GPT-DBR: Decoding-Based Language-Model Regression for CT-Derived Lung-Function Estimation

YoungSeok Kim[1,2], Kwangsuk Park[3,4], WoongJae Jo[3,4], Yerin Hong[3], and Yun-Hyeon Kim[1,5⋆]

[1] Department of Radiology, Chonnam National University Hospital, Gwangju, Republic of Korea
[2] Department of Convergence Medicine, University of Ulsan College of Medicine, Asan Medical Center, Seoul, Republic of Korea
[3] AA LAB, MODULABS, Republic of Korea
[4] Aiffel, MODULABS, Republic of Korea
[5] Department of Radiology, Chonnam National University Medical School, Gwangju, Republic of Korea
yhkim001@jnu.ac.kr

Abstract. Spirometry is the clinical standard for pulmonary-function assessment but depends on patient effort, operator expertise, and test availability. We present GPT-DBR, a small-cohort framework that applies decoding-based language-model regression to quantitative CT (qCT) and radiomic biomarkers. In 164 paired CT–spirometry cases, GPT-DBR(qCT) achieved the lowest FEV_1/FVC (%) error (MAE 7.81) and competitive FEV_1 (% predicted) error (MAE 15.28). In the held-out split, qCT label-recovery AUC increased from 0.469 to 0.768 across adaptation stages. These proof-of-concept findings support the feasibility of language-model regression over structured CT biomarkers in data-scarce settings, while external validation remains necessary before clinical translation.

Keywords: Decoding-based regression · CT biomarkers · COPD · language models · qCT

1 Introduction

Chronic obstructive pulmonary disease (COPD) is a major global health burden [1, 2]. Spirometry remains central to diagnosis and follow-up but depends on patient cooperation and test availability. Chest CT provides complementary structural information on emphysema, airway narrowing, and lung-volume change. Accordingly, qCT, radiomics, and deep-learning approaches have been used to estimate COPD-related pulmonary function from CT [3–6]. Whether language-model regression can exploit such structured CT biomarkers in small paired CT–PFT cohorts remains unclear.

⋆ Corresponding author

Language models are primarily used for text, classification, and report generation tasks in medicine [7–9], while tabular applications have also focused largely on classification [10]. Continuous regression is less natural for autoregressive decoders because they generate tokens rather than scalar values. Decoding-based regression (DBR) addresses this mismatch by representing numeric targets as generated token sequences [11].

We therefore evaluate GPT-DBR using standardized qCT and radiomic features to estimate FEV_1 (% predicted) and FEV_1/FVC (%). The contributions are threefold: (1) formulation of CT-derived spirometric estimation as structured continuous-value generation; (2) evaluation of a two-stage DBR adaptation of a version-pinned GPT-family decoder without a task-specific regression head; and (3) comparison with AutoML, an untuned language-model baseline, radiomics variants, and an exploratory 3-D CNN comparator.

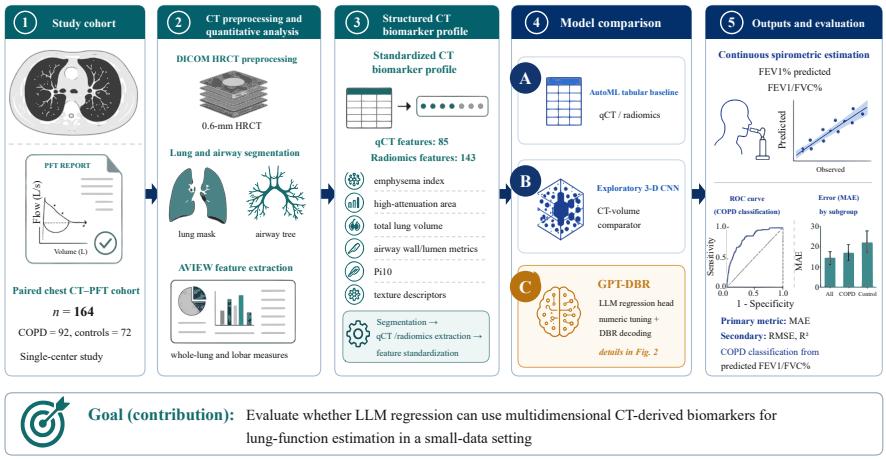


Fig. 1. Study workflow and contribution. Paired CT–PFT data were processed through DICOM HRCT preprocessing, lung/airway segmentation, AVIEW-based qCT and radiomics extraction, and feature standardization. The resulting CT biomarker profile was used for AutoML baselines, an exploratory 3-D CNN comparator, and GPT-DBR to estimate spirometric indices in a small-data setting.

2 Materials and Methods

2.1 Cohort and Data Use

We used an anonymized single-center cohort of 164 subjects collected between January 2011 and April 2014: 92 COPD patients and 72 non-COPD controls. Participants underwent pulmonary-function testing and high-resolution chest CT. COPD was defined in the source dataset using $FEV_1/FVC < 0.70$; because

GOLD recommends post-bronchodilator $FEV_1/FVC < 0.70$ for diagnostic confirmation [2], unavailable bronchodilator status is treated as a limitation.

The study was approved by the Institutional Review Board of Chonnam National University Hospital (IRB no. CNUH-2012-34), and the need for individual informed consent was waived for this retrospective analysis of de-identified data. Direct identifiers, DICOM metadata, accession numbers, free-text reports, and non-approved date fields were removed before analysis. Raw CT and pulmonary-function data are not publicly released.

Table 1. Baseline characteristics of the study cohort; P-values were calculated using Welch's t-test.

Variable	COPD (n=92)	Non-COPD (n=72)	P-value
Age (yr)	68.7 ± 8.3	62.8 ± 11.4	< 0.001
BMI (kg m^{-2})	22.8 ± 3.05	23.6 ± 2.74	0.077
Height (cm)	164.4 ± 8.12	161.6 ± 15.61	0.176
Body weight (kg)	61.9 ± 11.25	62.8 ± 10.90	0.614
FEV_1/FVC (%)	55.2 ± 12.69	78.3 ± 5.39	< 0.001
FEV_1 (% predicted)	80.9 ± 24.00	101.9 ± 23.70	< 0.001

2.2 CT-Derived Feature Extraction

Using AVIEW software (Coreline Soft, Seoul, Republic of Korea; v1.0.2.5.9), imaging features were extracted locally from chest CT. We used 85 quantitative CT (qCT) features, including emphysema index, high-attenuation area, lung volume, and airway measurements such as Pi10, together with 143 radiomic features. Only the resulting structured numerical features were used as inputs to the language-model regression framework; raw CT images were not provided to the language model.

2.3 Data Split and Preprocessing

The 164 subjects were divided using a fixed subject-level stratified split into 98 training, 16 validation, and 50 held-out evaluation subjects. The corresponding COPD/non-COPD counts were 55/43, 9/7, and 28/22, respectively. The same 50-patient evaluation set was used for all six GPT candidates. All data-dependent preprocessing, including imputation, standardization, scaling, and feature selection, was fit using the training split only and applied without refitting to validation and evaluation subsets.

2.4 GPT-DBR and Baselines

Figure 1 summarizes the study workflow. CT images were converted locally to qCT and radiomic biomarkers and assembled into standardized CT biomarker profiles. Language-model experiments used the version-pinned OpenAI model

`gpt-4o-2024-08-06` and its fine-tuned descendants [12]. The methodological contribution is the DBR protocol and structured-biomarker interface rather than the recency or unique optimality of this checkpoint.

We compared GPT-DBR with AutoML tabular ensembles implemented with AutoGluon [13], an untuned base language model, radiomics-based variants, and an exploratory Inflated 3-D ConvNet-style baseline. The AutoML models used structured tabular features, whereas the 3-D CNN operated directly on CT volumes. Because the cohort is small for end-to-end volumetric training, the 3-D CNN is interpreted only as an exploratory comparator.

2.5 GPT-DBR Training and Inference

GPT-Num first learned direct numeric decoding of the spirometric targets, and GPT-DBR was subsequently initialized from the corresponding GPT-Num model for decoding-based regression.

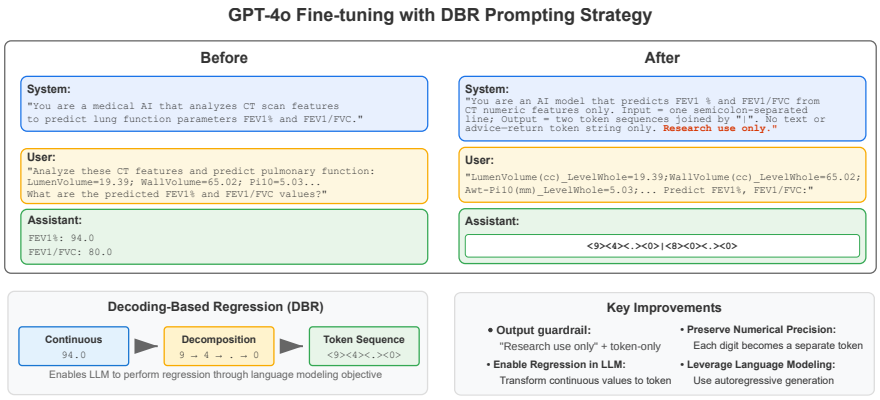


Fig. 2. Two-stage GPT-DBR fine-tuning. Direct numeric decoding (GPT-Num) is followed by numeric-token decoding (GPT-DBR), aligning regression with the autoregressive objective.

Figure 2 details the fine-tuning strategy. Stage 1 trained the decoder to output continuous spirometric values as direct numeric responses. Stage 2 used a fixed numeric-token response schema, aligning continuous regression with the autoregressive decoding objective without a task-specific regression head [11].

Fine-tuning used two epochs, batch size 1, a learning-rate multiplier of 2.0, and seed 42. A validation dataset was provided during fine-tuning and validation performance was monitored. The final fine-tuned model produced at the end of training was used for evaluation. No systematic grid search or automated hyperparameter optimization was performed.

During held-out evaluation, ten stochastic outputs were generated per patient and averaged to obtain one patient-level prediction before metric calculation.

2.6 Evaluation and Statistical Analysis

MAE was the primary regression metric, with RMSE and signed R^2 as secondary metrics. To account for repeated stochastic outputs, ten GPT predictions were first averaged within each patient, and metrics were calculated across the same 50 held-out patients. MAE uncertainty was estimated using 10,000 patient-level bootstrap resamples, with paired resampling for GPT-Num(qCT) versus GPT-DBR(qCT). Generation-level SD was interpreted only as within-patient output variability.

The FEV_1/FVC threshold analysis was secondary and exploratory because both the reference label and prediction were derived from FEV_1/FVC. ROC-AUC confidence intervals were estimated using 10,000 class-stratified patient-bootstrap resamples. During metric auditing, signed metrics such as R^2 were corrected where an absolute-value transform had been applied in the AutoML summary.

2.7 Data Transfer and Reproducibility

All CT processing and feature extraction were performed locally. Only de-identified numerical features and training/validation targets were transmitted for fine-tuning; raw CT images, direct patient identifiers, and held-out evaluation targets were not transmitted. Evaluation inference used `store=false`. We used the version-pinned `gpt-4o-2024-08-06` model and report metric-level rather than exact output reproducibility.

3 Results

Table 2 summarizes held-out performance after restoring signed R^2 values. For FEV1% predicted, GPT-DBR(qCT) had the lowest point-estimate MAE among the language-model variants (15.28), while GPT-Num(qCT) showed similar performance (MAE 15.45, RMSE 19.76, $R^2 = 0.291$). GPT-DBR(qCT) achieved MAE 15.28, RMSE 19.72, and $R^2 = 0.294$. ML(qCT) had a similar MAE (15.84) but a larger RMSE (26.82) and negative signed R^2 (-0.866), consistent with large squared-error sensitivity in the held-out split.

For FEV1/FVC%, GPT-DBR(qCT) yielded the lowest point-estimate error in the held-out split, with MAE 7.81 (95% patient-bootstrap CI: 5.46–10.69), RMSE 12.34, and $R^2 = 0.464$. GPT-Num(qCT) achieved MAE 10.07 (95% CI: 7.45–13.36), RMSE 14.76, and $R^2 = 0.234$. The paired improvement in MAE for GPT-DBR(qCT) over GPT-Num(qCT) was 2.26 percentage points (95% CI: 1.13–3.44). The corresponding values were higher for ML(qCT) (MAE 9.44, RMSE 11.90, $R^2 = 0.221$), ML(RM) (MAE 11.49, RMSE 14.11, $R^2 = -0.095$), and the exploratory 3-D CNN baseline (MAE 14.14, RMSE

Table 2. Held-out regression performance in the 50-patient evaluation set.

Model	Input	FEV1% predicted			FEV1/FVC%		
		MAE	RMSE	R^2	MAE	RMSE	R^2
ML(RM)	Radiomic feature	16.29	21.19	-0.165	11.49	14.11	-0.095
ML(qCT)	qCT feature	15.84	26.82	-0.866	9.44	11.90	0.221
3D-CNN	CT volume	20.45	25.60	-0.167	14.14	17.50	-0.248
GPT-base(RM)	Radiomic feature	26.79	32.38	-0.904	12.84	17.83	-0.118
GPT-base(qCT)	qCT feature	23.84	29.24	-0.552	13.17	17.52	-0.079
GPT-Num(RM)	Radiomic feature	17.80	23.87	-0.035	12.96	18.38	-0.188
GPT-Num(qCT)	qCT feature	15.45	19.76	0.291	10.07	14.76	0.234
GPT-DBR(RM)	Radiomic feature	19.47	25.84	-0.213	14.01	19.85	-0.385
GPT-DBR(qCT)	qCT feature	15.28	19.72	0.294	7.81	12.34	0.464

ML: AutoML tabular ensemble implemented with AutoGluon; RM: radiomic feature; qCT: quantitative CT; 3D-CNN: exploratory volumetric CT comparator; GPT-base: untuned base language-model decoder; GPT-Num: first-stage numeric tuning; GPT-DBR: second-stage decoding-based regression. For GPT variants, 10 stochastic outputs were averaged within each patient before calculation of patient-level metrics. Signed R^2 values are reported without absolute-value transformation.

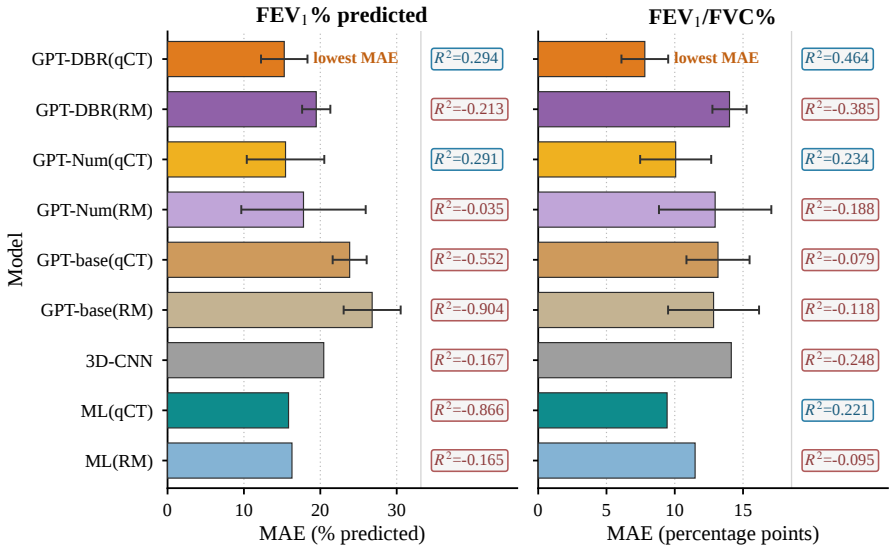
17.50, $R^2 = -0.248$). The untuned base language-model decoder showed substantially higher FEV1/FVC% error than GPT-DBR(qCT) (MAE 13.17, RMSE 17.52, $R^2 = -0.079$).

Figure 3 highlights the main regression comparison with signed R^2 annotations. GPT-DBR(qCT) was most distinctive for FEV1/FVC%, whereas FEV1% predicted showed a narrower spread between GPT-DBR(qCT), GPT-Num(qCT), and ML(qCT). This target-dependent behavior argues for cautious interpretation: DBR is not uniformly superior for every target, but the clearest reduction in error was observed for FEV1/FVC% with qCT inputs.

Figure 4 summarizes the secondary classification analysis for FEV1/FVC < 0.70 using predicted FEV1/FVC values. AUC was calculated using the negative patient-level mean predicted FEV1/FVC% as the continuous score, with measured FEV1/FVC < 0.70 defined as the positive class. Because both the reference class and the prediction were derived from FEV1/FVC, this secondary analysis should not be interpreted as an independent COPD diagnostic endpoint. With qCT inputs, AUC increased from 0.469 for GPT-base(qCT) to 0.666 for GPT-Num(qCT) and 0.768 for GPT-DBR(qCT). The paired GPT-DBR minus GPT-Num AUC difference was 0.102 (95% CI: -0.011–0.222). Radiomic-feature inputs remained close to chance across the same progression.

4 Discussion and Conclusion

This study frames language-model use in medical imaging as regression over structured CT biomarkers. GPT-DBR operates on qCT and radiomic biomarkers extracted from segmented CT scans rather than directly on raw CT volumes. This design keeps the methodological focus on structured imaging biomarkers



Error bars on GPT-based models denote SD across 10 stochastic generations. Signed R^2 is preserved after metric-audit correction.

Fig. 3. Target-wise MAE comparison using a unified model naming scheme. Each bar reports held-out MAE; right-side badges report signed target-specific R^2 values. For GPT variants, error bars denote 95% patient-bootstrap confidence intervals after averaging 10 stochastic outputs within each patient.

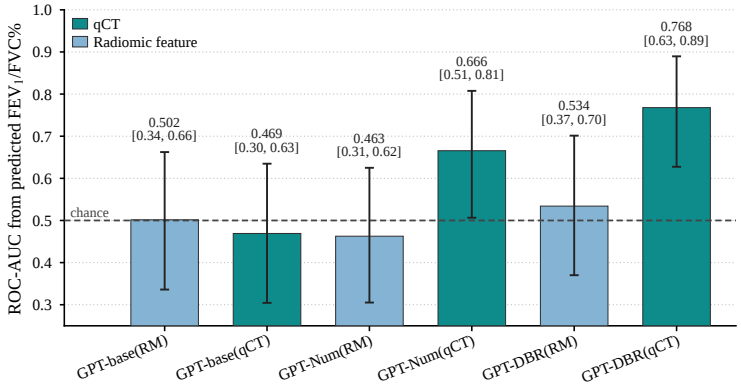


Fig. 4. Secondary classification performance for FEV1/FVC < 0.70 using predicted FEV1/FVC values. Error bars denote 95% class-stratified patient-bootstrap CIs.

and provides a proof-of-concept setting for evaluating language-model regression in a relatively small paired CT–spirometry cohort.

The results support the feasibility of language-model-based numeric regression for structured medical-imaging biomarkers. Language models have primarily been applied to text, reports, or multimodal description tasks, whereas here the target is a continuous physiologic value. The numeric-plus-token DBR strategy translates continuous spirometric outcomes into a sequence-generation format that can be optimized by an autoregressive decoder. The clearest effect was observed for FEV_1/FVC (%) with qCT inputs: GPT-DBR(qCT) achieved an MAE of 7.81, compared with 10.07 for GPT-Num(qCT), corresponding to a paired MAE reduction of 2.26 percentage points (95% CI: 1.13–3.44). In contrast, FEV_1 (% predicted) showed only a small difference between GPT-DBR(qCT) and GPT-Num(qCT), with MAEs of 15.28 and 15.45, respectively. These findings indicate that the effect of DBR was target dependent rather than uniformly superior across regression tasks.

The difference between FEV_1 (% predicted) and FEV_1/FVC (%) performance is clinically and statistically plausible. FEV_1/FVC is directly tied to airflow obstruction, whereas FEV_1 (% predicted) reflects a broader severity measure normalized by reference equations and influenced by demographic and non-imaging factors. In this cohort, FEV_1/FVC (%) also showed stronger COPD-control separation, which may partly explain the stronger qCT-based regression signal.

The comparison between qCT and radiomics suggests, but does not prove, a feature-quality versus feature-quantity trade-off. qCT features are physiologically interpretable and directly linked to COPD phenotypes, whereas high-dimensional radiomics may introduce redundancy in a small cohort. Consistent with this interpretation, the regression improvement observed with qCT was not reproduced with radiomic inputs: GPT-DBR(RM) had a FEV_1/FVC (%) MAE of 14.01 compared with 12.96 for GPT-Num(RM). Radiomics should therefore be reassessed in larger or externally validated designs.

In the held-out split, the AUC for classifying subjects according to $FEV_1/FVC < 0.70$ increased from 0.469 for GPT-base(qCT) to 0.666 for GPT-Num(qCT) and 0.768 for GPT-DBR(qCT). The paired GPT-DBR minus GPT-Num AUC difference was 0.102 (95% CI: -0.011–0.222), with the interval including zero. Because both the reference class and the prediction were derived from FEV_1/FVC, this secondary classification analysis should not be interpreted as independent COPD diagnostic performance. Radiomic-feature inputs remained close to chance, and spirometry remains the reference test.

Several limitations remain. First, this is a small, single-center proof-of-concept study without external validation. Validation across scanner vendors, reconstruction protocols, and patient populations is required before clinical translation [14]. Second, COPD labeling depends on spirometry, and post-bronchodilator status should be verified when available [2]. Third, the 3-D CNN baseline is exploratory and likely underpowered for end-to-end volumetric learning. Fourth, patient-level bootstrap intervals reflect sampling uncertainty within the fixed held-out cohort and do not capture model-retraining or external-generalization uncertainty. Finally, the GPT-based models depend on a closed, version-pinned API; repro-

ducibility is therefore reported at the metric level rather than as exact output reproduction.

GPT-DBR provides an experimental framework for estimating spirometric indices from CT-derived structured biomarkers using decoding-based regression. In this cohort, GPT-DBR(qCT) achieved competitive FEV_1 (% predicted) prediction and the lowest FEV_1/FVC (%) point-estimate error among the reported variants, whereas the same regression benefit was not observed with radiomic inputs. These proof-of-concept findings support further investigation while prospective multicenter external validation remains necessary before clinical translation.

AI-use disclosure. Generative AI was used for language editing; all scientific content and analyses were verified by the authors.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Boers, E., Barrett, M., Su, J.G., et al.: Global burden of chronic obstructive pulmonary disease through 2050. *JAMA Network Open* 6(12), e2346598 (2023). https://doi.org/10.1001/jamanetworkopen.2023.46598
2. Global Initiative for Chronic Obstructive Lung Disease: Global strategy for prevention, diagnosis and management of COPD: 2024 report. https://goldcopd.org/2024-gold-report/ (2024)
3. Zhou, T.-H., Zhou, X.-X., Ni, J., et al.: CT whole-lung radiomic nomogram: a potential biomarker for lung function evaluation and identification of COPD. *Military Medical Research* 11, 14 (2024). https://doi.org/10.1186/s40779-024-00516-9
4. Amudala Puchakayala, P.R., Sthanam, V.L., Nakhmani, A., et al.: Radiomics for improved detection of chronic obstructive pulmonary disease in low-dose and standard-dose chest CT scans. *Radiology* 307(5), e222998 (2023). https://doi.org/10.1148/radiol.222998
5. Park, H., Yun, J., Lee, S.M., et al.: Deep learning-based approach to predict pulmonary function at chest CT. *Radiology* 307(2), e221488 (2023). https://doi.org/10.1148/radiol.221488
6. Chen, B., Liu, Z., Lu, J., et al.: Deep learning parametric response mapping from inspiratory chest CT scans: a new approach for small airway disease screening. *Respiratory Research* 24, 299 (2023). https://doi.org/10.1186/s12931-023-02611-2
7. Li, C.-Y., Chang, K.-J., Yang, C.-F., Wu, H.-Y., Chen, W., Bansal, H., et al.: Towards a holistic framework for multimodal LLM in 3D brain CT radiology report generation. *Nature Communications* 16, 2258 (2025). https://doi.org/10.1038/s41467-025-57426-0
8. Zhao, Z., Liu, Y., Wu, H., Wang, M., Li, Y., Wang, S., et al.: CLIP in Medical Imaging: A Survey. *Medical Image Analysis* 102, 103551 (2025). https://doi.org/10.1016/j.media.2025.103551
9. Tu, T., Azizi, S., Driess, D., et al.: Towards generalist biomedical AI. arXiv:2307.14334 (2023). https://arxiv.org/abs/2307.14334
10. Hegselmann, S., Buendia, A., Lang, H., Agrawal, M., Jiang, X., Sontag, D.: TabLLM: few-shot classification of tabular data with large language models. arXiv:2210.10723 (2022). https://arxiv.org/abs/2210.10723

11. Song, X., Bahri, D.: Decoding-based Regression. *Transactions on Machine Learning Research* (2025). arXiv:2501.19383. https://arxiv.org/abs/2501.19383
12. OpenAI: Fine-tuning now available for GPT-4o. https://openai.com/index/gpt-4o-fine-tuning/ (2024)
13. Erickson, N., Mueller, J., Shirkov, A., et al.: AutoGluon-Tabular: robust and accurate AutoML for structured data. arXiv:2003.06505 (2020). https://arxiv.org/abs/2003.06505
14. Yang, Y., Zhang, H., Gichoya, J.W., et al.: The limits of fair medical imaging AI in real-world generalization. *Nature Medicine* 30, 2838–2848 (2024). https://doi.org/10.1038/s41591-024-03113-4