arXiv:2506.23298v3 [eess.IV] 17 Jul 2025

Exposing and Mitigating Calibration Biases and Demographic Unfairness in MLLM Few-Shot In-Context Learning for Medical Image Classification

Xing Shen[1,2], Justin Szeto[1,2], Mingyang Li[3], Hengguan Huang[4], and Tal Arbel[1,2]

[1] Centre for Intelligent Machines, McGill University, Montreal, Canada
xing.shen@mail.mcgill.ca, tal.arbel@mcgill.ca
[2] Mila – Quebec AI Institute, Montreal, Canada
[3] Stanford University, Stanford, USA
[4] University of Copenhagen, Copenhagen, Denmark
hengguan.huang@sund.ku.dk

Abstract. Multimodal large language models (MLLMs) have enormous potential to perform few-shot in-context learning in the context of medical image analysis. However, safe deployment of these models into real-world clinical practice requires an in-depth analysis of the accuracies of their predictions, and their associated calibration errors, particularly across different demographic subgroups. In this work, we present the first investigation into the calibration biases and demographic unfairness of MLLMs' predictions and confidence scores in few-shot in-context learning for medical image classification. We introduce **CALIN**, an inference-time calibration method designed to mitigate the associated biases. Specifically, CALIN estimates the amount of calibration needed, represented by calibration matrices, using a bi-level procedure: progressing from the population level to the subgroup level prior to inference. It then applies this estimation to calibrate the predicted confidence scores during inference. Experimental results on three medical imaging datasets: PAPILA for fundus image classification, HAM10000 for skin cancer classification, and MIMIC-CXR for chest X-ray classification demonstrate CALIN's effectiveness at ensuring fair confidence calibration in its prediction, while improving its overall prediction accuracies and exhibiting minimum fairness-utility trade-off. Our codebase can be found at https://github.com/xingbpshen/medical-calibration-fairness-mllm.

Keywords: Fairness · Bias · Confidence calibration · Uncertainty · Foundation models · Large language models

1 Introduction

Image-text to text foundation models, particularly multimodal large language models (MLLMs, or referred to as large multimodal models, LMMs), such as

OpenAI GPT-4o and Google Gemini [8,19], have demonstrated strong generalization capabilities and achieved state-of-the-art performance across numerous tasks. Furthermore, advances in few-shot in-context learning (FS-ICL) enables MLLMs to solve new tasks by simply being *prompted* with a few examples of question-answer pairs [1,22,2]. The success of MLLMs and FS-ICL methods has led to applications in medical imaging contexts, including cancer pathology classification [4], where they have shown promising results while reducing or eliminating the need for the extensive training or fine-tuning typically required by traditional deep learning methods. However, in the context of medical imaging, ensuring debiased and fair machine learning models, particularly with respect to both prediction utility and confidence calibration across different demographic subgroups, is essential in order to safely deploy these models in real clinical contexts [28,9,18]. The associated risks include trusting prediction uncertainties that can potentially indicate high confidence in wrong assertions, for example, or presenting disparities in model performance across groups which can lead to potential harm to underrepresented groups. Despite these risks, investigations into calibration biases in MLLMs under FS-ICL setting, as well as strategies to accurately overcome their errors and biases in medical imaging, remains unexplored. This limits their practical use and reliability in real-world clinical settings.

Enforcing calibration fairness under FS-ICL setting poses unique methodological challenges. The lack of an additional training/validation set with an adequate amount of labeled data for different subgroups renders widely adopted optimization-based calibration methods impractical [12,5]. In addition, the most powerful state-of-the-art MLLMs are typically large-scale black-box models (e.g., GPT-4o, Gemini 1.5, Claude 3.5 Sonnet), making debiasing methods requiring additional access of their internal parameters infeasible [7].

In order to fill the gap and enable the trustworthy deployment of FS-ICL methods, this work investigates the calibration unfairness of MLLM under FS-ICL, exposing their biases and limitations in the context of medical image classification. To address current challenges, we propose **CALIN**, a novel training-free algorithm that automatically calibrates MLLM's predictions and their associated confidence scores, and enforces fairness across demographic subgroups at inference. CALIN (see Fig. 1) uses a *bi-level* procedure: progressing from the *population level* to the *subgroup level*, ensuring an accurate and stable adjustment estimation procedure for fair calibration across subgroups. Extensive experiments are performed on three publicly available medical imaging datasets–PAPILA [11] for fundus image classification, HAM10000 [21] for skin cancer classification, and MIMIC-CXR [10] for chest X-ray classification. Experimental results expose calibration biases in the MLLM under FS-ICL, and validate CALIN's effectiveness at: (i) mitigating the calibration gap between demographic subgroups, (ii) providing more reliable confidence scores over the entire population, (iii) improving prediction accuracies, and (iv) exhibiting a minimum fairness-utility trade-off. Detailed ablation studies further validate the necessity of the bi-level method in producing reliable and fair confidence calibrations.

2 Background on FS-ICL and Calibration Biases

We formally define the few-shot in-context learning (FS-ICL) setting and demographic calibration biases. At inference, a few-shot exemplar dataset with N samples (e.g., $N \leq 5$) is presented, represented as 3-tuples $\mathcal{D}_{\text{fs}} := \{(X_i, A_i, Y_i)\}_{i=1}^{N}$, where X_i is a random variable representing the medical image of the patient, A_i is a random variable representing the sensitive attribute (e.g., sex, age) of that patient, Y_i is a random variable representing the label of the image. Every tuple in $\mathcal{D}_{\text{fs}}$ follows the same task distribution denoted as $(X_i, A_i, Y_i) \sim P_\tau(X, A, Y)$. Given a new query $(X, A) \sim P_\tau(X, A)$ and a predictive model $f(\cdot)$ with fixed parameters, the new prediction $\hat{Y}$ from few-shot in-context learning is $\hat{Y} = f(\mathcal{D}_{\text{fs}}, X, A)$.

The demographic calibration bias can be defined as the confidence calibration error gap (CCEG, Δ_ε) between subgroups under a sensitive attribute [9]. Formally, for a given demographic attribute A, the gap Δ_ε can be expressed as:

$$\varepsilon(a) = \mathbb{E}\left[\left|\Pr[Y = \hat{Y} \mid \hat{p}, A = a] - \hat{p}\right|\right], \tag{1}$$

$$\Delta_\varepsilon(A) = \mathop{\mathbb{E}}_{(a,b)\sim\mathcal{U}(\{(a,b)|a,b\in\mathcal{A},a\neq b\})}\left[|\varepsilon(a) - \varepsilon(b)|\right], \tag{2}$$

where $\hat{p}$ is the predicted confidence for the prediction $\hat{Y}$, Y is the ground-truth label, $\mathcal{A} = \text{Val}(A)$ is the support of A, and (a, b) is a 2-tuple of values sampled uniformly from $\mathcal{A} \times \mathcal{A}$ such that $a \neq b$. A perfectly fair model has $\Delta_\varepsilon(A) = 0$.

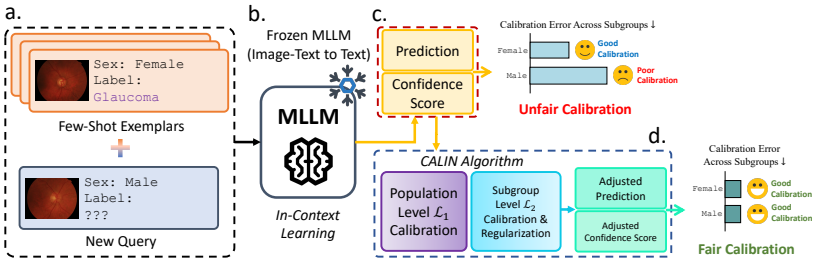

Fig. 1. Overview of CALIN for medical image classification with FS-ICL: (a) The MLLM takes as input a set of few-shot exemplars, each comprising an image, an associated attribute, and a label, along with a new query image and its attribute for label prediction. (b) MLLM predicts the label for the query image and the associated confidence score is calculated. (c) The predictions from the MLLM exhibit confidence calibration biases, leading to demographic disparities. (d) CALIN adjusts the confidence scores to mitigate calibration errors and improves fairness across demographic groups.