ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL

LMM-IQA: IMAGE QUALITY ASSESSMENT FOR LOW-DOSE CT IMAGING

M.Sc. THESIS

Kağan ÇELİK

Department of Electronics and Communication Engineering

Electronics Engineering Programme

FEBRUARY 2026

ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL

LMM-IQA: IMAGE QUALITY ASSESSMENT FOR LOW-DOSE CT IMAGING

M.Sc. THESIS

Kağan ÇELİK
(504221233)

Department of Electronics and Communication Engineering

Electronics Engineering Programme

Thesis Advisor: Assoc. Prof. Dr. İsa YILDIRIM

FEBRUARY 2026

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

LMM-IQA: DÜŞÜK DOZLU BT GÖRÜNTÜLERİNDE GÖRÜNTÜ KALİTESİ DEĞERLENDİRMESİ

YÜKSEK LİSANS TEZİ

Kağan ÇELİK
(504221233)

Elektronik ve Haberleşme Mühendisliği Anabilim Dalı

Elektronik Mühendisliği Programı

Tez Danışmanı: Doç. Dr. İsa YILDIRIM

ŞUBAT 2026

Kağan ÇELİK, a M.Sc. student of ITU Graduate School student ID 504221233 successfully defended the thesis entitled "LMM-IQA: IMAGE QUALITY ASSESSMENT FOR LOW-DOSE CT IMAGING", which he prepared after fulfilling the requirements specified in the associated legislations, before the jury whose signatures are below.

Thesis Advisor :	**Assoc. Prof. Dr. İsa YILDIRIM** Istanbul Technical University	.....
Jury Members :	**Prof. Dr. Ender Mete EKŞİOĞLU** Istanbul Technical University	.....
	Dr. Metin ERTAŞ Turkish Airlines	.....

Date of Submission : **15 January 2026**
Date of Defense : **18 February 2026**

To my family,

FOREWORD

I extend my deepest thanks to my advisor, Assoc. Prof. İsa Yıldırım, for his invaluable guidance and contributions throughout the development of this research. I am also grateful to Dr. Metin Ertaş and Mehmet Ozan Ünal for their constructive contributions and valuable feedbacks during the research process.

I am deeply thankful to my family for their unwavering love and support. Their presence has always been a source of strength and motivation throughout this journey.

February 2026 Kağan ÇELİK

TABLE OF CONTENTS

ABBREVIATIONS

AI	**:** Artificial Intelligence
BT	**:** Bilgisayarlı Tomografi
CT	**:** Computed Tomography
DL	**:** Deep Learning
FBI	**:** Fast Blind Image Denoiser
ICC	**:** Intraclass Correlation Coefficient
IQA	**:** Image Quality Assessment
KROCC	**:** Kendall Rank Correlation Coefficient
LDCT	**:** Low-Dose Computed Tomography
LLM	**:** Large Language Model
LMM	**:** Large Multimodal Model
MICCAI	**:** Medical Image Computing and Computer Assisted Intervention
NR-IQA	**:** No-Reference Image Quality Assessment
PLCC	**:** Pearson Linear Correlation Coefficient
PSNR	**:** Peak Signal-to-Noise Ratio
RL	**:** Reinforcement Learning
SROCC	**:** Spearman Rank Order Correlation Coefficient
SSIM	**:** Structural Similarity Index Measure

SYMBOLS

$\mathbf{y}$	: Radiologist score (ground truth)
$\hat{\mathbf{y}}$	: Model-predicted image quality score
$\mathbf{e}$	: Absolute prediction error ($\lvert y - \hat{y} \rvert$)
$\mathbf{x}$	: Input CT image
$\mathbf{f}_\theta$	: LMM prediction function parameterized by θ
$\boldsymbol{\theta}$	: Model parameters
$\mathbf{D}$	: Training data distribution
$\mathbf{n}$	: Total number of test images
$\mathbf{R_i}$	: Rank of radiologist score
$\hat{\mathbf{R}}_\mathbf{i}$	: Rank of model-predicted score
$\boldsymbol{\mu}$	: Mean radiologist score
$\hat{\boldsymbol{\mu}}$	: Mean model-predicted score
$\mathbf{J}(\theta)$	: Objective function minimizing prediction error
$\mathbf{n_i}$	: Estimated noise level of image i

LIST OF TABLES

LIST OF FIGURES

LMM-IQA: IMAGE QUALITY ASSESSMENT FOR LOW-DOSE CT IMAGING

SUMMARY

Low-Dose Computed Tomography (LDCT) reduces the amount of ionizing X-rays to which the patient is exposed; however, the accompanying increase in noise, blurring, and loss of contrast leads to a degradation in the diagnostic quality of the images. Therefore, consistency and robustness in image quality assessment have become critical requirements for clinical applications. Although numerous successful quantitative and qualitative methods are used in evaluating image quality, these methods and metrics also have various limitations. Image quality assessment approaches more commonly used in clinics rely heavily on human observation and cannot provide a standard level of assessment due to inter-observer differences. This situation highlights the need for more objective and reproducible assessment strategies.

In this study, an LLM-based image quality assessment system is proposed as an alternative solution to this problem. In the proposed method, a model has been developed that generates both numerical scores and textual descriptions of clinical images, taking into account degradations such as noise, blurring, and contrast loss present in the images. The model directly explains the degradations in the images with textual and numerical feedback, making quality differences more interpretable. The main purpose of this model is to provide clinicians with a consistent decision support tool and to make the evaluation process more objective.

Within the proposed method, different inference strategies such as zero-shot approach, metadata integration, and error feedback are systematically examined. The integration of contextual information allows the model to interpret visual changes from a medical perspective, while feedback mechanisms enable the system to refine its scoring accuracy. By evaluating the gains provided by these strategies individually, it has been demonstrated under what conditions the model produces output more reliably. This approach bridges the gap between purely mathematical metrics and the complex semantic requirements of radiological diagnosis. By leveraging the inherent reasoning capabilities of these models, the study moves beyond simple pattern recognition to provide a context-aware analysis of medical image integrity.

The results indicate that the proposed system not only provides highly consistent scores but also offers interpretable and clear analyses that remain compatible with clinical workflows. Unlike traditional models, this methodology focuses on explainability and flexibility, providing a strategic step toward minimizing clinical workload. In this respect, the system offers a practical framework for quality assessment in Low-Dose CT imaging, which can serve as a solid foundation for the integration of artificial intelligence into future clinical practice.

LMM-IQA: DÜŞÜK DOZLU BT GÖRÜNTÜLERİNDE GÖRÜNTÜ KALİTESİ DEĞERLENDİRMESİ

ÖZET

Bilgisayarlı Tomografi (BT), tıp dünyasının en kritik tanı araçlarından biridir. İnsan vücudunun iç yapısını invaziv olmayan bir yöntemle, yüksek çözünürlüklü kesitler halinde sunar. Ancak bu teknolojik görüntüleme yöntemi, doğası gereği hastayı iyonize radyasyona maruz bırakmaktadır. X-ışınlarının dokular üzerindeki stokastik etkileri tıp dünyasını mümkün olan en düşük radyasyon dozuyla en yüksek tanısal verimi elde etme prensibine yöneltmiştir. Bu etik sorumluluk doğrultusunda geliştirilen Düşük Dozlu Bilgisayarlı Tomografi (LDCT) protokolleri, radyasyon maruziyetini minimize ederek hasta güvenliğini önceliklendirmeyi hedefler. Fakat burada fiziksel bir paradoksla karşılaşılır: Radyasyon dozu azaldıkça, görüntüleme fiziği gereği dedektöre ulaşan foton sayısı düşer ve bu durum görüntülerde kuantum gürültüsünün artmasına, yapay çizgilerin (streak artifacts) oluşmasına ve doku kontrastının kaybolmasına neden olur.

Düşük dozlu bilgisayarlı tomografi, her ne kadar hastanın maruz kaldığı iyonize X-ışını miktarını azaltmış olsa da; doz azalımı ile beraber oluşan gürültü, bulanıklık ve kontrast kaybındaki artış görüntülerin tanısal kalitesini düşürmektedir. Görüntüleme fiziğinin temel prensipleri gereği, radyasyon dozundaki azalma, dedektör seviyesinde toplanan foton sayısını düşürür; bu da sinyal-gürültü oranının bozulmasına ve görüntülerde kuantum gürültüsü ile karakteristik "streak" (çizgi) artefaktlarının belirginleşmesine neden olur. Bu teknik bozulmalar, yumuşak doku kontrastını azaltarak anatomik detayların kaybolmasına ve dolayısıyla hekimin tanısal güveninin zedelenmesine yol açar. Bu nedenle görüntü kalitesi değerlendirmesinde tutarlılık ve sağlamlık, klinik uygulamalar için kritik bir gereksinim haline gelmektedir.

Görüntü kalitesinin ölçümünde nicel ve nitel birçok başarılı yöntem kullanılmasına rağmen bu yöntemler ve metrikler farklı olumsuz yanlar da barındırmaktadır. Klinikte daha sıklıkla kullanılan görsel değerlendirme yaklaşımları büyük oranda insan gözlemine dayanmakta ve gözlemciler arası farklılıklar nedeniyle standart bir değerlendirme düzeyi sunamamaktadır. Ayrıca radyologların manuel olarak yaptıkları bu değerlendirme süreci son derece zaman alıcı olmakla beraber klinik iş yükünü de artırmaktadır. Bu durum, daha nesnel ve tekrarlanabilir yaklaşımları gerekli kılmaktadır.

Literatürdeki mevcut Görüntü Kalitesi Değerlendirme (IQA) metrikleri incelendiğinde, PSNR ve SSIM gibi geleneksel tam referanslı yöntemlerin klinik gerçeklikle örtüşmediği görülmektedir; zira bu metrikler piksellerin matematiksel benzerliğine odaklanırken, radyologların algıladığı tanısal ayrıntıları ve semantik bilgiyi çoğu zaman göz ardı etmektedir. Öte yandan, referanssız (NR-IQA) çalışan derin öğrenme modelleri, radyolog skorlarıyla yüksek korelasyon kurabilse de, bu modellerin eğitilmesi için

devasa ve titizlikle etiketlenmiş veri setlerine ihtiyaç duyulmaktadır. Dahası, derin öğrenme modelleri bir görüntüye neden düşük kalite puanı verdiklerini tıbbi bir dille açıklayamazlar. Bu da klinik ortamda hekimin modele olan güvenini sınırlamaktadır. Bu noktada, hem sayısal skor üretebilen hem de bu skoru mantıksal ve dilsel bir çerçeveye oturtan Çok Modelli Büyük Dil Modelleri (LMM), modern radyolojide yeni bir çığır açma potansiyeline sahiptir.

Bu çalışmada, bu problemin çözümüne alternatif olarak geliştirilen LLM tabanlı bir görüntü kalitesi değerlendirme sistemi önerilmiştir. Yapay zekadaki son gelişmeler, modellerin sadece metin işlemesini değil, aynı zamanda görsel verileri de bir uzman gibi yorumlayabilmesini sağlamıştır. Önerilen yöntemde görüntü üzerindeki gürültü, bulanıklık ve kontrast kaybı gibi bozulmaları dikkate alarak oluşturulan klinik görüntülerin hem sayısal skorlamalarını hem de metinsel açıklamalarını üreten bir model geliştirilmiştir. Model, görüntülerdeki bozulmaları doğrudan metinsel ve sayısal geri bildirimlerle açıklar ve kalite farklılıklarını daha anlaşılır hale getirir. Oluşturulan bu modelin temel amacı, klinik uzmanlara tutarlı bir karar destek aracı sunmak ve değerlendirme sürecini daha nesnel bir yapıya kavuşturmaktır.

Çalışmanın metodolojik kurgusu, modelin gradyan tabanlı bir eğitime (retraining) ihtiyaç duymadan, sahip olduğu genel bilgiyi tıbbi bağlama nasıl aktarabileceği üzerine kurulmuştur. Bu kapsamda, sektör lideri OpenAI (GPT-o3, GPT-4o), Google (Gemini, Gemma), Meta (Llama), Anthropic (Claude) ve xAI (Grok) gibi farklı teknoloji sağlayıcılarının modelleri, LDCT-IQA bağlamında geniş bir alanda test edilmiştir.

Çalışmada kullanılan veri setinin niteliği, elde edilen sonuçların klinik geçerliliği açısından büyük önem taşımaktadır. Bu kapsamda kullanılan 1000 eğitim ve 300 test görüntüsünden oluşan MICCAI 2023 LDCT-IQA Challenge veri seti, Amerika Birleşik Devletleri'ndeki Mayo Clinic ve Güney Kore'deki National Cancer Center gibi prestijli kurumlardan elde edilen heterojen bir popülasyonu temsil etmektedir. Görüntülerdeki düşük doz koşulları, sadece basit bir gürültü ekleme işlemiyle değil, Poisson gürültüsü ve streak artefaktlarını içeren fizik tabanlı bir simülasyon hattı (pipeline) aracılığıyla gerçeğe en yakın şekilde üretilmiştir. Bu yöntem, LDCT görüntülerinde sıkça karşılaşılan ve tanısal doğruluğu etkileyen karmaşık bozulma paternlerinin model tarafından öğrenilmesine olanak tanımıştır. Ayrıca, 5-6 farklı uzman radyoloğun 0-4 Likert ölçeğinde verdiği puanların ortalamasıyla oluşturulan "ground-truth" değerleri, test setinde 0.96 gibi oldukça yüksek bir ICC (Intraclass Correlation Coefficient) değerine sahiptir; bu da modellerin eğitilmesi ve test edilmesi için kullanılan referans değerlerin bilimsel güvenilirliğini kanıtlamaktadır.

Önerilen yöntemde zero shot yaklaşımı, meta veri entegrasyonu ve hata geri bildirimi gibi farklı çıkarım stratejileri sistematik olarak incelenmiş ve bu yaklaşımların genel performansa nasıl katkı sağladığı en basit düzenden daha kapsamlı düzene doğru adım adım gösterilmiştir. İlk aşamada uygulanan "Zero-shot" çıkarımı, modellerin hiçbir ön bilgi veya örnek görmeden sadece test edilecek görüntüler verilmiştir. Önce fizibilite değerlendirmeleri için veri setinin küçük bir kısmına tabi tutulmuş yaptıkları tahminler değerlendirildiğinde tıbbi doğruluk için yeterli olmadığı gözlemlenmiştir. Bu nedenle zero-shot testleri tüm veri setine uygulanmamıştır. Bunu aşmak için geliştirilen "Few-shot" (az atışlı) öğrenme modelinde, prompt içerisine birkaç örnek görüntü-puan çifti eklenerek modelin radyolojik ölçekle hizalanması sağlanmıştır. Küçük veri setinde

iyi performans gösteren few-shot yaklaşımı ise veri seti genişletilerek çalışmalar bir adım öteye taşınmıştır. Ancak burada veri çeşitliliği sebebiyle genellenebilirliğin düşmesiyle birlikte performans düşüşü gözlemlenmiştir. Bu sebeple bu da diğer bir yeniliğe motivasyon kaynağı olmuştur.

Çalışmanın bu en önemli yeniliklerinden birisi ise, "Metadata Entegrasyonu"dur. Bu kapsamda, yine LLM' ler yardımıyla görüntünün ait olduğu anatomik bölge (örneğin karın, böbrek vb.) etiketi ve görüntü kalitesi değerlendirilmesinde en önemli metriklerden birisi olan gürültü seviyesi, FBI-Denoiser yöntemiyle hem test hem de örnek (train) görsellerine etiketlenerek girdi olarak verilmiştir. Bu bağlamsal veriler, modelin piksellerdeki değişimi tıbbi bir perspektifle yorumlamasına olanak tanımıştır. Son aşamada ise, modelin yaptığı hataları bir geri bildirim mekanizmasıyla analiz eden "Hata Geri Bildirimli Çıkarım" (Error Feedback-Guided Inference) stratejisi uygulanmıştır. Bu yöntemle modelin, geçmişteki yanlış tahminlerinden ders çıkararak gelecekteki skorlarını iyileştirmesi hedeflenmiştir. Bu stratejilerin ayrı ayrı sağladığı kazanımlar değerlendirilerek modelin hangi koşullarda daha güvenilir şekilde çıktı ürettiği ortaya konmuştur.

Elde edilen sonuçlara bakıldığında, önerilen sistemin yalnızca yüksek tutarlılıkta puanlar vermekle kalmadığı, aynı zamanda yorumlanabilir ve açık bir analiz sağladığı gösterilmiştir. Özellikle OpenAI o3 modelinin, meta veri ve hata geri bildirimi desteğiyle, radyolog puanlarıyla olan korelasyonunu (PLCC, SROCC ve KROCC metrikleri üzerinden) başlangıç seviyesinden çok daha yukarıya taşıdığı gözlemlenmiştir.

Modelin ürettiği metinsel açıklamalar, modelin görüntülere verdiği puanları tıbbi terminolojiye uygun şekilde tanımlayabilmesi amaçlanmıştır. Bu durum, sistemin sadece sayısal bir çıktı değil, hekim için açıklanabilir bir rapor sunulmasını da hedeflediğini ortaya koymaktadır. Bu yönüyle önerilen sistem ile düşük dozlu BT görüntülerinde kalite değerlendirmesine yönelik esnek, pratik ve gelecekteki klinik uygulamalara temel oluşturabilecek alternatif bir yapı sunulmuştur.

Sonuç olarak, bu araştırma LMM tabanlı yaklaşımların tıbbi görüntü kalitesi değerlendirmesi süreçlerindeki dönüştürücü potansiyelini kapsamlı bir şekilde ortaya koymaktadır. Geleneksel derin öğrenme modellerinin aksine, açıklanabilirlik ve uygulama esnekliğini odağına alan bu metodoloji, radyolojik değerlendirmelerde nesnel bir standardizasyonun sağlanması ve klinik iş yükünün minimize edilmesi yolunda stratejik bir adım teşkil etmektedir. Çalışmamız, yapay zekayı kapalı bir sistem olmaktan çıkarıp, sadece sayısal bir kalite puanı üretmekle kalmayıp, bu puanı destekleyen dilbilimsel ve anlamsal açıklamalar sunmuştur. Bu bağlamda, MICCAI 2023 gibi geniş ölçekli bir veri seti üzerinde, sektör lideri çok sayıda farklı modelle gerçekleştirilen testler, çalışmanın bilimsel derinliğini ve alanındaki öncü konumunu pekiştirmektedir. Sistemin sunduğu çeviklik, yeni klinik senaryolara ve farklı modalitelere hızla uyarlanabilmesine imkan tanıyabilecektir.

Günümüzde yapay zeka sağlayıcılarının modellerini her geçen gün geliştirdiği göz önüne alındığında, bu tezin sunduğu esnek yapı, gelecekteki teknolojik iyileşmelerle birlikte radyologların tanısal güvenini artıracak ve yapay zekanın tıbbi iş akışlarına tam entegrasyonu için sağlam bir temel oluşturacaktır.

1. INTRODUCTION

Computed tomography (CT) is one of the most commonly employed imaging modalities in contemporary medicine and it is essential to diagnostic accuracy and patient management. However, the hazards associated with ionizing radiation represent a serious challenge to patient safety. Thus, the development and utilization of Low-Dose Computed Tomography protocols has increased [1]. LDCT provides a value in that it utilizes minimal radiation doses, but it can be associated with changes in image quality, noise and artifacts, which have a very real impact on radiologists' diagnostic outcomes. Therefore, the dependable and high-quality evaluation of LDCT must be guaranteed for accurate clinical judgments.

Image quality relies on subjective assessment by radiologists. Although this represents the gold-standard approach, it has several limitations, including inter-observer variability and a time-consuming evaluation process. Automatically applied Image Quality Assessment (IQA) methods therefore have arisen as an extremely targeted research activity in recent years [2]–[6]. PSNR and SSIM, used in early full-reference approaches, often failed to capture diagnostic relevance, which led to the development of no-reference (NR) IQA methods.

Recently, in addition to these full-reference metrics, no-reference (NR) IQA approaches, and particularly deep learning-based methods, have also made significant progress. Various deep learning methods show high correlation with radiologists' scores [3]–[6]. However, these approaches require large and balanced datasets, which often results in generalization issues when imaging modalities change, thereby necessitating retraining.

In recent years, a significant paradigm shift in artificial intelligence has taken place. Large language models (LLM) and their multimodal variants achieve unique performance not only for language processing approaches but also for more complicated visual-text-based problems [7]. Because of their unique abilities in text generation, logical inference, and contextual understanding, LLMs can be an alternative to

traditional deep learning approaches with a more flexible and generalized solution. Multimodal LLMs (LMMs), thanks to their capacity to evaluate images alongside linguistic explanations, not only produce a score but also provide justifications like human. [8]–[10]. These features hold significant potential in terms of both explainability and relevance to clinical contexts in IQA tasks [11]–[14].

While several attempts at LMM-based IQA exist in the literature [7]–[9,15], to our knowledge no research has yet explored such a comprehensive dataset and multi-model comparison in the context of LDCT. Our study is based on a large-scale dataset with high confidence (ICC=0.96) consisting of 1000 training and 300 test images provided by the MICCAI 2023 LDCT-IQA Challenge [1].

In this study, we systematically investigate LMMs for image quality assessment on a large-scale LDCT dataset, introducing key innovations such as a no-reference evaluation setup, zero-shot and few-shot inference testing, an Error Feedback–Guided Inference strategy, and metadata-based explainability enhancement. Furthermore, the comparative evaluation of multiple AI models strengthened the reliability of the obtained results. In summary, these advancements introduce a flexible, training-free few-shot inference framework that is readily adaptable to clinical contexts and extends beyond conventional deep learning–based IQA methods.

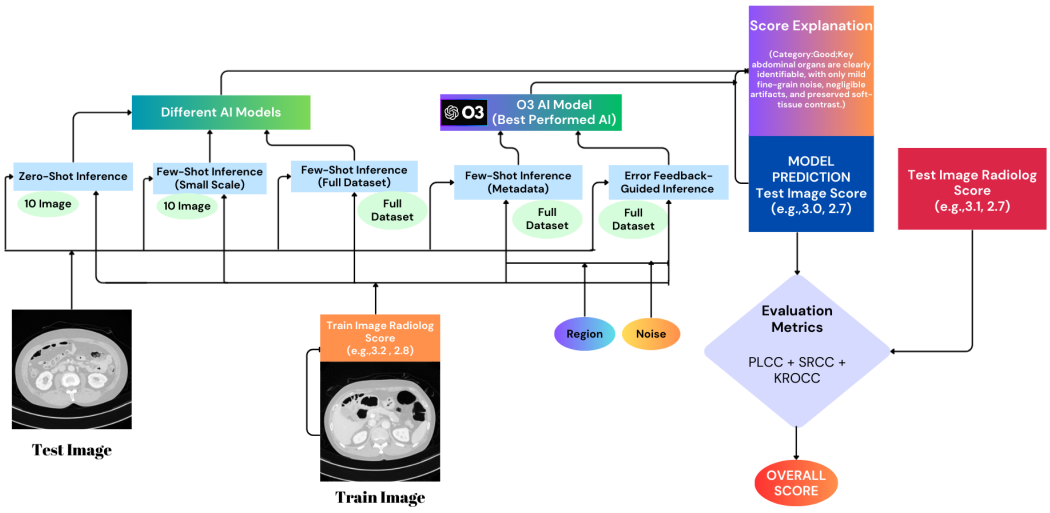

Figure 1.1: Flowchart of the LMM-IQA methodology.

2. METHODOLOGY

2.1 Dataset

In this study, we used the open-access dataset generated as part of the MICCAI 2023 Low-Dose Computed Tomography Perceptual Image Quality Assessment (LDCT-IQA) Challenge [1]. The dataset consists of 1,000 training and 300 test images. The images were sourced from the Mayo Clinic (USA) and the National Cancer Center (South Korea) and were obtained from different populations. A physics-based artifact addition pipeline was applied to the images to reflect low-dose conditions by introducing varying levels of Poisson noise and streak artifacts. Radiologists scored all images on a 0–4 Likert scale (0: very poor, 4: excellent). For each image, the average of the scores provided by 5–6 radiologists was used to create a ground-truth. These annotations were found to be highly reliable, with an intraclass correlation coefficient (ICC) value of 0.96 in the test set.

2.2 Evaluated Large Language and Multimodal Models

The study tested large-scale language and multimodal models from different vendors: OpenAI: GPT-o3, GPT-4o, GPT-4o-mini; Google: Gemini 2.5 Pro; Meta: Llama-4 Maverick; Qwen: VL Max; Anthropic: Claude Sonnet-4; xAI: Grok 2 Vision; Google DeepMind: Gemma 3. These models were tested in both zero-shot and few-shot scenarios, consistent with recent multimodal IQA benchmarks [8,12]–[14].

2.3 Proposed LMM-Based IQA Frameworks

Fig. 1.1 illustrates the general operation of the different approaches in the proposed LMM-IQA methods. The model follows a process that includes zero-shot, few-shot, metadata-aided, and error-feedback-guided inference steps. In these stages, the contextual accuracy of the predictions is increased by adding external region and noise information. Ultimately, the model produces both numerical quality scores

and descriptive text, demonstrating a gradual increase in correlation with radiologist assessments.

2.3.1 Zero-shot inference

Models scored test images without any sample images. Scoring was done with 10 test images. Results of clustering were saved to a separate file for analysis.

2.3.2 Few-shot inference (small-scale)

10 training images and their related radiologist scores were given to the model, and then 10 test images were requested as scored from model, consistent with few-shot setups in prior work [14].

2.3.3 Few-shot inference (full dataset)

In this setup, the models were evaluated on the entire dataset (1000 training and 300 test images) using a training-free few-shot inference paradigm. Rather than parameter optimization, a limited number of training examples were provided in-context to guide the model's reasoning. For each inference step, 10 representative examples were included in the prompt, while 34 examples were used in the final tests to further stabilize performance. Since there were 300 test images, the models produced one prediction per test image across 300 inference steps. In addition to numerical scores, the models generated brief textual explanations describing the criteria used for scoring, similar to the descriptive IQA frameworks proposed in [9].

2.3.4 Error feedback–guided inference

In this study, Error Feedback–Guided Inference approach was applied to model evaluation. In the first stage, the model evaluated several randomly selected training images with the corresponding noise level (n_i) that was generated from noise estimator and the true score (y_i) was determined by the radiologist. The absolute error, was calculated as the difference between the model's prediction ($\hat{y}_i$) and the true score (y_i).

$$e_i = |y_i - \hat{y}_i|, \tag{2.1}$$

These error values were provided as feedback to the model for each sample and were used as guidance for subsequent test samples. Thus, the model attempted to minimize the error in its new predictions by considering the errors in its past predictions. The reward–punishment loop can be defined similarly to the expected reward function in the classical Reinforcement Learning (RL) literature. The model's performance metric is to minimize the total error:

$$J(\theta) = \arg\min_{\theta} \mathbb{E}_{(x,y)\sim D}[\,|y - f_\theta(x)|\,] \tag{2.2}$$

where $(x,y) \sim D$ denotes the data sampled from the training distribution, and $f_\theta(x)$ represents the model's prediction parameterized by θ. The resulting e_i errors at each new prediction step are passed on to the model as a negative penalty. Unlike classical supervised learning, which relies solely on assigning correct labels, this method improves the model's decision-making process by converting error magnitude into information [13].

2.3.5 Metadata integration for context-aware scoring

Anatomical *Region* labels were added to each image to increase the model's scoring accuracy, enabling it to better understand structural differences and make more accurate comparisons between similar anatomical areas. For this purpose, the LDCT images were provided to the LLMs to determine the corresponding anatomical *Region* (e.g., abdomen, kidney, etc.). Both training and testing images were assigned with these labels and used as input in LMM few-shot scenarios. This aimed to allow the model to more accurately capture quality differences between the same anatomical regions.

Noise level is one of the most critical factors determining the quality of CT images. Therefore, Poisson-Gaussian-based noise estimation was performed for each image. No reference-noise metadata were estimated quickly and accurately using the Fast Blind Image Denoiser (FBI-Denoiser) method [16]. For example, values such as 0.003 or 0.005 were added to each image as noise metadata, and '*Region* + *Noise*' were incorporated into the LMM input to produce more consistent predictions with radiologist.

Consequently, the addition of Region information allowed the model to better understand structural differences, contributing to more accurate comparisons between similar anatomical areas. The addition of Noise information allowed it to account for noise variations inherent in low-dose CT and helped produce predictions more consistent with radiologist scores.

2.3.6 Evaluation metrics

Three correlation coefficients were used to evaluate model performance, measuring the relationship between predicted scores and radiologist scores: PLCC, SROCC, and KROCC. These metrics assess model reliability by capturing both linear correlation and rank-based similarities, in line with prior LDCT-IQA definitions [1].

PLCC (Pearson Linear Correlation Coefficient) measures the linear correlation between model predictions and radiologist scores. It is defined mathematically as follows:

$$\text{PLCC} = \frac{\sum(s_i - \mu)(\hat{s}_i - \hat{\mu})}{\sqrt{\sum(s_i - \mu)^2 \sum(\hat{s}_i - \hat{\mu})^2}}, \tag{2.3}$$

where s_i represents the score given by the radiologist, and $\hat{s}_i$ represents the score predicted by the model. The terms μ and $\hat{\mu}$ denote the averages of the radiologist and model scores, respectively.

SROCC (Spearman Rank Order Correlation Coefficient) measures the rank order similarity of scores and is used to evaluate the relative consistency between rankings:

$$\text{SROCC} = 1 - \frac{6\sum(R_i - \hat{R}_i)^2}{n(n^2 - 1)}, \tag{2.4}$$

where R_i and $\hat{R}_i$ represent the ranks of the actual and predicted scores, respectively, and n denotes the total number of images.

KROCC (Kendall Rank Correlation Coefficient) is calculated based on the difference between the concordant pairs (M_c) and discordant (M_d) pairs of two rankings:

$$\text{KROCC} = \frac{M_c - M_d}{\sqrt{(M_c + M_d + T_s)(M_c + M_d + T_t)}}, \tag{2.5}$$

where T_s and T_t represent the terms accounting for ties in the rankings.

In the final evaluation of the results, these three metrics were considered together, and their sum was reported as the *Overall Score*, defined as:

$$\text{Overall Score} = \text{SRCC} + \text{KROCC} + \text{PLCC} \tag{2.6}$$

Thus, model performance was comprehensively measured by considering both linear and rank-based similarities simultaneously. [1].

3. EXPERIMENTAL RESULTS

As shown in Fig. 3.1, the model outputs do not only provide a numerical image quality score, but also include textual explanations that clearly indicate which visual characteristics the score is based on. In the example, the dominant factors determining image quality are described in terms of noise level, the distinguishability of anatomical structures, and contrast properties. Such outputs make it possible to directly observe how the numerical scores relate to visual image quality, thereby enabling a more meaningful interpretation of the quantitative results in a clinical context.

```
[1/300] Evaluating test000.png (Region: abdomen, Noise: 0.003)...
Score (float): 3.2
Category: Good – Key abdominal organs are clearly identifiable, with only mild fine-grain noise,
negligible artifacts, and preserved soft-tissue contrast.

[2/300] Evaluating test001.png (Region: abdomen, Noise: 0.005)...
Score (float): 1.6
Category: Poor; Explanation: Marked graininess and conspicuous concentric/radial streak artifacts
obscure fine detail, soft-tissue contrast is reduced and organ margins (liver, kidneys, bowel, spine)
appear blurred though still recognizable, limiting confident diagnostic use.
```

Figure 3.1: Example output of LMM generated image quality assessment.

Table 3.1: Zero-shot inference results of different AI models

Model	PLCC	SROCC	KROCC	Overall Score
GPT-4o Scores	0.5015	0.2730	0.1673	0.9418
Claude-Sonnet4 Scores	0.4221	0.3597	0.2948	1.0766
Gemma3 Scores	0.4941	0.4912	0.4138	1.3990
Grok2 Scores	0.6179	0.5704	0.4800	1.6683
Qwen Scores	0.5796	0.6174	0.4945	1.6915
Gemini Scores	0.6729	0.6863	0.5196	1.8788
Meta-Llama4 Scores	0.6595	0.6743	0.5700	1.9038
O3 Scores	0.7521	0.6796	0.5512	1.9828
GPT-4o-mini Scores	0.7325	0.7082	0.5457	**1.9864**

3.1 Zero-Shot Inference Results

All AI models performed poorly in the zero-shot scenario as seen in Table 3.1. Although the GPT-4o-mini and O3 models produced relatively better results, a strong correlation was not established with radiologist scores. Therefore, given the low scores obtained

even with 10 test images, a zero-shot experiment was deemed unnecessary on all test images. This finding demonstrates that the zero-shot approach might not be a good candidate for medical IQA.

3.2 Few-Shot Inference Results (Small Scale - 10 Images)

The Small Scale Few Shot test results are shown in Table 3.2.In few-shot experiments with a limited number of images (10 training + 10 test), the model performance improved significantly. The top-performing models were O3, GPT-4o, and Qwen, with Overall Scores in the range of approximately 2.60–2.67. The medium-performing models included GPT-4o-mini, Gemini, and Llama4, achieving Overall Scores between 2.34–2.54. Finally, the low-performing models were Gemma-3, Grok-2, and Claude Sonnet-4. This result demonstrates that a small number of samples allows models to make predictions closer to radiologist scores.

Table 3.2: Few-shot inference results (small scale)

Model	PLCC	SROCC	KROCC	Overall Score
Gemma3 Scores	0.2848	0.4431	0.3568	1.0847
Grok2 Scores	0.6760	0.6896	0.5552	1.9208
Claude Sonnet4 Scores	0.8170	0.8006	0.6838	2.3014
Meta Llama4 Scores	0.8123	0.8115	0.7230	2.3468
Gemini Scores	0.8950	0.8329	0.7389	2.4669
GPT-4o-mini Scores	0.9222	0.8697	0.7483	2.5402
Qwen Scores	0.9095	0.8867	0.8040	2.6003
GPT-4o Scores	0.9606	0.8785	0.8015	2.6405
O3 Scores	0.9630	0.8927	0.8142	**2.6698**

3.3 Few-Shot Inference Results (Full Dataset – 1000 Training / 300 Test)

In experiments using the full dataset, the models were tested with 300 inference steps, with 10 training images per inference step and 1 test image as the output. However, performance decreased compared to the smaller-scale experiments. The initial test results are provided in Table 3.3.The performance degradation stems from the model's limited contextual capacity and generalization ability. While the model learned relationships easily with small data because the examples were similar, the increased diversity across the entire dataset (different regions and noise levels) disrupted this consistency. Because LMMs do not train, they only infer short contexts; this reduces

the representativeness of the examples in the large dataset, leading to a decrease in correlation. The top-performing models were O3, GPT-4o, and Gemini, with Overall Scores in the range of approximately 1.85–1.88. The medium-performing models included GPT-4o-mini and Llama4, achieving Overall Scores between 1.02–1.26. Finally, the low-performing model was Gemma-3. These findings suggest that LMM models, which achieve high correlation in small-scale tests, experience performance degradation on larger and more diverse datasets.

Table 3.3: Few-shot inference results (full dataset)

Model	PLCC	SROCC	KROCC	Overall Score
Gemma3 Scores	0.0677	0.0400	0.0300	0.1378
GPT-4o-mini Scores	0.3808	0.3664	0.2733	1.0205
Meta-Llama4 Scores	0.4765	0.4591	0.3261	1.2616
GPT-4o Scores	0.6781	0.6709	0.4996	1.8486
Gemini Scores	0.6758	0.6906	0.5093	1.8757
O3 Scores	0.6863	0.6854	0.5127	**1.8844**

3.4 Few-Shot Inference Results with Metadata

The performance degradation observed in the full-dataset experiments was addressed by incorporating the metadata-based enhancement strategies introduced earlier in the study. The resulting performance gains, obtained through these incremental strategies, are presented step by step in Table 3.4 . Furthermore, in the final tests conducted to maximize the performance, a maximum of 34 training images were given to the O3 model instead of 10 training images per inference step to allow the system to learn better. Tests were conducted using the entire dataset.

For both cost-effectiveness and optimal results achieved from previous experiments, test scenarios were developed using the O3 AI model in the optimization efforts. The final test results are shown in Table 3.4.

Table 3.4: O3 model few-shot inference results with different approaches

Model	PLCC	SROCC	KROCC	Overall Score
O3 Scores	0.6863	0.6854	0.5127	1.8844
O3-Metadata Scores	0.8137	0.8051	0.6347	2.2535
O3-Error-Feedback Scores	0.8122	0.8180	0.6347	**2.2649**

When the results were examined based on the O3 model, the score performance, which started at 1.88 was increased to the 2.25–2.26 range with the improved methods and metadata. This indicates that providing additional metadata to LMMs significantly increased performance.

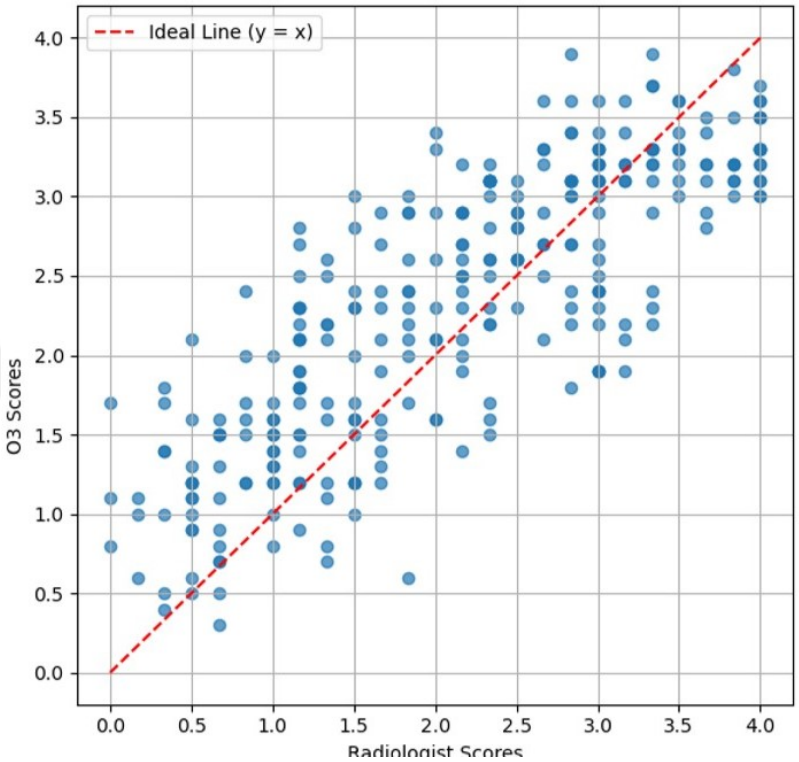

Figure 3.2: Scatter plot comparing radiologist scores with o3-error feedback scores.

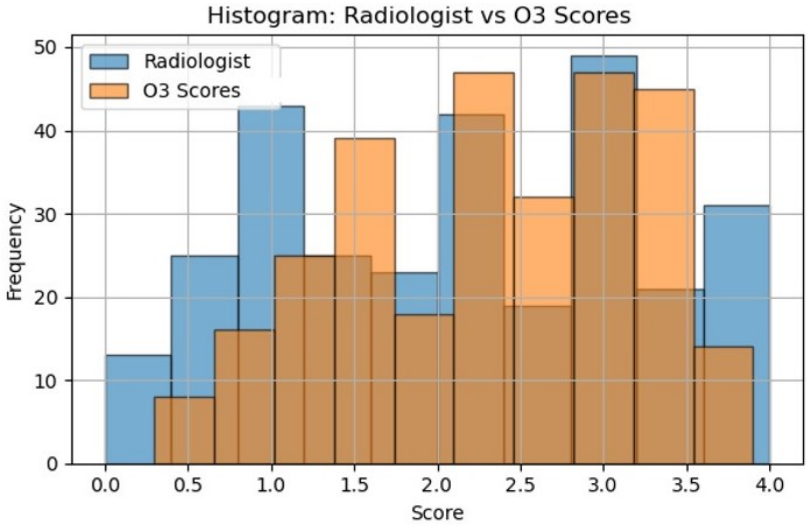

Figure 3.3: Distribution of radiologist vs o3-error feedback scores.

3.5 Error Feedback–Guided Inference Results

O3-Error Feedback consistently improved across all correlation metrics, yielding the best LMM-based result in this set of experiments with an Overall Score of 2.2649. The improvement is particularly pronounced in the order-sensitive metrics (SROCC and KROCC), indicating that reasoning patterns selected with reward/punishment can more consistently utilize the Region/Noise context, reducing outlier scores.

O3-Error Feedback demonstrated a strong correlation with radiologist scores, as confirmed by both shown in Fig. 3.2 and Fig. 3.3.

Distribution between radiologist scores and O3-Error Feedback scores shown in Fig. 3.2. The majority of points cluster near the ideal line ($y = x$), indicating a strong correlation between model scores and radiologist scores. The agreement is particularly good in the medium- and high-quality regions (range 2.0–3.5), where the model performs similarly to radiologists. Some deviations appear in the low-quality scores (range 0–1), where the model tends to rate images slightly higher than radiologists. This suggests a more tolerant behavior toward noise and artifacts.

Comparison of the score distributions for radiologists and O3-Error Feedback is shown in Fig. 3.3. The radiologist score distribution is broader and more spread out toward the extreme values (0 and 4), while the O3-Error Feedback distribution is concentrated in a narrower band (1–3.5). The two distributions match well, especially in the ranges 1.0–2.0 and 2.5–3.5. However, the model used very low (0–0.5) and very high (4.0) scores less frequently, suggesting that O3-Error Feedback is more cautious about assigning extreme values and generally leans toward mid-range scores.

4. CONCLUSION

In this thesis, a large multimodal model (LMM)-based methodology was developed for assessing quality in low-dose CT acquisitions, offering a flexible and training-free alternative to deep learning-based NR-IQA methods. While deep learning paradigms can generate accurate scores, they are often limited to numerical outputs without fully characterizing multidimensional perceptual factors or explaining the reasoning behind a specific rating. In this investigation, the LMM's linguistic inference ability was harnessed to address this limitation; the system both generated a score as well as identified quality degrading factors with textual linguistic explanations. By bridging the gap between purely mathematical metrics and complex radiological requirements, the new methodology generated both numerical as well as interpretative outputs which is more meaningful in the broad clinical context.

The robustness of the proposed framework was tested on a large-scale dataset from the MICCAI 2023 LDCT-IQA Challenge, representing diverse populations from the Mayo Clinic and the National Cancer Center. The high reliability of the ground-truth scores, evidenced by an Intraclass Correlation Coefficient (ICC) of 0.96, provided a solid foundation for evaluating the alignment between human perception and model predictions. Experimental results demonstrated strong correlation with radiologist annotations and improved robustness across anatomical regions when incorporating metadata and error feedback. The incremental performance gains observed throughout the study validate the effect of providing the model with context-aware.

Specifically, the integration of anatomical region labels and physical noise levels estimated via FBI-Denoiser allowed the model to better understand structural differences and account for noise variations inherent in low-dose CT. The application of the Error Feedback-Guided Inference strategy further refined these predictions by creating a reward-punishment loop that minimized outlier scores without the need for gradient-based weight updates. The best-performing configuration, the O3 model with

error feedback, achieved an Overall Score of 2.2649, showing that reasoning patterns can be iteratively improved through strategic prompting and feedback loops. Detailed analysis of score distributions revealed that while radiologists utilize the full 0-4 range, the LMM tends to be more cautious, concentrating its assessments in the mid-range (1.0-3.5) and avoiding extreme values unless the quality clearly warrants such a rating.

Moreover, the ability to identify and describe low-dose CT artifacts further supports clinical applicability from our method. By providing a reasoning layer alongside numerical scores, the system offers a complementary layer of interpretability for clinical workflows and future decision support systems.

In addition, our research paves the way for future studies. The demonstrated "training-free" nature of the LMM-IQA framework ensures its scalability to new imaging protocols and various vendors without the high computational cost of retraining traditional deep learning models. More extensive and varied datasets, further modality data and the potential for LMMs to work interactively with radiologists may yield quality assessment that is better, more reliable, and standardized, and add new insights.

REFERENCES

[1] **Lee, W.**, **Wagner, F.**, **Galdran, A.**, **Shi, Y.**, **Xia, W.**, **Wang, G.**, **Mou, X.**, **Ahamed, M.A.**, **Imran, A.A.Z.**, **Oh, J.E.** *et al.* (2025). Low-dose computed tomography perceptual image quality assessment, *99*, 103343.

[2] **Bosse, S.**, **Maniry, D.**, **Müller, K.R.**, **Wiegand, T. and Samek, W.** (2017). Deep neural networks for no-reference and full-reference image quality assessment, *27*(1), 206–219.

[3] **Zhu, H.**, **Li, L.**, **Wu, J.**, **Dong, W. and Shi, G.** (2020). MetaIQA: Deep meta-learning for no-reference image quality assessment, 14143–14152.

[4] **Yang, S.**, **Wu, T.**, **Shi, S.**, **Lao, S.**, **Gong, Y.**, **Cao, M.**, **Wang, J. and Yang, Y.** (2022). Maniqa: Multi-dimension attention network for no-reference image quality assessment, pp.1191–1200.

[5] **Madhusudana, P.C.**, **Birkbeck, N.**, **Wang, Y.**, **Adsumilli, B. and Bovik, A.C.** (2022). Image quality assessment using contrastive learning, *31*, 4149–4161.

[6] **Talebi, H. and Milanfar, P.** (2018). NIMA: Neural image assessment, *27*(8), 3998–4011.

[7] **Wang, J.**, **Jiang, H.**, **Liu, Y.**, **Ma, C.**, **Zhang, X.**, **Pan, Y.**, **Liu, M.**, **Gu, P.**, **Xia, S.**, **Li, W.** *et al.* (2024). A comprehensive review of multimodal large language models: Performance and challenges across different tasks.

[8] **You, Z.**, **Li, Z.**, **Gu, J.**, **Yin, Z.**, **Xue, T. and Dong, C.** (2024). Depicting beyond scores: Advancing image quality assessment through multi-modal language models, Springer, pp.259–276.

[9] **You, Z.**, **Gu, J.**, **Cai, X.**, **Li, Z.**, **Zhu, K.**, **Dong, C. and Xue, T.** Enhancing Descriptive Image Quality Assessment with a Large-scale Multi-modal Dataset.

[10] **Chen, Z.**, **Zhang, X.**, **Li, W.**, **Pei, R.**, **Song, F.**, **Min, X.**, **Liu, X.**, **Yuan, X.**, **Guo, Y. and Zhang, Y.** (2024). Grounding-iqa: Multimodal language grounding model for image quality assessment.

[11] **Zhu, H.**, **Wu, H.**, **Li, Y.**, **Zhang, Z.**, **Chen, B.**, **Zhu, L.**, **Fang, Y.**, **Zhai, G.**, **Lin, W. and Wang, S.** (2024). Adaptive image quality assessment via teaching large multimodal model to compare, *37*, 32611–32629.

[12] **Varga, D.** (2025). Comparative Evaluation of Multimodal Large Language Models for No-Reference Image Quality Assessment with Authentic Distortions: A Study of OpenAI and Claude. AI Models, *9*(5), 132.

[13] **Wu, H.**, **Zhang, Z.**, **Zhang, W.**, **Chen, C.**, **Liao, L.**, **Li, C.**, **Gao, Y.**, **Wang, A.**, **Zhang, E.**, **Sun, W.** *et al.* (2023). Q-align: Teaching lmms for visual scoring via discrete text-defined levels.

[14] **You, Z.**, **Cai, X.**, **Gu, J.**, **Xue, T. and Dong, C.** (2025). Teaching large language models to regress accurate image quality scores using score distribution, pp.14483–14494.

[15] **Chen, Z.**, **Hu, B.**, **Niu, C.**, **Chen, T.**, **Li, Y.**, **Shan, H. and Wang, G.** (2024). IQAGPT: computed tomography image quality assessment with vision-language and ChatGPT models, *7*(1), 20.

[16] **Byun, J.**, **Cha, S. and Moon, T.** (2021). Fbi-denoiser: Fast blind image denoiser for poisson-gaussian noise, pp.5768–5777.

CURRICULUM VITAE

Kağan ÇELİK

EDUCATION:

- **B.Sc.:** 2022, Istanbul Kültür University, Faculty of Engineering, Department of Electrical and Electronics Engineering
- **M.Sc.:** 2026, Istanbul Technical University, Faculty of Electrical and Electronics Engineering, Department of Electronics Engineering (Thesis Programme)

PROFESSIONAL EXPERIENCE:

- **Turkish Airlines – Turkish Cargo**
  Data Analyst & IT Projects Specialist (09.2022–11.2024)
  Worked on software–focused projects to improve logistics and warehouse process efficiency using data–driven decision support and automation systems. Analyzed large datasets for KPI monitoring, and contributed to the optimization of storage operations.