

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 Li³, Hengguan Huang⁴, 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: PAPA 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-mlm>.

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}_{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:

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

$$\Delta_{\varepsilon}(A) = \mathbb{E}_{(a,b) \sim \mathcal{U}(\{(a,b) \mid a,b \in \mathcal{A}, a \neq b\})} [\varepsilon(a) - \varepsilon(b)], \quad (2)$$

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 $\Delta_{\varepsilon}(A) = 0$.

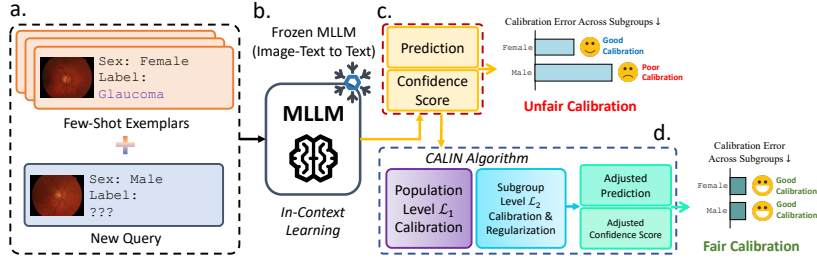


Fig. 1. Overview of CALIN for medical image classification with FS-ICL: (a) The MLLM takes as input a set of few-shot exemplars, each comprising an image, an associated attribute, and a label, along with a new query image and its attribute for label prediction. (b) MLLM predicts the label for the query 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.

3 CALIN: Intergroup Confidence Alignment From Null-Input Calibration

To overcome calibration errors and biases under the FS-ICL setting and to ensure calibration fairness among subgroups, we propose **CALIN**, an *inference-time calibration* method that contains a bi-level procedure – from *population-level* to *subgroup-level*. The goal is to provide fair and reliable confidence without requiring an additional training/validation set or access to the MLLM’s parameters.

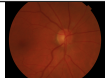
3.1 Notations

We assume that the predictive model is implemented by a pretrained frozen multimodal large language model $f_{\text{MLLM}}(\cdot)$ (e.g., GPT-4o and Gemini-1.5 [8,19]) that takes as input a set of *multimodal prompts* (image and text). We define a template φ that has fields for an image X , attributes A , and the label Y , though some may be left empty, to generate multimodal prompts. For example, $\varphi(X = \mathbf{x}, A = \text{Male}, Y = \text{Negative})$ is mapped to: “Does the fundus \mathbf{x} of a male show glaucoma? Negative” (see Table. 1 for more examples). During inference, the model is provided with the multimodal prompt for the new query $\varphi(X = \mathbf{x}, A = a, \cdot)$ along with FS-ICL (few-shot) exemplars $\mathbf{D} := \{(X_i = \mathbf{x}_i, A_i = a_i, Y_i = y_i) | (X_i, A_i, Y_i) \in \mathcal{D}_{\text{fs}}\}$. The MLLM’s predicted probability for \hat{Y} being y given the inputs is denoted $\hat{p}_y(\mathbf{D}, \mathbf{x}, a)$ and estimated as follows:

$$\underbrace{\Pr[\hat{Y} = y | \mathbf{D}, X = \mathbf{x}, A = a]}_{\hat{p}_y(\mathbf{D}, \mathbf{x}, a)} = \frac{\Pr[\hat{T} = y | \mathbf{D}, X = \mathbf{x}, A = a]}{\sum_{y_j \in \mathcal{Y}} \Pr[\hat{T} = y_j | \mathbf{D}, X = \mathbf{x}, A = a]}. \quad (3)$$

Here, $\hat{T} = f_{\text{MLLM}}(\{\varphi(X_i, A_i, Y_i) | (X_i, A_i, Y_i) \in \mathcal{D}_{\text{fs}}\} \cup \{\varphi(X, A, \cdot)\})$ is a random variable denoting the predicted next-token, and $\mathcal{Y} = \text{Val}(Y)$. We additionally define a vector $\hat{\mathbf{p}}_{\mathcal{Y}}(\mathbf{D}, \mathbf{x}, a) \in \mathbb{R}^{|\mathcal{Y}|}$, where each dimension represents the probability of the prediction belonging to a specific class $y_j \in \mathcal{Y}$.

Table 1. Multimodal prompts φ under different inputs for fundus image classification. The left illustrates a datapoint containing the fundus image $X = \mathbf{x}$, the value of the attribute $A = \text{Male}$, and the label $Y = \text{Negative}$. The right illustrates the corresponding prompts. For cases $\varphi(\cdot, A, \cdot)$ and $\varphi(\cdot, \cdot, \cdot)$, we do not input the image to the MLLM.

Example Prompts φ	
 Male with no glaucoma	$\varphi(X, A, Y)$ Does the fundus of a male show glaucoma? Negative
	$\varphi(X, A, \cdot)$ Does the fundus of a male show glaucoma?
	$\varphi(\cdot, A, \cdot)$ Does an arbitrary fundus of a male show glaucoma?
	$\varphi(\cdot, \cdot, \cdot)$ Does an arbitrary fundus show glaucoma?

3.2 Bi-Level Confidence Calibration

The bi-level procedure used by CALIN can be intuitively thought of as first inferring the “amount of calibration” needed for the entire population (*population-level*), then inferring the “coarse” amount of calibration needed for each subgroup (*subgroup-level*). Information flows from the upper *population-level* to regularize the lower *subgroup-level* to provide an accurate and fair confidence calibration.

Population-Level Calibration \mathcal{L}_1 . Inspired by the findings of language model’s prediction bias presented in [27,6], CALIN first infers the “amount of calibration” for the entire population to avoid prediction bias under FS-ICL. In this work, the amount of population-level calibration is defined by a diagonal matrix $\mathbf{U} \in \mathbb{R}^{|\mathcal{Y}| \times |\mathcal{Y}|}$ (we call it *calibration matrix* in this work), then the softmaxed linear transformation of $\hat{\mathbf{p}}_{\mathcal{Y}}(\mathbf{D}, \mathbf{x}, a)$, determined by \mathbf{U} , is the \mathcal{L}_1 post-calibration confidence, given by $\tilde{\mathbf{p}}_{\mathcal{Y}}(\mathbf{D}, \mathbf{x}, a) = \text{softmax}(\mathbf{U}\hat{\mathbf{p}}_{\mathcal{Y}}(\mathbf{D}, \mathbf{x}, a))$.

To determine \mathbf{U} without the need of extra training/validation set, CALIN adopts a *multimodal null-input probing* technique. Specifically, we ensure that the predicted confidence $\hat{\mathbf{p}}_{\mathcal{Y}}(\mathbf{D}, \mathbf{x}, a)$ is aligned with a uniform distribution when a null (or “content-free”, “semantic-free” [27,15]) query $\varphi(\cdot, \cdot, \cdot)$ is fed into the MLLM. For a concrete binary classification example in Table. 1, when we neither provide the fundus image nor specify the sex of the patient, the MLLM’s predicted confidence distribution should be uniform⁵. To this end, \mathbf{U} is calculated based on the observed predicted confidence $\hat{\mathbf{p}}_{\mathcal{Y}}(\mathbf{D}, \cdot, \cdot)$ by the MLLM when we send null query $\varphi(\cdot, \cdot, \cdot)$ to it, given by $\mathbf{U} = (\text{diag}(\hat{\mathbf{p}}_{\mathcal{Y}}(\mathbf{D}, \cdot, \cdot)))^{-1}$.

Subgroup-Level Calibration \mathcal{L}_2 . \mathcal{L}_1 improves confidence calibration over the entire population. To capture the potential variations across subgroups, we propose *subgroup-wise multimodal null-input probing* which aims to infer a set of calibration matrices $S := \{\mathbf{S}_a | a \in \mathcal{A}\}$ for \mathcal{L}_2 calibration. Each matrix in S focuses on calibrating one specific subgroup with sensitive attribute $A = a$. Borrowing from the intuition of multimodal null-input probing, subgroup-wise multimodal null-input probing finds S such that the predicted confidence given an attribute-conditioned null query $\varphi(\cdot, A = a, \cdot)$ is uniform for all subgroups. Specifically, we calculate them based on the observed predicted confidence $\hat{\mathbf{p}}_{\mathcal{Y}}(\mathbf{D}, \cdot, a)$ by the MLLM, given by $\mathbf{S}_a = (\text{diag}(\hat{\mathbf{p}}_{\mathcal{Y}}(\mathbf{D}, \cdot, a)))^{-1}$ for all $a \in \mathcal{A}$. Then, $\tilde{\mathbf{p}}_{\mathcal{Y}}(\mathbf{D}, \mathbf{x}, a) = \text{softmax}(\mathbf{S}_a \hat{\mathbf{p}}_{\mathcal{Y}}(\mathbf{D}, \mathbf{x}, a))$ is the \mathcal{L}_2 post-calibration confidence for any new query.

Regularizing \mathcal{L}_2 with \mathcal{L}_1 . While \mathcal{L}_2 calibration aims to achieve subgroup level confidence alignment, relying solely on \mathcal{L}_2 may not guarantee accurate calibration. This is because the language model’s inherent prompt bias [26,3] can lead to inaccurate and unstable estimation of calibration matrices, particularly since the \mathcal{L}_2 ’s probing prompt $\varphi(\cdot, A, \cdot)$ includes additional semantic information by conditioning on sensitive attributes. To mitigate this issue, we leverage \mathcal{L}_1 as

⁵ We assume that it is impossible to identify the ground-truth label without observing the medical image \mathbf{x} .

a regularization mechanism, allowing the final calibration to capture subgroup variability and also penalizing anomalies. Specifically, we calculate a new set of calibration matrices $C := \{C_a | a \in \mathcal{A}\}$ using exponential decay: When the estimated \mathcal{L}_2 calibration \mathbf{S}_a extremely diverges (due to unstable estimation) from \mathcal{L}_1 calibration \mathbf{U} , the final calibration will be more aligned with \mathcal{L}_1 , otherwise, the final calibration will be more aligned with \mathcal{L}_2 . The decay rate is governed by $(\sqrt{\alpha} + 1)^{-1}$ where α is the maximum observed deviation across subgroups, calculated by $\alpha = \max_a \{\|\mathbf{S}_a \mathbf{i} - \mathbf{U} \mathbf{i}\|_\infty\}$ where $\|\cdot\|_\infty$ denotes the infinity-norm. The final calibration matrices are given by:

$$\mathbf{c}_a = \mathbf{U} \mathbf{i} + (\mathbf{S}_a \mathbf{i} - \mathbf{U} \mathbf{i}) \odot \exp\left(-(\sqrt{\alpha} + 1)^{-1} \cdot \|\mathbf{S}_a \mathbf{i} - \mathbf{U} \mathbf{i}\|_\infty\right), \quad (4)$$

$$\mathbf{C}_a = \text{diag}(\mathbf{c}_a), \quad \forall a \in \mathcal{A}, \quad (5)$$

where $\mathbf{i} = \mathbf{1}_{|\mathcal{Y}|}$ is a vector with $|\mathcal{Y}|$ ones, \odot denotes the element-wise product. We obtain the post-calibration confidence $\hat{\mathbf{p}}_{\mathcal{Y}}(\mathbf{D}, \mathbf{x}, a) = \text{softmax}(\mathbf{C}_a \hat{\mathbf{p}}_{\mathcal{Y}}(\mathbf{D}, \mathbf{x}, a))$. Given the denoted vector construction $\hat{\mathbf{p}}_{\mathcal{Y}}(\mathbf{D}, \mathbf{x}, a) = [\hat{p}_{y_j}(\mathbf{D}, \mathbf{x}, a) | y_j \in \mathcal{Y}]$ we can get the adjusted predicted label $\tilde{y} = \arg \max_{y_j \in \mathcal{Y}} \{\hat{p}_{y_j}(\mathbf{D}, \mathbf{x}, a)\}$. The algorithm of CALIN is shown in Algorithm 1.

Algorithm 1 CALIN for Fair Confidence Calibration Under FS-ICL

Require: Few-shot \mathbf{D} , model f_{MLLM} , prompt template φ , demographic values \mathcal{A}

Ensure: Calibration matrices C

- 1: Compute $\hat{\mathbf{p}}_{\mathcal{Y}}(\mathbf{D}, \cdot, \cdot)$ using (3) with f_{MLLM} , \mathbf{D} , $\varphi(\cdot, \cdot, \cdot)$
- 2: Compute $\mathbf{U} = (\text{diag}(\hat{\mathbf{p}}_{\mathcal{Y}}(\mathbf{D}, \cdot, \cdot)))^{-1}$ #Population-Level#
- 3: **for** a in \mathcal{A} **do**
- 4: Compute $\hat{\mathbf{p}}_{\mathcal{Y}}(\mathbf{D}, \cdot, a)$ using (3) with f_{MLLM} , \mathbf{D} , $\varphi(\cdot, A = a, \cdot)$
- 5: Compute $\mathbf{S}_a = (\text{diag}(\hat{\mathbf{p}}_{\mathcal{Y}}(\mathbf{D}, \cdot, a)))^{-1}$ #Subgroup-Level#
- 6: **end for**
- 7: Compute \mathbf{C}_a using (4) and (5) with \mathbf{U} and \mathbf{S}_a , add \mathbf{C}_a to C . For all $a \in \mathcal{A}$
- 8: **return** C

Require: Few-shot \mathbf{D} , new query medical image \mathbf{x}^* , demographic value $a^* \in \mathcal{A}$, model f_{MLLM} , prompt template φ , calibration matrix $\mathbf{C}_{a^*} \in C$

Ensure: Adjusted prediction \tilde{y} and its calibrated confidence \tilde{p}

- 9: Compute $\hat{\mathbf{p}}_{\mathcal{Y}}(\mathbf{D}, \mathbf{x}^*, a^*)$ using (3) with f_{MLLM} , \mathbf{D} , $\varphi(X = \mathbf{x}^*, A = a^*, \cdot)$
 - 10: Compute $\hat{\mathbf{p}}_{\mathcal{Y}}(\mathbf{D}, \mathbf{x}^*, a^*) = \text{softmax}(\mathbf{C}_{a^*} \hat{\mathbf{p}}_{\mathcal{Y}}(\mathbf{D}, \mathbf{x}^*, a^*))$ #Inference-Time#
 - 11: Assign vector elements $[\hat{p}_{y_j}(\mathbf{D}, \mathbf{x}^*, a^*) | y_j \in \mathcal{Y}] = \hat{\mathbf{p}}_{\mathcal{Y}}(\mathbf{D}, \mathbf{x}^*, a^*)$
 - 12: **return** $\tilde{y} = \arg \max_{y_j \in \mathcal{Y}} \{\hat{p}_{y_j}(\mathbf{D}, \mathbf{x}^*, a^*)\}$ and $\tilde{p} = \hat{p}_{\tilde{y}}(\mathbf{D}, \mathbf{x}^*, a^*)$
-

4 Experiments and Results

Experiments are designed to showcase the effectiveness of CALIN in mitigating confidence calibration bias in MLLM under FS-ICL on 3 medical imaging datasets: (i) PAPILA [11], (ii) HAM10000 [21], (iii) MIMIC-CXR [10].

Datasets, Configuration and Implementation Details. The PAPILA dataset [11] is a glaucoma classification dataset consisting of patient fundus images, along with their sex and ages. 364 patients are randomly chosen as the test set, including 118 male and 246 female patients, with 146 young (age < 60) patients, and 218 elder (age ≥ 60) patients. The images are binary-labeled, indicating the diagnosis of glaucoma. The HAM10000 dataset [21] is a large-scale skin lesion classification dataset, consisting of dermatoscopic images of pigmented skin lesions, along with the patients’ sex and ages. 1,062 patients are randomly chosen for the test set, including 566 male and 496 female patients, with 472 young patients and 590 elder patients. A binary label indicates the diagnosis as malignant or benign. The MIMIC-CXR dataset [10] is large-scale dataset of chest radiographs with structured labels. 1,062 randomly chosen patients make up the test set, including 547 male and 488 female patients, with 439 young patients and 623 elder patients. A binary label indicates the diagnosis of pleural effusion. All experiments are conducted using GPT-4o-mini. Human review of input data is disabled on Azure OpenAI Service to comply with PhysioNet’s guidelines for responsible use of MIMIC-CXR with GPT [17]. We treat both sex and age as sensitive attributes. In each context, 4 additional patients are randomly selected from the dataset, apart from the test set, to serve as few-shot exemplars.

4.1 Main Results

We consider 5 different metrics in the experiments: (i) classification accuracy (**Acc.**), (ii) expected calibration error (**ECE**) [5] for quantifying the reliability of predicted confidence, (iii) mean equalized odd ratio [16] between sex and age (**EOR**) for fairness evaluation, (iv) confidence calibration error gap (**CCEG**, Δ_ϵ) for calibration fairness evaluation, and (v) equity-scaling measure [14,20,9] of calibration error (**ESCE**) for accessing the overall calibration performance adjusted by subgroups’ performance. ECSE also quantifies the fairness-utility trade-off. Note that for CCEG, which is the main focus of this work, we consider fairness for the attributes of **sex**, **age**, and intersectional (**Inter.**) fairness [25] of both attributes. Due to the lack of other valid baseline methods in this new problem setting, we compare the proposed method, CALIN, with the vanilla FS-ICL [2]. Experiments on both methods were performed with the same exemplars and queries. Given GPT’s inherent stochasticity, and any future updates to GPT, might result in slight variability in the exact metric values. Experimental results in Table. 2 and in Fig. 2 indicate that CALIN consistently outperforms the vanilla method on all metrics (metric values are scaled by $\times 10^2$). Specifically, CALIN improves confidence calibration across the entire population, as indicated by a substantial reduction in ECE. More importantly, as evidenced by the notable decrease in CCEG across all datasets, CALIN effectively mitigates confidence calibration bias associated with demographic attributes. This is particularly evident for age and attribute intersection, where vanilla FS-ICL struggles with fairness issues across these demographic groups. In Fig. 2, CALIN demonstrates superior performance in the equity-scaling measure (ESCE), validating its minimal fairness-utility trade-off.

Table 2. Main results for 3 datasets. The proposed method consistently outperforms the vanilla FS-ICL [2] baseline method according to all metrics, especially in terms of calibration across the population (ECE) and fair calibration across subgroups (CCEG).

Method	Acc. \uparrow	ECE \downarrow	$\overline{\text{EOR}} \uparrow$	CCEG $\Delta_\epsilon \downarrow$		
				Sex	Age	Inter.
<i>PAPILA</i> [11]						
Vanilla	78.30	19.13	20.00	4.84	19.37	15.15
CALIN (proposed)	78.57	5.97	34.38	1.53	9.52	6.14
<i>HAM10000</i> [21]						
Vanilla	74.76	23.70	70.51	5.01	30.25	20.66
CALIN (proposed)	74.76	2.68	74.24	4.43	3.14	3.11
<i>MIMIC-CXR</i> [10]						
Vanilla	66.38	28.09	59.48	4.92	23.28	16.33
CALIN (proposed)	68.55	17.12	64.32	3.65	1.60	3.48

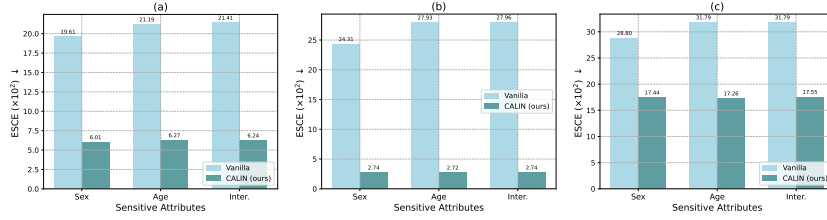


Fig. 2. Results on equity-scaling measure of calibration error (ESCE) on 3 datasets: (a) PAPILA [11], (b) HAM10000 [21], and (c) MIMIC-CXR [10]. The proposed method consistently outperforms baseline method, vanilla FS-ICL, in terms of the fairness-utility trade-off.

Table 3. Ablation study results on HAM10000. The bi-level approach outperforms baselines using single-level in most of the metrics.

Method	Acc. \uparrow	ECE \downarrow	$\overline{\text{EOR}} \uparrow$	CCEG $\Delta_\epsilon \downarrow$		
				Sex	Age	Inter.
\mathcal{L}_1 only	74.29	22.55	70.91	5.49	29.79	20.19
\mathcal{L}_2 only	64.88	14.13	72.66	0.83	22.73	16.43
Bi-level	74.76	2.68	74.24	4.43	3.14	3.11

4.2 CALIN Ablation Experiments

Ablation experiments are chosen to validate the effectiveness of CALIN’s bi-level framework, comparing its performance with baselines that use a single level (\mathcal{L}_1 or \mathcal{L}_2). Results in Table. 3 illustrate that the \mathcal{L}_2 baseline consistently outperforms \mathcal{L}_1 in both fairness metrics, highlighting the importance of modeling subgroup variability. The bi-level framework further improves by a large margin across most metrics, demonstrating the effectiveness of regularizing \mathcal{L}_2 with \mathcal{L}_1 .

5 Limitations

Model Limitation. This study focuses on identifying and addressing calibration biases in modern multimodal large language models (MLLMs), specifically using the GPT family model in our experiments. Although our findings reveal the presence of such biases in this model, a comprehensive analysis across alternative MLLM architectures and varying model sizes remains an open direction for future research.

Task Limitation. This study is restricted to medical image classification tasks where each label is represented by a single token. Such a formulation may be inadequate for tasks requiring multi-token labels. A possible workaround is to reformulate the task using single-token categorical options (e.g., *A*, *B*, *C*, *D*). Additionally, a thorough investigation into the impact of different exemplar combinations is left to future work.

6 Conclusion and Future Work

In this paper, we examine MLLM’s confidence calibration biases across demographic subgroups under FS-ICL, an area that remains unexplored in existing research. To address these biases, we introduce CALIN, a novel inference-time confidence calibration method. CALIN operates through a bi-level calibration procedure, effectively mitigating unfairness. Experimental results on three medical imaging datasets demonstrate that CALIN not only enhances fairness but also improves overall predictive performance and exhibits minimum fairness-utility trade-off. Future work should explore verbalized confidence [24], and integrate prompt optimization [13,23] to refine in-context exemplars for each demographic subgroup for improved fairness.

7 Ethics Statement

This work investigates fairness and calibration in modern multimodal large language models when used as clinical decision support tools. To promote equitable and reliable outcomes, we propose a training-free mitigation approach that reduces potential biases in model predictions. All experiments involving health-related data were conducted in compliance with relevant ethical guidelines and regulatory standards. Necessary permissions were obtained to access and process the datasets used in this study, and care was taken to ensure that all data handling adhered to principles of privacy, security, and responsible AI use.

Acknowledgments. This work was supported in part by the Natural Sciences and Engineering Research Council of Canada, in part by the Canadian Institute for Advanced Research (CIFAR) Artificial Intelligence Chairs Program, in part by the Mila – Quebec Artificial Intelligence Institute, in part by the Mila-Google Research Grant, and in part by the Canada First Research Excellence Fund, awarded to the Healthy Brains, Healthy Lives initiative at McGill University.

References

1. Baldassini, F.B., Shukor, M., Cord, M., Soulier, L., Piwowarski, B.: What makes multimodal in-context learning work? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1539–1550 (2024)
2. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. *Advances in Neural Information Processing Systems* **33**, 1877–1901 (2020)
3. Cao, B., Lin, H., Han, X., Sun, L., Yan, L., Liao, M., Xue, T., Xu, J.: Knowledgeable or educated guess? revisiting language models as knowledge bases. 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. 1860–1874 (2021)
4. Ferber, D., Wölflin, G., Wiest, I.C., Liger, M., Sainath, S., Ghaffari Laleh, N., El Nahhas, O.S., Müller-Franzes, G., Jäger, D., Truhn, D., et al.: In-context learning enables multimodal large language models to classify cancer pathology images. *Nature Communications* **15**(1), 10104 (2024)
5. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International Conference on Machine Learning. pp. 1321–1330. PMLR (2017)
6. Han, Z., Hao, Y., Dong, L., Sun, Y., Wei, F.: Prototypical calibration for few-shot learning of language models. In: The Eleventh International Conference on Learning Representations (2023)
7. He, K., Long, Y., Roy, K.: Prompt-based bias calibration for better zero/few-shot learning of language models. In: Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 12673–12691. Association for Computational Linguistics (2024)
8. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. *arXiv preprint arXiv:2410.21276* (2024)
9. Jin, R., Xu, Z., Zhong, Y., Yao, Q., Dou, Q., Zhou, S.K., Li, X.: FairmedFM: Fairness benchmarking for medical imaging foundation models. In: The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2024)
10. Johnson, A.E., Pollard, T.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Peng, Y., Lu, Z., Mark, R.G., Berkowitz, S.J., Horng, S.: Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042* (2019)
11. Kovalyk, O., Morales-Sánchez, J., Verdú-Monedero, R., Sellés-Navarro, I., Palazón-Cabanes, A., Sancho-Gómez, J.L.: Papila: Dataset with fundus images and clinical data of both eyes of the same patient for glaucoma assessment. *Scientific Data* **9**(1), 291 (2022)
12. Kull, M., Perello Nieto, M., Kängsepp, M., Silva Filho, T., Song, H., Flach, P.: Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. *Advances in Neural Information Processing Systems* **32** (2019)
13. Lu, Y., Bartolo, M., Moore, A., Riedel, S., Stenetorp, P.: Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 8086–8098 (2022)

14. Luo, Y., Tian, Y., Shi, M., Pasquale, L.R., Shen, L.Q., Zebardast, N., Elze, T., Wang, M.: Harvard glaucoma fairness: a retinal nerve disease dataset for fairness learning and fair identity normalization. *IEEE Transactions on Medical Imaging* (2024)
15. Ma, H., Zhang, C., Bian, Y., Liu, L., Zhang, Z., Zhao, P., Zhang, S., Fu, H., Hu, Q., Wu, B.: Fairness-guided few-shot prompting for large language models. *Advances in Neural Information Processing Systems* **36** (2024)
16. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. *ACM Computing Surveys* **54**(6), 1–35 (2021)
17. PhysioNet: Responsible use of mimic data with online services like GPT (2023), <https://physionet.org/news/post/gpt-responsible-use>
18. Shui, C., Szeto, J., Mehta, R., Arnold, D.L., Arbel, T.: Mitigating calibration bias without fixed attribute grouping for improved fairness in medical imaging analysis. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 189–198. Springer (2023)
19. Team, G., Georgiev, P., Lei, V.I., Burnell, R., Bai, L., Gulati, A., Tanzer, G., Vincent, D., Pan, Z., Wang, S., et al.: Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530* (2024)
20. Tian, Y., Shi, M., Luo, Y., Kouhana, A., Elze, T., Wang, M.: Fairseg: A large-scale medical image segmentation dataset for fairness learning using segment anything model with fair error-bound scaling. In: *The Twelfth International Conference on Learning Representations* (2024)
21. Tschandl, P., Rosendahl, C., Kittler, H.: The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data* **5**(1), 1–9 (2018)
22. Wang, X., Zhang, X., Cao, Y., Wang, W., Shen, C., Huang, T.: Seggpt: Towards segmenting everything in context. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 1130–1140 (2023)
23. Wu, Z., Lin, X., Dai, Z., Hu, W., Shu, Y., Ng, S.K., Jaillet, P., Low, B.K.H.: Prompt optimization with EASE? efficient ordering-aware automated selection of exemplars. In: *The Thirty-eighth Annual Conference on Neural Information Processing Systems* (2024)
24. Xiong, M., Hu, Z., Lu, X., LI, Y., Fu, J., He, J., Hooi, B.: Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In: *The Twelfth International Conference on Learning Representations* (2024)
25. Xu, G., CHEN, Q., Ling, C., Wang, B., Shui, C.: Intersectional unfairness discovery. In: *Forty-first International Conference on Machine Learning* (2024)
26. Xu, Z., Peng, K., Ding, L., Tao, D., Lu, X.: Take care of your prompt bias! investigating and mitigating prompt bias in factual knowledge extraction. In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. pp. 15552–15565 (2024)
27. Zhao, Z., Wallace, E., Feng, S., Klein, D., Singh, S.: Calibrate before use: Improving few-shot performance of language models. In: *International Conference on Machine Learning*. pp. 12697–12706. PMLR (2021)
28. Zong, Y., Yang, Y., Hospedales, T.: MEDFAIR: Benchmarking fairness for medical imaging. In: *The Eleventh International Conference on Learning Representations* (2023)