nature computational science

Article

https://doi.org/10.1038/s43588-025-00789-7

Evaluating and mitigating bias in AI-based medical text generation

Received: 17 August 2024

Accepted: 12 March 2025

Published online: 23 April 2025

Check for updates

Xiuying Chen[1,2,4] ✉, Tairan Wang[2,4], Juexiao Zhou[2], Zirui Song[1], Xin Gao[2] ✉ & Xiangliang Zhang[2,3] ✉

Artificial intelligence (AI) systems, particularly those based on deep learning models, have increasingly achieved expert-level performance in medical applications. However, there is growing concern that such AI systems may reflect and amplify human bias, reducing the quality of their performance in historically underserved populations. The fairness issue has attracted considerable research interest in the medical imaging classification field, yet it remains understudied in the text-generation domain. In this study, we investigate the fairness problem in text generation within the medical field and observe substantial performance discrepancies across different races, sexes and age groups, including intersectional groups, various model scales and different evaluation metrics. To mitigate this fairness issue, we propose an algorithm that selectively optimizes those underserved groups to reduce bias. Our evaluations across multiple backbones, datasets and modalities demonstrate that our proposed algorithm enhances fairness in text generation without compromising overall performance.

Artificial intelligence (AI) systems, particularly those based on deep learning models, have been widely adopted in healthcare, consistently demonstrating expert-level performance across various domains, presenting a clear incentive for real-world deployment due to the global medical expert shortage and to AI algorithms matching specialist performance[1–6]. However, the issue of fairness has arisen in medical image classification tasks, with biases observed in deep learning models related to race[7–9], sex[10,11] and age[10]. The bias also exists in models trained from different types of medical datum, such as chest X-rays[8], CT scans[12] and skin dermatology images[13]. For instance, chest X-ray classifiers trained to predict the presence of disease systematically underdiagnose black patients[14], potentially leading to delays in care. A biased decision-making system is socially and ethically detrimental, especially in life-changing scenarios such as healthcare[15,16].

This has motivated a growing body of work to understand bias and pursue fairness in image classification tasks[17–20]. For example, ref. 10 proposed an algorithm that leverages the marginal pairwise equal opportunity to reduce bias in medical image classification. However, the fairness of text generation in medical contexts remains largely underexplored. This is particularly concerning given the rapid advancements in text generation using large language models (LLMs)[21–23]. Valuable applications of text generation in healthcare include generating detailed radiology-report descriptions[24] for more accurate diagnosis and distilling lengthy medical reports into concise summaries[25] for quicker decision-making. It also aids in creating personalized patient education materials[26] and automating clinical trial protocols[27], enhancing patient engagement and research efficiency. As these applications become widely adopted[23,28–32], it is crucial to consider the unfairness problem and bias. For example, as the real example in Supplementary Fig. 1 shows, if a summarization model misses an important cardiomegaly observation from the doctor's findings for a patient, this can lead to misdiagnosis or delayed treatment, potentially compromising patient care. This raises an important question: does unfairness exist in AI-generated medical text, and if so, how can it be mitigated? Investigating this problem presents a greater challenge than a typical classification task, as generation is harder to evaluate, and maintaining fairness in the generation process is more difficult than simply outputting classification labels.

[1]Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, United Arab Emirates. [2]King Abdullah University of Science and Technology, Thuwal, Saudi Arabia. [3]University of Notre Dame, Notre Dame, IN, USA. [4]These authors contributed equally: Xiuying Chen, Tairan Wang
✉e-mail: xiuying.chen@mbzuai.ac.ae; xin.gao@kaust.edu.sa; xzhang33@nd.edu

https://doi.org/10.1038/s43588-025-00789-7

Table 1 | Dataset characteristics

	Subgroup	Attribute	Total	Percentage
MIMIC dataset	Age	<65 yr	46,336	54.67%
		≥65 yr	55,902	45.32%
	Sex	Male	50,273	49.17%
		Female	51,965	50.82%
	Race	White	64,783	63.36%
		Black	19,568	19.13%
	Split	Train	102,238	97.94%
		Val.	800	0.76%
		Test	1,341	1.29%
Task			**Input**	**Output**
	Report generation		Images	Text
	Report summarization		Text	Text
PubMed dataset	Age	Adolescent	6,749	42.29%
		Young adult	2,077	13.01%
		Middle aged	6,749	42.29%
		Aged	4,887	30.62%
	Species	Humans	19,638	68.88%
		Animals	6,906	24.22%
	Split	Train	71,062	84.26%
		Val.	6,633	7.86%
		Test	6,635	7.86%
Task			**Input**	**Output**
	Paper summarization		Text	Text

The characteristics of MIMIC-CXR and PubMed datasets for the tasks of radiology-report generation, report summarization and paper summarization.

In this study, we first evaluate the presence of unfairness issues in image-based computer-aided diagnosis and text-based radiology-report and medical-paper summarization using publicly available datasets (Table 1). Our evaluation spans six generation evaluation metrics and three different scale models, and considers both individual and intersectional groups across dimensions such as race, sex and age. We also propose a metric-aware unfairness indicator to evaluate the unfairness from different aspects. Our experimental result shows that the unfairness problem exists against certain groups. We also find that intersectional subgroups exhibit compounded biases in text generation, with patients who belong to two underserved subgroups receiving lower-quality diagnoses and experiencing larger discrepancies. To address the issue of unfairness, we propose a selection optimization framework. Our first selection criterion relies on the intuitive cross-entropy loss function, where cases with higher loss in underrepresented groups are given more emphasis during the training process. Apart from general quality considerations, we also want to consider domain-specific fairness enhancements. There are metrics specifically designed for evaluating the accuracy of pathology concepts in medical text, and we prioritize training on cases that receive lower medical evaluation scores, thereby ensuring that the generated text precisely describes pathology terms. We demonstrate that our selective optimization can mitigate unfairness across various metrics, datasets and model scales for both individual and intersectional groups, without compromising the model's overall performance. Our approach is not task specific or model specific, and can be applied to various areas of text generation, potentially effectively reducing bias issues. An illustration of our model is presented in Fig. 1.

Results

Model overview

The pipeline of the proposed model is depicted in Fig. 1. The input images or text are passed into a neural network, which generates prediction results. The proposed method is versatile and not confined to specific deep learning models.

Generally speaking, vanilla generation models are trained by considering all cases in a batch, processing them sequentially using cross-entropy loss to predict words one by one. In this context, we denote the cases within a batch as B_i, where i represents the index number of the case.

In our approach, we introduce two paradigms for selecting cases for backpropagation, rather than using all cases. Our first selection criterion is simple and intuitive: we prioritize cases that exhibit a larger cross-entropy loss. Formally, this can be expressed as

$$B^*_{\text{selected}} = \text{Top}\,\gamma\{B_i \in B \mid \mathcal{L}_{\text{CE}}(B_i)\}. \quad (1)$$

Here, $\mathcal{L}_{\text{CE}}(B_i)$ denotes the cross-entropy loss of case B_i, and B^*_{selected} represents the subset of cases selected for backpropagation on the basis of their higher loss values.

The cross-entropy criterion emphasizes word-level accuracy. However, given our focus on medical applications, we also want to underscore the significance of accurately detecting pathology observations. To address this, as depicted in Fig. 2, we modify the model to provide a prediction score not only for the ground-truth reference but also for a set of reference candidates. These candidates are generated by base models such as R2Gen, which are predefined and sorted according to their ROUGE and CheXpert scores, covering a range of quality levels.

The model then learns a ranking function that assigns higher prediction scores to candidates of higher quality. To achieve this, we employ a ranking loss that penalizes the model when it fails to rank high-quality candidates above lower-quality ones. The ranking loss is defined as follows:

$$\mathcal{L}_{\text{Ranking}} = \sum_{i=1}^{n-1} \max\left(0, \Delta_i - \text{score}(C_i) + \text{score}(C_{i+1})\right), \quad (2)$$

where $\mathcal{L}_{\text{Ranking}}$ is the ranking loss, n is the number of candidates, Δ_i is the allowed margin between the scores of the ith candidate C_i and the $(i+1)$th candidate C_{i+1}, and score(C_i) is the model's prediction score for the ith candidate. The loss function encourages the model to learn that the score of the ith candidate should be at least Δ_i higher than the score of the $(i+1)$th candidate. If the model's predictions do not meet this criterion, the loss is non-zero and the model is penalized.

In essence, for cases where the candidates are not ranked correctly, it indicates that the input case is more challenging and the model does not fully comprehend the input. Therefore, by integrating the ranking loss with the generation loss, we select the cases for training as follows:

$$B^*_{\text{selected}} = \text{Top}\,\gamma\{B_i \in B \mid \mathcal{L}_{\text{CE}}(B_i) + \mathcal{L}_{\text{Ranking}}(B_i)\}. \quad (3)$$

This combined approach ensures that the model is trained to not only generate accurate words but also maintain pathology accuracy, thereby improving the overall quality of the generated text.

Evaluation metrics

The evaluation of generated text quality is a complex and arduous topic, presenting greater challenges than the predominantly studied image classification task[33–35]. To reasonably compare performance across different groups, it is crucial to first determine how to evaluate the quality of the generation. On the one hand, evaluation metrics based on n-gram overlap are the most commonly used and straightforward approach to assess text quality. Previous studies indicate that

https://doi.org/10.1038/s43588-025-00789-7

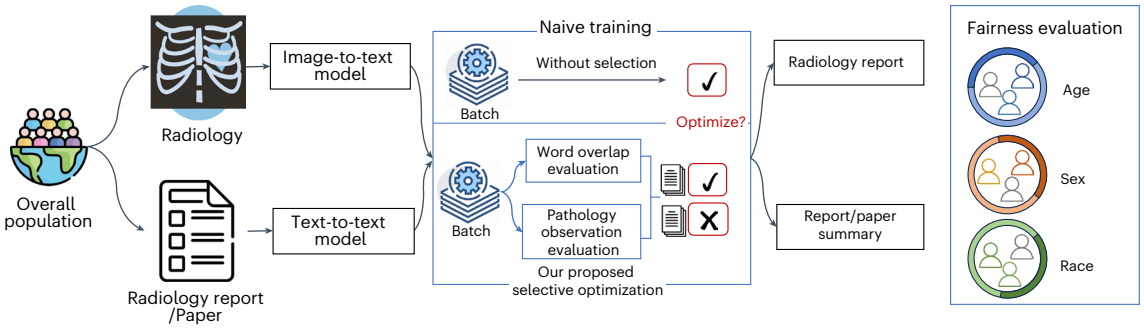

Fig. 1 | Model pipeline overview. Overview of the model pipeline consisting of radiology-report generation, report summarization and paper summarization.

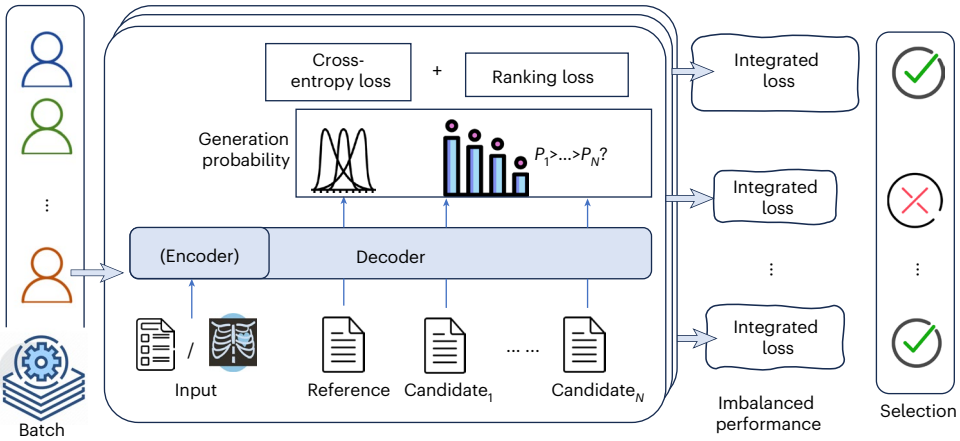

Fig. 2 | Details of the selection algorithm. Encoder–decoder or decoder-only model outputting generation probabilities for references and candidate sets. Ranking loss evaluating whether higher-quality candidates receive higher generation probabilities. Cases with high cross-entropy and ranking loss are selected for optimization.

metrics within this category, such as ROUGE[36], exhibit a high correlation with human evaluation results[37]. Other metrics such as BERTScore[33], ACU[38] and QuestEval[39] also demonstrate performance comparable to that of ROUGE[40]. Therefore, in this paper, we choose ROUGE scores as one of the primary evaluation metrics. Another paradigm of metrics is designed for radiology domains. Here, we use CheXpert scores[41], a common method in the radiology field[42–44], to evaluate generated text. Researchers define 14 chest pathologies as labels and assess the quality of the generated text by checking how well it detects and classifies these labels. The results are compared with ground-truth labels. The labels include No Finding, Enlarged Cardiomediastinum, Cardiomegaly, Lung Opacity, Lung Lesion, Edema, Consolidation, Pneumonia, Atelectasis, Pneumothorax, Pleural Effusion, Pleural Other, Fracture and Support Devices. Compared with ROUGE scores, CheXpert focuses more on the terms describing results of chest pathologies, rather than taking into account word overlap without differentiating the general and domain-specific terms.

Since previous work has not investigated the unfairness problem in text generation, in this study we also introduce a metric-aware fairness difference (MFD) to quantitatively measure unfairness. Inspired by the 'pairwise fairness difference' (PFD)[10] in the classification domain, which subtracts the score of the lowest-performing group from the highest score within the subgroups, our MFD adapts this concept for text generation. MFD is metric aware, calculating a differential score between subgroups for each specific metric by subtracting the score of the lowest-performing group from the highest score within the subgroups. Compared with PFD, MFD can assess the degree of unfairness from various perspectives in the generated text. The formal calculation of MFD can be found in 'Evaluation metrics'.

The justification for using MFD lies in its ability to capture unfairness across multiple dimensions relevant to text generation, which is inherently more complex than classification. In text generation, unfairness can manifest in subtle ways across various aspects of the generated output, such as stylistic inconsistencies or biases in content distribution. By providing a granular and metric-specific view of performance disparities, MFD facilitates a more comprehensive and targeted assessment of fairness in text-generation systems. A large MFD indicates disparities across subgroups, highlighting areas where fairness interventions are needed.

Compared models

Our study investigates three tasks: radiology-report generation, report summarization and paper summarization. For the first task, we selected the specialized model R2Gen[45], which is tailored for report generation. For the second task, we used the pretrained language model BART-large[46]. For the last task, we utilized both BART-base[46] and the LLM Llama-2-13B[47]. Our proposed paradigm is also applied to and compared with all the above models. In this way, the sizes of the models we investigated range from 100 million to 13 billion parameters, which provides a comprehensive investigation of fairness, and also allows us to thoroughly test our proposed method.

Datasets

We assessed the proposed and baseline models using the MIMIC-CXR[48] and PubMed[49] datasets. MIMIC-CXR[48] is a large public dataset of 377,110 chest X-rays associated with 227,827 free-text radiology reports and summaries for patients presenting to the Beth Israel Deaconess Medical Center Emergency Department between 2011 and 2016.

https://doi.org/10.1038/s43588-025-00789-7

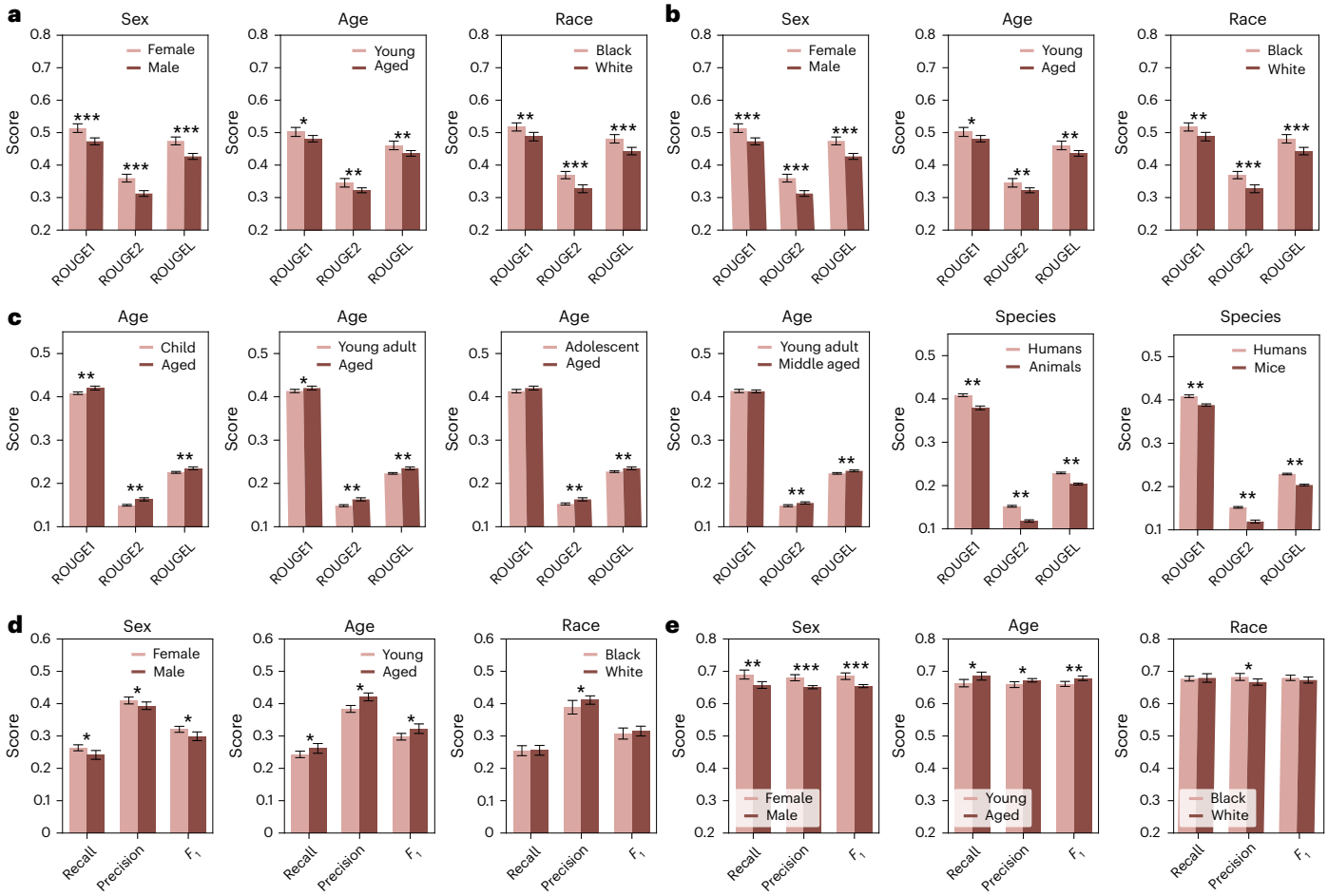

Fig. 3 | Performance disparities across demographics. a–e, Performance disparities across various categories, including sex (female versus male), race (black versus white), age (young versus aged) and species (humans versus animals), for ROUGE comparison in radiology-report generation (**a**), summarization on the CXR dataset (**b**) and summarization on the PubMed dataset (**c**), and for CheXpert comparison in radiology-report generation (**d**) and report summarization (**e**). The sample size is 6, and the data are presented as mean ± s.e.m. Error bars represent 95% confidence intervals. Significant differences, denoted by * ($P < 0.05$), ** ($P < 0.01$) and *** ($P < 0.001$), were identified using the two-sided Mann–Whitney *U* test.

With the release of its fourth version[50], the dataset now includes corresponding patient information. The race and sex data are self-reported, and age is documented at the time of a patient's first admission. We filtered out unpaired cases and accounted for instances where a single patient may have multiple X-rays and reports by randomly sampling one from each set. PubMed[49] is a summarization dataset consisting of 133,215 full-text papers as documents and their abstracts as summaries. We collected Medical Subject Headings (MeSH) labels using the PubMed API, resulting in 29,203 MeSH labels, and the evaluation is on the cases in test set with MeSH labels. Note that if a paper discusses both women and men, it is assigned to both categories.

Supplementary Fig. 2 provides an overview of the data selection process.

Text-generation bias in individual subpopulations on age, sex and race

As shown in Fig. 3, we find that the text-generation quality of baseline models for all datasets differs in most of the considered subpopulations.

First, in Fig. 3a–c, we show the performance of different subgroups under ROUGE metrics across report generation, report summarization and paper summarization, respectively. For the first two tasks, it is observed that female, young and black patients receive higher-quality summaries than their male, aged and white counterparts. This indicates that the generated text exhibits a higher word-level overlap with the ground-truth reference. These findings are consistent across the two tasks. For the third task, the analysis includes more granular age comparisons as well as comparisons between humans and animals. In the detailed age analysis, we observe that older individuals generally tend to receive better-performing summaries. Additionally, significant differences are observed when comparing human data with animal data.

Next, in Fig. 3d,e, we present the CheXpert scores of different groups on the first two tasks, which are based on the radiology domain. It is observed that, in 15 out of 18 settings, female, aged and white individuals achieve higher scores. This indicates that these groups receive higher-quality generated results for medical conditions. Detailed performances are given in Supplementary Tables 1–7.

While the CheXpert score calculates the overlap of clinical observations between the generated summary and the reference summary, ROUGE scores consider all words. Therefore, when we observe that females consistently achieve higher CheXpert scores and ROUGE scores than males, we also notice a distinct pattern: the generated text for young and black groups tends to score higher according to ROUGE metrics, whereas aged and white individuals score higher CheXpert scores. This discrepancy indicates that different metrics highlight different biases in model performance, making the alleviation of unfairness a challenging task that requires a multifaceted approach.

https://doi.org/10.1038/s43588-025-00789-7

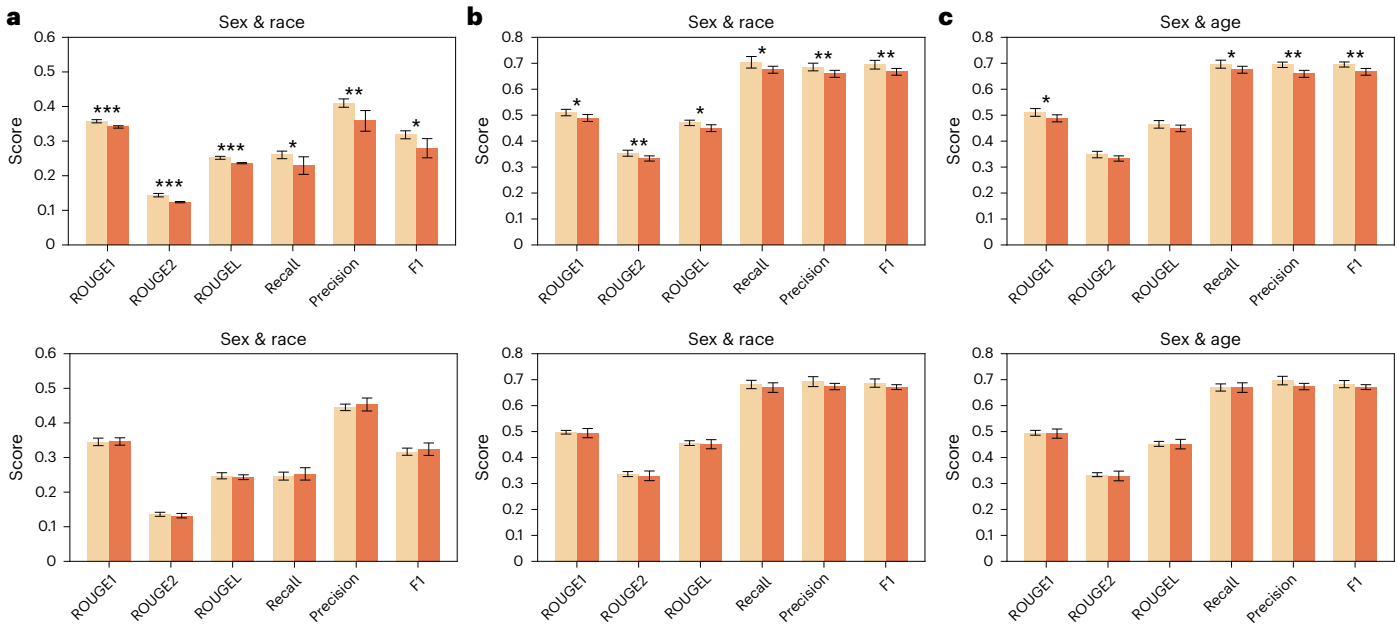

Fig. 4 | Performance disparities across intersectional groups. a–**c**, Performance disparities across intersectional groups for radiology-report generation (**a**) and report summarization for sex–race (**b**) and sex–age (**c**). Top row: baseline performance. Bottom row: enhanced performance using the proposed method. The sample size is 6, and the data are presented as mean ± s.e.m. Error bars represent 95% confidence intervals. Significant differences, denoted by * ($P < 0.05$), ** ($P < 0.01$) and *** ($P < 0.001$), were identified using the two-sided Mann–Whitney U test.

To offer a more intuitive understanding of the varying generation qualities, we present a representative example in Supplementary Fig. 1, with additional cases detailed in Supplementary Figs. 3 and 4. Despite similar inputs, males in these samples consistently received lower ROUGE and CheXpert scores, often resulting in missed or incorrect diagnoses. For instance, in the case depicted in Supplementary Fig. 1, both female and male cases involve cardiomegaly and edema. Yet, the predicted impression for females accurately includes these observations, whereas the male prediction fails to acknowledge the cardiopulmonary condition. It is crucial to emphasize that the observed biases could result in misdiagnoses or underdiagnoses for certain groups, potentially exacerbating existing health disparities. For example, if an AI system is biased against men or minorities, it might fail to accurately diagnose conditions that are more prevalent or present differently in these groups, such as cardiovascular diseases in men or skin conditions in people with darker skin tones. Addressing biases is crucial to ensure that AI systems are equitable and provide fair, accurate assessments for all patients, regardless of age, sex or race.

Text-generation bias in intersectional groups

We next investigate intersectional groups, defined as patients belonging to two subpopulations—for example, black female patients. We highlight the subpopulations with the largest fairness disparities in intersectional groups such as sex–race and sex–age in the first row of Fig. 4, with race–age comparisons shown in Supplementary Fig. 5. Out of the 16/18 metric comparisons, significant differences are revealed, indicating that intersectional subgroups frequently experience notable biases in text generation. To delve into the degree of bias, we compare the MFD between intersectional groups and subgroups in Supplementary Table 8. This comparison indicates that patients belonging to two underserved subgroups are more likely to receive lower-quality diagnoses and experience greater discrepancies between groups. For example, the disparity in CheXpert results between black males and white females is more pronounced than the disparities between black and white individuals or between females and males. Detailed scores from Fig. 4 are in Supplementary Table 9.

Why unfairness exists in radiology-report generation tasks

While the previous section focused on identifying biases across different groups, it is equally important to understand why such unfairness exists. By uncovering the underlying causes, we can better address these disparities and develop more equitable models.

First, we find that the ROUGE score is related to the target length. The Pearson correlation between the ROUGE score and the reference length is −0.21 with a P value of 3.90×10^{-16}, indicating a mild correlation where longer references tend to lead to lower ROUGE scores. This is intuitive because longer texts contain more information that needs to be generated, making the task more challenging.

Second, the CheXpert score is related to the original positive labels. If a group has more diseases classified as positive, its CheXpert score tends to be higher, with a correlation of 0.26 and a P value of 9.79×10^{-24}. This indicates that the model tends to generate text mentioning pathologies rather than plain text without any disease mentions. Meanwhile, we observe that different demographic groups have varying probabilities of developing certain diseases. For example, black patients have a 7% higher likelihood of being diagnosed with pneumonia compared with white patients. Related works, such as ref. 51, indicate that hospitals often provide lower-quality care to black patients for pneumonia. This highlights the need to further explore disparities in disease prevalence and healthcare quality among different groups, which we leave for future work.

Finally, the number of training cases is also crucial. The correlation between different group case numbers and the ROUGE performance is 0.26 with a P value of 0.02. This is also intuitive, as a larger number of training cases provides more data for the model to learn from, leading to better performance.

It is important to note that a group's final performance is determined by multiple factors combined. Sometimes, one factor may outweigh the others. For example, target text length is a decisive factor for ROUGE performance. However, in some cases, multiple factors work together to yield the final comparative outcomes. For instance, aged individuals have fewer training cases compared with young individuals (45% compared with 54%), more observation labels (5.13 compared

https://doi.org/10.1038/s43588-025-00789-7

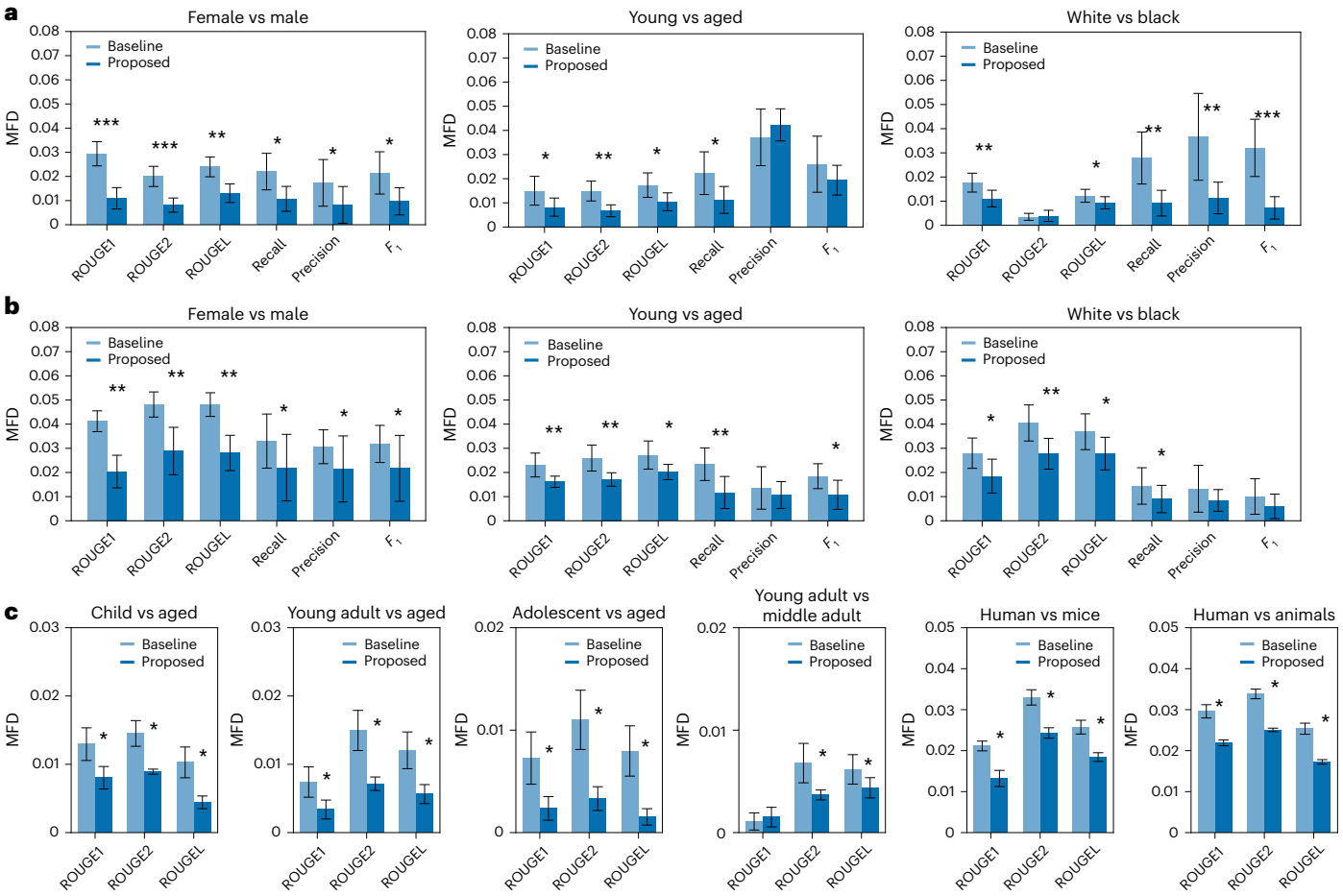

Fig. 5 | Reduction of MFD. a–**c**, Reduction of MFD across pairwise comparisons, including female versus male, for radiology-report generation (**a**), report summarization (**b**) and scholarly-paper summarization (**c**). The sample size is 6, and the data are presented as mean ± s.e.m. Error bars represent 95% confidence intervals. Significant differences, denoted by * ($P < 0.05$), ** ($P < 0.01$) and *** ($P < 0.001$), were identified using the two-sided Mann–Whitney U test.

with 5.01) and longer texts (62.89 words compared with 60.23 words). The longer text factor and fewer training samples lead to worse ROUGE performance, but having more labels results in a higher CheXpert score. Due to the complexity of these distributions, achieving balance with a simple adjustment of one factor is challenging. To validate this, we performed an oversampling study, detailed in Supplementary Table 11, where we increased the number of male cases to match that of female cases. However, the performance imbalance persisted, indicating that addressing such disparities requires more nuanced approaches.

In summary, these factors contribute to an unstable and imbalanced performance across different groups, making the alleviation of these disparities a challenging task.

Discussion

When equipping the baseline models with our proposed selective optimization, we find that our model is effective in reducing disparities across all datasets with respect to age, sex and race. In Fig. 5a,b, we present the MFD score for radiology-report generation and report summarization, respectively. Across almost all metrics and tasks, the MFD scores of our model are smaller than those of the baseline model, with the average MFD reduced by 35.27%. For the comparisons where the original discrepancy between groups is large, such as the female and male comparison, our model is particularly effective in alleviating bias. Moreover, although the discrepancies for different metrics vary—for example, black people have higher ROUGE scores but lower CheXpert scores compared with white people—our method is consistently useful for alleviating bias in these different metrics that measure different aspects. This suggests that, to improve fairness in other evaluation metrics, our framework can still be adopted. In Fig. 5c, we demonstrate that our method is also effective in alleviating bias not only between human groups but also across different species.

We also explore the effectiveness of our method for intersectional groups, as shown in Fig. 4. We highlight the subpopulations with the largest fairness disparities in intersectional groups such as sex–race and sex–age as shown in the three lower charts of the image, with race–age comparisons shown in Supplementary Fig. 5. As observed, there is no longer a significant difference between comparisons in most metrics. Our method even achieves superior performance, for example, on all three CheXpert metrics in the report-generation task. This suggests that our method, by maintaining balance, can indeed enhance the treatment received by intersectional groups, ensuring that they benefit from improved outcomes. Details of Fig. 4 are in Supplementary Table 9. We also provide a discussion of the original performance comparison in Supplementary Section 1.1, and performance on the LLM backbone in Supplementary Section 1.3.

While our proposed method demonstrates significant improvements in mitigating bias and enhancing fairness in medical text generation, several limitations remain. First, our approach relies on existing datasets, which may themselves contain inherent biases due to the demographic distributions or clinical practices present in the collected data. As a result, our method may not fully address biases that are not represented or underrepresented in the training data. Further research is needed to explore more diverse and inclusive datasets that better reflect the variability in real-world populations. Second, the

effectiveness of the proposed selective optimization depends on the quality of the fairness metrics used. While metrics such as ROUGE and CheXpert are effective in measuring text similarity and medical accuracy, they may not fully capture subtle forms of bias, such as stylistic or semantic disparities across subpopulations. Incorporating additional evaluation methods or human-in-the-loop assessments could further enhance bias mitigation. Additionally, while our approach generalizes across multiple models and scales, its performance in real-time or resource-constrained environments remains unexplored. The computational cost of selective optimization, especially when applied to LLMs, could pose challenges for deployment in clinical settings with limited resources. By addressing these limitations, our approach can be further refined to ensure more robust and equitable AI-driven solutions in medical text generation.

Methods

Evaluation metrics

We employ two types of evaluation metric. The first is a set of traditional text-generation metrics—ROUGE-1, ROUGE-2 and ROUGE-L, which respectively measure the matches of unigrams, bigrams and the longest common subsequence. These metrics directly reflect the similarity between the generated text and the ground-truth summary.

$$\text{ROUGE-1} = \frac{\sum_{S\in\text{References}}\sum_{\text{unigram}\in S}\text{Count}_{\text{match}}(\text{unigram})}{\sum_{S\in\text{References}}\sum_{\text{unigram}\in S}\text{Count}(\text{unigram})}, \quad (4)$$

$$\text{ROUGE-2} = \frac{\sum_{S\in\text{References}}\sum_{\text{bigram}\in S}\text{Count}_{\text{match}}(\text{bigram})}{\sum_{S\in\text{References}}\sum_{\text{bigram}\in S}\text{Count}(\text{bigram})}, \quad (5)$$

$$\text{ROUGE-L} = F_{\text{score}} = \frac{(1+\beta^2)R_{\text{LCS}}P_{\text{LCS}}}{R_{\text{LCS}}+\beta^2 P_{\text{LCS}}}, \quad (6)$$

where

$$R_{\text{LCS}} = \frac{\text{LCS}(X,Y)}{\text{length of reference}}, \quad (7)$$

$$P_{\text{LCS}} = \frac{\text{LCS}(X,Y)}{\text{length of candidate}}. \quad (8)$$

Additionally, we incorporate specially designed metrics, the CheXpert precision, recall and F_1 scores[41], which automatically detect the presence of 14 observations in radiology reports, capturing the uncertainties inherent in radiograph interpretation. This metric has been shown to outperform at least two of three radiologists in detecting four clinically relevant pathologies, demonstrating its capability in evaluating the accuracy of text.

From a fairness perspective, we introduce MFD, which computes the average absolute difference between two subgroups. Formally, MFD is defined as follows:

$$\text{MFD} = \frac{1}{n}\sum_{i=1}^{n}|\text{Metric}_{\text{subgroup1}}(i) - \text{Metric}_{\text{subgroup2}}(i)| \quad (9)$$

Here, n represents the number of instances, and $\text{Metric}_{\text{subgroup1}}(i)$ and $\text{Metric}_{\text{subgroup2}}(i)$ denote the metric values for the ith instance in subgroup 1 and subgroup 2, respectively. Being metric aware, MFD is adept at capturing the actual disparities among various groups on the basis of the inherent attributes of the metric itself, such as word-level accuracy or symptom detection accuracy. This allows for a nuanced understanding of fairness in the context of the specific task at hand.

The Mann–Whitney U test is a non-parametric statistical test used to determine whether there is a significant difference between two independent groups. A smaller P value indicates a significant difference, and better performance is determined by comparing the metric values of interest between the groups.

Experimental settings

We conducted our experiments using Hugging Face[52] on NVIDIA A100 graphics processing units. For the R2Gen[45] model, we employed a ResNet[53] pretrained on ImageNet[54] as the visual extractor to extract patch features, with each feature having a dimension of 2,048. For the relational memory, we set the dimension to 512, the number of heads in multihead attention to 8 and the number of memory slots to 3 by default. The model was trained using cross-entropy loss with the Adam optimizer[55]. We set the learning rate to 5×10^{-5} for the visual extractor and 1×10^{-4} for other parameters. We decayed this rate by a factor of 0.8 per epoch for each dataset and set the beam size to 3 to balance generation effectiveness and efficiency.

For the BART models (facebook/bart-base and facebook/bart-large)[46], we adhered to their hyperparameter settings as they yielded better performance. We used the Adam optimizer with ϵ set to 1×10^{-8} and β set to (0.9, 0.999). The learning rate was set to 3×10^{-5}, with a warm-up of 500 steps. The batch size was set to 8, with four gradient accumulation steps.

For fine-tuning the LLM Llama-2-13B, we used low-rank adaptation, which reduced the number of trainable parameters by learning pairs of rank-decomposition matrices while freezing the original weights. Specifically, we applied low-rank adaptation to the query projection layer and the value projection layer to enhance the model's adaptability without altering its structure. Additionally, we set the per-device training batch size to 16, utilized gradient accumulation with a step count of 1 to simulate larger batch sizes, and employed a cosine learning rate scheduler to optimize the learning rate adaptively throughout the training process.

We also include an ablation study in Supplementary Table 11, where we remove the cross-entropy-based selection and the ranking-loss-based selection, respectively. The results demonstrate the effectiveness of both components in alleviating bias. All experiments in this paper were repeated at least five times, following ref. 10. The average performance with 95% confidence intervals is reported for each evaluation.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The MIMIC-CXR dataset used in this study is available in the PhysioNet database[48] https://www.physionet.org/content/mimic-cxr-jpg/, which consists of de-identified chest X-ray images collected from the Beth Israel Deaconess Medical Center. The PubMed dataset is available at https://huggingface.co/datasets/ccdv/pubmed-summarization. It is a summarization and document pair dataset derived from PubMed, containing biomedical research abstracts and their corresponding summaries. All source datasets are public datasets that can be accessed on the basis of the links in this paper. Source data for Figs. 3–5 are available with this manuscript[56] under a Creative Commons license CC BY 4.0. Figures 1 and 2 do not contain associated data.

Code availability

The code supporting this study is publicly available[57] under a Creative Commons license CC BY 4.0. For development and version control, the source code is also hosted on GitHub: https://github.com/iriscxy/GenFair.

References

1. Jiang, F. et al. Artificial intelligence in healthcare: past, present and future. *Stroke Vasc. Neurol.* **2**, 230–243 (2017).

2. Rajpurkar, P. et al. CheXNet: radiologist-level pneumonia detection on chest x-rays with deep learning. Preprint at https://arxiv.org/abs/1711.05225 (2017).
3. Lin, M. et al. Automated diagnosing primary open-angle glaucoma from fundus image by simulating human's grading with deep learning. *Sci. Rep.* **12**, 14080 (2022).
4. Rimmer, A. Radiologist shortage leaves patient care at risk, warns royal college. *Br. Med. J.* **359**, j4683 (2017).
5. Chen, X. et al. Unveiling the power of language models in chemical research question answering. *Commun. Chem.* **8**, 4 (2025).
6. Wang, T. et al. Nature of metal–support interaction for metal catalysts on oxide supports. *Science* **386**, 915–920 (2024).
7. Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **366**, 447–453 (2019).
8. Seyyed-Kalantari, L., Liu, G., McDermott, M., Chen, I. Y. & Ghassemi, M. CheXclusion: fairness gaps in deep chest X-ray classifiers. In *Biocomputing 2021: Proc. Pacific Symposium* Vol. 26, 232–243 (World Scientific, 2020).
9. Zong, Y., Yang, Y. & Hospedales, T. MEDFAIR: benchmarking fairness for medical imaging. In *Eleventh International Conference on Learning Representations* (ICLR, 2022).
10. Lin, M. et al. Improving model fairness in image-based computer-aided diagnosis. *Nat. Commun.* **14**, 6261 (2023).
11. Larrazabal, A. J., Nieto, N., Peterson, V., Milone, D. H. & Ferrante, E. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proc. Natl Acad. Sci. USA* **117**, 12592–12594 (2020).
12. Zhou, Y. et al. RadFusion: benchmarking performance and fairness for multimodal pulmonary embolism detection from CT and EHR. Preprint at https://arxiv.org/abs/2111.11665 (2021).
13. Kinyanjui, N. M. et al. Fairness of classifiers across skin tones in dermatology. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* Vol. 12266, 320–329 (MICCAI, 2020).
14. Lin, M. et al. Evaluate underdiagnosis and overdiagnosis bias of deep learning model on primary open-angle glaucoma diagnosis in under-served populations. *AMIA Jt Summits Transl. Sci. Proc.* **2023**, 370–377 (2023).
15. Saarni, S. I. et al. Ethical analysis to improve decision-making on health technologies. *Bull. World Health Org.* **86**, 617–623 (2008).
16. Grote, T. & Berens, P. On the ethics of algorithmic decision-making in healthcare. *J. Med. Ethics* **46**, 205–211 (2020).
17. Zhang, H. et al. Improving the fairness of chest X-ray classifiers. *Proc. Mach. Learn. Res.* **174**, 204–233 (2022).
18. Lahoti, P. et al. Fairness without demographics through adversarially reweighted learning. *Adv. Neural Inf. Process. Syst.* **33**, 728–740 (2020).
19. Narasimhan, H., Cotter, A., Gupta, M. & Wang, S. Pairwise fairness for ranking and regression. In *Proceedings of the AAAI Conference on Artificial Intelligence* Vol. 34, 5248–5255 (AAAI Press, 2020).
20. Yang, Y., Zhang, H., Gichoya, J. W., Katabi, D. & Ghassemi, M. The limits of fair medical imaging AI in real-world generalization. *Nat. Med.* **30**, 2838–2848 (2024).
21. Chen, Q. et al. An extensive benchmark study on biomedical text generation and mining with ChatGPT. *Bioinformatics* **39**, btad557 (2023).
22. Li, J., Dada, A., Puladi, B., Kleesiek, J. & Egger, J. ChatGPT in healthcare: a taxonomy and systematic review. *Comput. Methods Programs Biomed.* **245**, 108013 (2024).
23. Tian, S. et al. Opportunities and challenges for ChatGPT and large language models in biomedicine and health. *Brief. Bioinform.* **25**, bbad493 (2024).
24. Tanida, T., Müller, P., Kaissis, G. & Rueckert, D. Interactive and explainable region-guided radiology report generation. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 7433–7442 (IEEE, 2023).
25. Van Veen, D. et al. RadAdapt: radiology report summarization via lightweight domain adaptation of large language models. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks* (eds Demner-Fushman, D. et al.) 449–460 (Association for Computational Linguistics, 2023).
26. Karabacak, M., Ozkara, B. B., Margetis, K., Wintermark, M. & Bisdas, S. The advent of generative language models in medical education. *JMIR Med. Educ.* **9**, e48163 (2023).
27. Subbiah, V. The next generation of evidence-based medicine. *Nat. Med.* **29**, 49–58 (2023).
28. Weidinger, L. et al. Ethical and social risks of harm from language models. Preprint at https://arxiv.org/abs/2112.04359 (2021).
29. Miner, A. S. et al. Smartphone-based conversational agents and responses to questions about mental health, interpersonal violence, and physical health. *JAMA Intern. Med.* **176**, 619–625 (2016).
30. Bickmore, T. W. et al. Patient and consumer safety risks when using conversational assistants for medical information: an observational study of Siri, Alexa, and Google Assistant. *J. Med. Internet Res.* **20**, e11510 (2018).
31. Sloan, P., Clatworthy, P., Simpson, E. & Mirmehdi, M. Automated radiology report generation: a review of recent advances. *IEEE Rev. Biomed. Eng.* 4225–4232 (2024).
32. Pang, T., Li, P. & Zhao, L. A survey on automatic generation of medical imaging reports based on deep learning. *Biomed. Eng. Online* **22**, 48 (2023).
33. Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q. & Artzi, Y. BERTScore: evaluating text generation with BERT. In *International Conference on Learning Representations* (ICLR, 2020).
34. Celikyilmaz, A., Clark, E. & Gao, J. Evaluation of text generation: a survey. Preprint at https://arxiv.org/abs/2006.14799 (2020).
35. Fu, J., Ng, S.-K., Jiang, Z. & Liu, P. GPTScore: evaluate as you desire. In *Proc. 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* Vol. 1, 6556–6576 (2024).
36. Lin, C.-Y. ROUGE: a package for automatic evaluation of summaries. In *Text Summarization Branches Out* 74–81 (Association for Computational Linguistics, 2004).
37. Chen, X. et al. Flexible and adaptable summarization via expertise separation. In *Proc. 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* 2018–2027 (Association for Computing Machinery, 2024).
38. Liu, Y. et al. Revisiting the gold standard: grounding summarization evaluation with robust human evaluation. *Proc. 61st Annual Meeting of the Association for Computational Linguistics* Vol. 1 (eds Rogers, A. et al.) 4140–4170 (Association for Computational Linguistics, 2023).
39. Scialom, T. et al. QuestEval: summarization asks for fact-based evaluation. In *Proc. 2021 Conference on Empirical Methods in Natural Language Processing* (eds Moens, M.-F. et al.) 6594–6604 (Association for Computational Linguistics, 2021).
40. Chen, X. et al. Rethinking scientific summarization evaluation: grounding explainable metrics on facet-aware benchmark. Preprint at https://arxiv.org/abs/2402.14359 (2024).
41. Irvin, J. et al. CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison. *Proc. Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence* 590–597 (AAAI Press, 2019).

42. Wang, X. et al. CXPMRG-Bench: pre-training and benchmarking for X-ray medical report generation on CheXpert Plus dataset. Preprint at https://arxiv.org/abs/2410.00379 (2024).
43. Endo, M., Krishnan, R., Krishna, V., Ng, A. Y. & Rajpurkar, P. Retrieval-based chest x-ray report generation using a pre-trained contrastive language-image model. *Proc. Mach. Learn. Res.* **158**, 209–219 (2021).
44. Boag, W. et al. Baselines for chest x-ray report generation. *Proc. Mach. Learn. Res.* **116**, 126–140 (2020).
45. Chen, Z., Song, Y., Chang, T.-H. & Wan, X. Generating radiology reports via memory-driven transformer. In *Proc. 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (eds Jurafsky, D. et al.) 1439–1449 (2020).
46. Lewis, M. et al. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proc. 58th Annual Meeting of the Association for Computational Linguistics* 7871–7880 (Association for Computational Linguistics, 2020).
47. Touvron, H. et al. Llama 2: open foundation and fine-tuned chat models. Preprint at https://arxiv.org/abs/2307.09288 (2023).
48. Johnson, A. E. et al. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. Preprint at https://arxiv.org/abs/1901.07042 (2019).
49. Cohan, A. et al. A discourse-aware attention model for abstractive summarization of long documents. In *Proc. 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* Vol. 2 (eds Walker, M. et al.) 615–621 (2018).
50. Johnson, A. et al. MIMIC-IV. *PhysioNet* https://physionet.org/content/mimiciv/1.0/ 49–55 (2020).
51. Mayr, F. B. et al. Do hospitals provide lower quality of care to black patients for pneumonia? *Crit. Care Med.* **38**, 759–765 (2010).
52. Wolf, T. et al. Transformers: State-of-the-art natural language processing. In *Proc. 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (eds Liu, Q. & Schlangen, D.) 38–45 (Association for Computational Linguistics, 2020).
53. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 770–778 (IEEE, 2016).
54. Deng, J. et al. ImageNet: a large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* 248–255 (IEEE, 2009).
55. Kingma, D. P. and Ba, J. Adam: a method for stochastic optimization. Preprint at https://arxiv.org/abs/1412.6980 (2014).
56. Chen, X. support data.zip. *Figshare* https://doi.org/10.6084/m9.figshare.28516889.v1 (2025).
57. Chen, X. Code for GenFair. *Figshare* https://doi.org/10.6084/m9.figshare.28516898.v1 (2025).

Acknowledgement

X.C. was supported by Mohamed bin Zayed University of Artificial Intelligence (MBZUAI) through grant award 8481000078.

Author contributions

X.C. and T.W. contributed to the development of the idea, experiments and manuscript writing. J.Z. and Z.S. were responsible for conducting experiments. X.G. and X.Z. provided supervision and contributed to the manuscript writing.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43588-025-00789-7.

Correspondence and requests for materials should be addressed to Xiuying Chen, Xin Gao or Xiangliang Zhang.

Peer review information *Nature Computational Science* thanks Jiangning Song and Wenbin Zhang for their contribution to the peer review of this work. Primary Handling Editor: Ananya Rastogi, in collaboration with the *Nature Computational Science* team. Peer reviewer reports are available.

nature portfolio

Corresponding author(s): Xiuying Chen

Last updated by author(s): Mar 11, 2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	We used benchmark datasets, the details are provided in the Data section.
Data analysis	No software was used for data collection.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The MIMIC-CXR dataset used in this study is available in the PhysioNet database (https://www.physionet.org/content/mimic-cxr-jpg), which consists of de-identified chest X-ray images collected from the Beth Israel Deaconess Medical Center. The PubMed dataset is available at (https://huggingface.co/datasets/ccdv/pubmed-summarization). It is a summarization and document-pair dataset derived from PubMed, containing biomedical research abstracts and their corresponding summaries.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	N/A
Reporting on race, ethnicity, or other socially relevant groupings	The race distribution is reported in the statistic table in the paper.
Population characteristics	The population characteristics are reported in the statistic table in the paper.
Recruitment	n/a
Ethics oversight	n/a

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	100,000
Data exclusions	We didn't modify the data distribution.
Replication	We release all the code for reproducing the experiment result.
Randomization	We use 6 different seeds for experiment.
Blinding	No human evaluation is conducted. All evaluated by automatic metrics.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	*Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*
Novel plant genotypes	*Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.*
Authentication	*Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*