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 ($\overline{\textbf{EOR}}$) 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. ↑	ECE ↓	$\overline{\text{EOR}}$ ↑	CCEG Δ_ε ↓		
				Sex	Age	Inter.
PAPILA [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**
HAM10000 [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**
MIMIC-CXR [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**

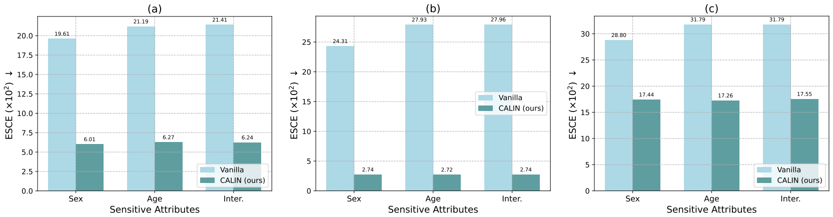

Fig. 2. Results on equity-scaling measure of calibration error (ESCE) on 3 datasets: (a) PAPILA [11], (b) HAM10000 [21], and (c) MIMIC-CXR [10]. The proposed method consistently outperforms baseline method, vanilla FS-ICL, in terms of the fairness-utility trade-off.

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

Method	Acc. ↑	ECE ↓	$\overline{\text{EOR}}$ ↑	CCEG Δ_ε ↓		
				Sex	Age	Inter.
$\mathscr{L}_1$ only	74.29	22.55	70.91	5.49	29.79	20.19
$\mathscr{L}_2$ 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 ($\mathscr{L}_1$ or $\mathscr{L}_2$). Results in Table. 3 illustrate that the $\mathscr{L}_2$ baseline consistently outperforms $\mathscr{L}_1$ 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 $\mathscr{L}_2$ with $\mathscr{L}_1$.

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.