bundan sonraki aşamalarda ve gelecek çalışmalarda, öncelikle Müşteri temsilcilerine ait çok daha fazla sayıda görüşmeye ait verilerin toplanması ve veri setinin büyütülmesi ayrıca farklı modellerinde dahil edilerek daha kapsamlı bir çalışma gerçekleştirilmesi hedeflenmektedir. Bu sayede sadece Kalite Değerlendirme Birimlerine yardımcı olacak bir model yerine bu değerlendirme işlemini otonom olarak geliştirebilecek bir sistemin oluşturulması mümkün olacaktır.

## KAYNAKÇA

Braggaar, A., Liebrecht, C., van Miltenburg, E., & Krahmer, E. (2023). Evaluating task-oriented dialogue systems: A systematic review of measures, constructs and their operationalisations. *arXiv preprint arXiv:2312.13871*.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, *20*(1), 37-46.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., ... & Ganapathy, R. (2024). The llama 3 herd of models. *arXiv e-prints*, arXiv-2407.

Fraile Navarro, D., Coiera, E., Hambly, T. W., Triplett, Z., Asif, N., Susanto, A., ... & Berkovsky, S. (2025). Expert evaluation of large language models for clinical dialogue summarization. *Scientific reports*, *15*(1), 1195.

Gao, T., Fisch, A., & Chen, D. (2021, August). Making pre-trained language models better few-shot learners. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the*

*11th international joint conference on natural language processing (volume 1: long papers)* (pp. 3816-3830).

Jia, J., Komma, A., Leffel, T., Peng, X., Nagesh, A., Soliman, T., ... & Kumar, A. (2024). Leveraging LLMs for dialogue quality measurement. *arXiv preprint arXiv:2406.17304*.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., ... & Amodei, D. (2020). Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in neural information processing systems*, *35*, 22199-22213.

Long, Y., Luo, H., & Zhang, Y. (2024). Evaluating large language models in analysing classroom dialogue. *npj Science of Learning*, *9*(1), 60.

Mohammadabadi, S. M. S., Kara, B. C., Eyupoglu, C., Uzay, C., Tosun, M. S., & Karakuş, O. (2025). A survey of large language models: evolution, architectures, adaptation, benchmarking, applications, challenges, and societal implications. *Electronics*.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023, July). Robust speech recognition via large-scale weak supervision. In *International conference on machine learning* (pp. 28492-28518). PMLR.

Team, G., Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., ... & Iqbal, S. (2025). Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.