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

Anonymous Authors

Anonymous Affiliation
anonymous@example.com

Abstract. Spirometry is the clinical standard for pulmonary-function assessment, but its results depend on patient effort, operator expertise, and test availability. Chest CT provides effort-independent structural information, yet direct volumetric learning usually requires large annotated cohorts. We present GPT-DBR, a small-cohort experimental framework that applies decoding-based language-model regression to standardized CT-derived biomarkers after quantitative CT and radiomic feature extraction. In 164 paired CT–spirometry cases, GPT-DBR(qCT) achieved the lowest FEV1/FVC% error (MAE 6.90) and competitive FEV1% predicted error (MAE 15.01). In the held-out split, qCT inputs showed an adaptation-associated increase in FEV1/FVC-based label-recovery AUC from 0.578 to 0.857. These proof-of-concept findings suggest that language-model regression heads can use multidimensional CT biomarkers in data-scarce settings, but they are not sufficient for diagnostic or treatment decisions; 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 it depends on patient cooperation, trained operators, and local test availability. Chest CT offers a complementary view by capturing emphysema, airway narrowing, and lung-volume change without forced expiratory effort. Quantitative CT (qCT) and radiomics have therefore been used for COPD identification and lung-function estimation [3, 4]. Large CT-to-spirometry deep-learning studies also show that pulmonary function can be inferred from imaging when sufficient paired data are available [5, 6]. Prior work therefore indicates that CT can encode pulmonary-function-related structural information; the remaining question is whether a compact model can use multidimensional CT biomarkers when paired CT–PFT data are scarce.

Table 1. Baseline characteristics of the anonymized study cohort. Welch's t-test was computed using individual-level values before rounding.

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
FEV1/FVC (%)	55.2 ± 12.69	78.3 ± 5.39	< 0.001
FEV1 (% predicted)	80.9 ± 24.00	101.9 ± 23.70	< 0.001

Language models are most natural for text, classification, and report-generation tasks in medicine [7–9]. Prior tabular applications have also focused largely on classification [10]. Continuous regression is less straightforward because autoregressive decoders predict tokens rather than scalar values. Decoding-based regression (DBR) addresses this mismatch by representing numeric targets as generated strings or token sequences [11]. This motivates a different use of language models in medical imaging: a regression head over structured measurements extracted from images.

We investigate this idea in a paired chest CT–spirometry cohort. Instead of training a large volumetric model, standardized qCT and radiomic feature profiles are used as inputs to a language-model decoder that emits FEV1% predicted and FEV1/FVC%. The study makes three contributions. First, it reformulates spirometric estimation from CT-derived features as structured continuous-value generation. Second, it evaluates a two-stage DBR adaptation of a version-pinned GPT-family decoder without adding a dense regression head. Third, it compares this regression head with AutoML tabular ensembles, an untuned language-model baseline, radiomics variants, and an exploratory 3-D CNN comparator.

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 FEV1/FVC < 0.70 (70%); because GOLD recommends post-bronchodilator FEV1/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 the source institution (IRB no. ****-2012-34, partially anonymized for double-blind review), 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. The private source dataset is governed by institutional access terms; raw CT/PFT data are not publicly released.

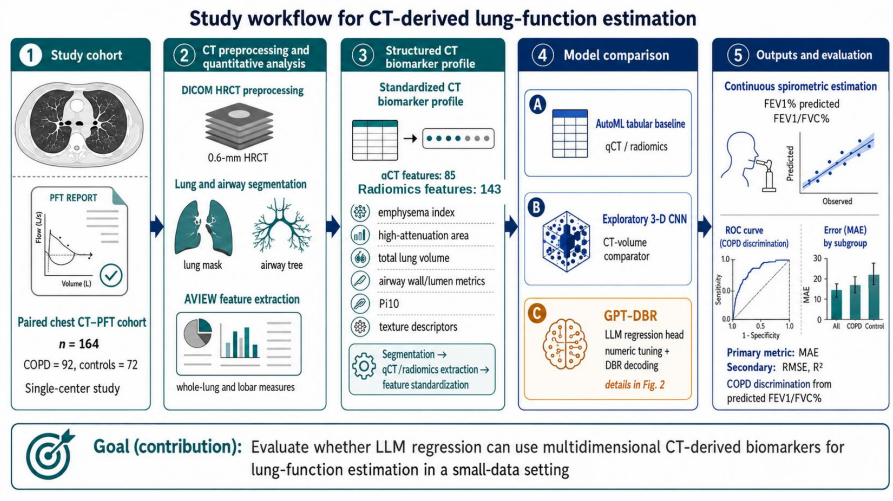

Fig. 1. Study workflow and contribution. Paired CT–PFT data were processed through DICOM HRCT preprocessing, lung/airway segmentation, AVIEW-based qCT and radiomics extraction, and feature standardization. The resulting CT biomarker profile was used for AutoML baselines, an exploratory 3-D CNN comparator,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

2.2 CT-Derived Feature Extraction

Using AVIEW software (Coreline Soft, Seoul, Korea; v1.0.2.5.9), imaging features were automatically extracted and subsequently confirmed by an experienced radiologist. We extracted 85 qCT features, including emphysema index, high-attenuation area, Pi10, and airway wall/lumen measurements across whole-lung, lobar, and segmental regions. We also extracted 143 radiomic features: first-order statistics (18), GLCM (24), GLRLM (16), GLSZM (16), NGTDM (5), GLDM (14), histogram (10), gradient (5), moment (7) and percentile features (28). qCT was treated as the main physiologically interpretable representation, while radiomics was used as a high-dimensional comparator.

2.3 GPT-DBR and Baselines

Figure 1 summarizes the study workflow. CT images are converted to qCT/radiomic biomarkers, assembled into a standardized CT biomarker profile, and provided to model families. For reproducibility, the reported experiments used GPT-4o-2024-08-06 as a version-pinned decoder via the OpenAI API [12]. We report this checkpoint because reviewers may need to identify the model used in the experiments. The methodological contribution is the DBR protocol and the structured-biomarker interface, rather than the recency or unique optimality of this checkpoint. Only de-identified numeric CT-derived feature vectors and target-token strings were submitted to the API; no PHI/PII, free text, DICOM metadata,

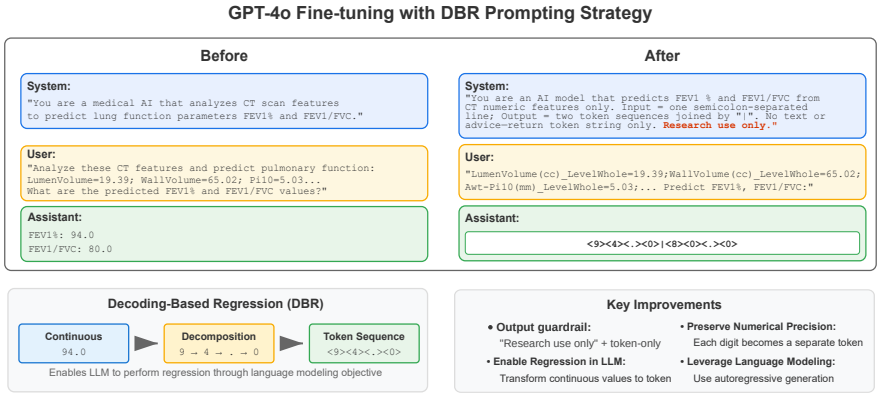

Fig. 2. GPT-DBR prompting and fine-tuning strategy. Free-form numeric responses are replaced by a structured numeric-feature prompt and a token-only response contract. Continuous spirometric targets are decomposed into digit-level token sequences, aligning regression with the autoregressive language-model objective.

accession numbers, dates, or institution identifiers were submitted. Institutional privacy and business-associate-agreement applicability were checked prior to API use.

Figure 2 details the fine-tuning strategy. Stage 1 trained the decoder to output raw decimal spirometric values. Stage 2 revisited the same instances after serializing continuous targets into numeric tokens, enforcing the DBR objective without a task-specific regression head [11]. Fine-tuning used two epochs, automated hyperparameter selection, batch size 1, learning-rate multiplier 2, and seed 42. 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. Because the cohort is small for end-to-end volumetric training, the 3-D CNN is interpreted only as an exploratory image-based comparator.

2.4 Split, Preprocessing, and Metrics

The dataset was stratified at the subject level into 98 training, 16 validation, and 50 held-out test subjects (approximately 60/10/30%). All preprocessing steps that could learn from data, including imputation, standardization, scaling, and any feature selection, were fit using the training split only and applied without refitting to validation and test subsets. MAE was the primary metric; MSE, RMSE, signed R^2, Pearson correlation, and median absolute error were secondary metrics. During metric auditing, we identified that the AutoML summary script applied an absolute-value transform to all returned metrics. We therefore restored the sign of signed metrics, including R^2, before reporting Table 2. For language-model outputs, each test input was generated ten times; the reported

standard deviations quantify stochastic generation variability, not patient-level confidence intervals. Inputs were semicolon-separated feature-name=value lines; DBR outputs were two digit-token sequences separated by "|" and decoded by concatenating digits, decimal points, and optional signs. AutoGluon used the same tabular matrices and validation split for model selection, and the 3-D CNN was an exploratory I3D-style CT-volume comparator trained only on the training split. AVIEW v1.0.2.5.9, GPT-4o-2024-08-06, prompts, decoding scripts, metric code, environment files, and exact generation settings are archived for review where policy permits.

3 Results

Table 2 summarizes held-out performance after restoring signed R^2 values. For FEV1% predicted, GPT-Num(qCT) had the lowest point-estimate MAE among the language-model variants (14.24±0.70), while GPT-DBR(qCT) remained competitive (MAE 15.01±1.04, RMSE 19.21±1.34, $R^2 = 0.340$). 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 6.90±0.54, RMSE 8.15±0.62, and $R^2 = 0.728$. 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 17.50, $R^2 = -0.248$). The untuned base language-model decoder failed for FEV1/FVC% (MAE 63.47±0.003), consistent with scale or output-format instability before DBR adaptation.

Table 2. Held-out regression performance after signed-R^2 audit. MAE is the primary metric. R^2 is reported with its sign preserved; AutoML R^2 values were corrected after identifying that the original summary script applied an absolute-value transform to all metrics.

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	24.11	29.43	-1.106	63.47	64.94	-21.265
GPT-base(qCT)	qC featureT	20.95	25.79	-0.617	63.47	64.94	-21.266
GPT-Num(RM)	Radiomic feature	15.81	21.58	-0.139	13.07	17.43	-0.607
GPT-Num(qCT)	qCT feature	14.24	19.07	0.114	7.96	9.63	0.509
GPT-DBR(RM)	Radiomic feature	22.80	40.71	-5.072	14.26	18.63	-0.416
GPT-DBR(qCT)	qCT feature	15.01	19.21	0.340	6.90	8.15	0.728

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 numeric-plus-token decoding-based regression. GPT entries are means across 10 stochastic generations; their standard deviations are shown in the figures/text. Signed R^2 was not converted to an absolute value.

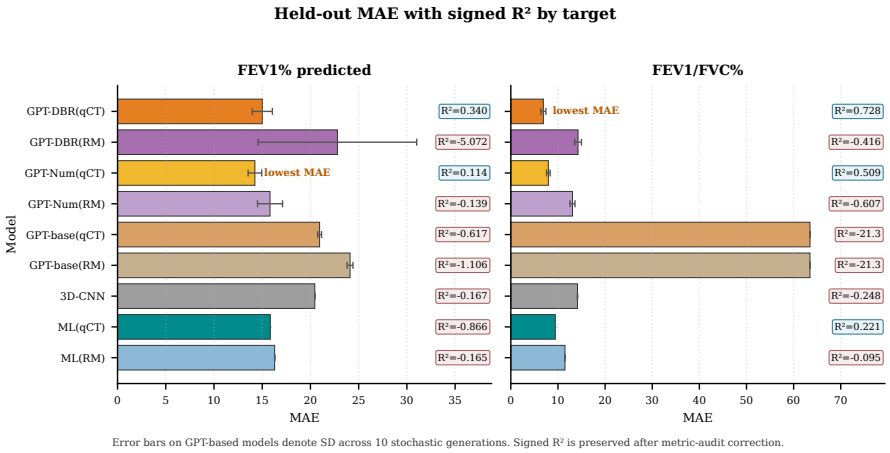

Fig. 3. Target-wise MAE comparison using a unified model naming scheme. Each bar reports held-out MAE; right-side badges report signed target-specific R^2 values after correcting the AutoML summary sign issue. Error bars on GPT variants denote SD across 10 stochastic generations.

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 it substantially stabilized the more scale-sensitive FEV1/FVC% task.

Figure 4 summarizes ROC-AUC progression for COPD discrimination derived from predicted FEV1/FVC%. AUC was computed using the inverse predicted FEV1/FVC% as a continuous score against the spirometry-defined COPD label. This evaluates recovery of the label-defining physiologic index, not an independent diagnostic endpoint. With qCT inputs, AUC increased from 0.578±0.079 for GPT-base(qCT) to 0.825±0.016 after GPT-Num(qCT) and 0.857±0.039 after GPT-DBR(qCT). Radiomic-feature inputs remained close to chance across the same progression.

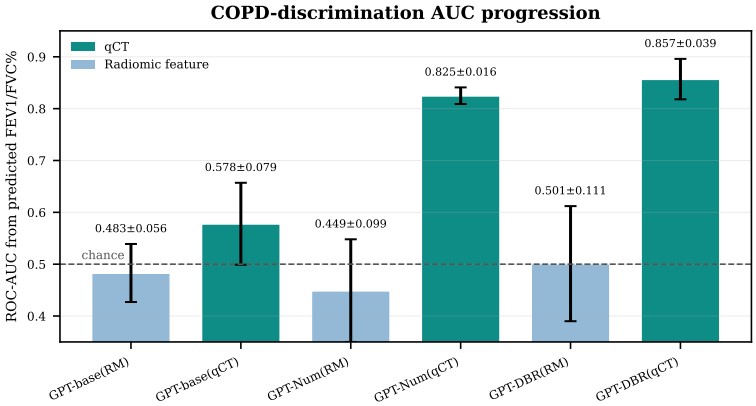

Fig. 4. ROC-AUC progression for COPD discrimination derived from predicted FEV1/FVC%. AUC uses $-\widehat{\text{FEV1/FVC}}\%$ as a continuous score against the spirometry-defined COPD label. Error bars denote SD across 10 stochastic generations, not patient-level confidence intervals.

4 Discussion and Conclusion

This study frames language-model use in medical imaging as post-extraction structured biomarker regression. GPT-DBR operates on qCT and radiomic biomarkers extracted from segmented CT scans. This design reduces the data burden and makes the method suitable for early-stage, data-scarce medical image computing studies, while keeping the claims focused on structured biomarkers.

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 is therefore important: it translates continuous spirometric outcomes into a sequence-generation format that can be optimized by an autoregressive decoder. The strongest effect appeared for FEV1/FVC%, where the base language model was unstable and DBR reduced the MAE to 6.90. This supports the idea that explicit numeric adaptation is essential for scale-sensitive biomedical regression.

The difference between FEV1% predicted and FEV1/FVC% performance is clinically and statistically plausible. FEV1/FVC% is the ratio most directly tied to airflow obstruction and is the index used to establish airflow limitation, whereas FEV1% predicted reflects a broader severity scale normalized by reference equations and influenced by demographic and non-imaging factors. In this cohort, FEV1/FVC% showed much stronger COPD-control separation than FEV1% predicted, suggesting a stronger structural phenotype signal for qCT-based learning.

The comparison between qCT and radiomics suggests, but does not prove, a feature-quality versus feature-quantity trade-off. qCT features such as emphy-

sema index, high-attenuation area, lung volume, airway wall/lumen measurements, and Pi10 are physiologically interpretable and directly linked to COPD phenotypes. By contrast, high-dimensional radiomics can capture texture and distributional information, but in a small cohort it may add redundancy and overfitting risk. Radiomics should therefore be reassessed in larger hybrid or externally validated designs.

In the held-out split, the qCT DBR model achieved a label-recovery AUC of 0.857 when COPD labels were scored by predicted FEV1/FVC%. This result should be interpreted as recovery of the spirometric index used to define the label, not as an independent diagnostic endpoint. The present proof-of-concept analysis is therefore not sufficient for diagnostic or treatment decisions, FEV1% MAE remained too large for treatment decisions, and spirometry remains the reference test.

Several limitations remain. First, this is a single-center proof-of-concept without external validation. External 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 because the dataset was small for end-to-end volumetric learning. Fourth, stochastic generation variability should not be interpreted as patient-level confidence intervals, and no prospective clinical utility analysis was performed. Fifth, the GPT-based models depend on a closed, version-pinned API; prompts, decoding rules, generation settings, and metric scripts should therefore be archived with the review package. Finally, although signed R^2 values were restored in this revision, raw held-out prediction files should be archived so that all metrics can be recomputed under a single aggregation rule during review or camera-ready preparation.

GPT-DBR provides a lightweight experimental framework for estimating spirometric indices from CT-derived structured biomarkers using DBR. In this cohort, GPT-DBR(qCT) achieved competitive FEV1% prediction and the lowest FEV1/FVC% point-estimate error among the reported variants. These proof-of-concept findings support further investigation of language-model regression heads for structured medical imaging biomarkers, while external validation remains necessary before any clinical translation.

AI-use disclosure. Large-language-model tools were used only as general-purpose writing assistance, copy-editing, and consistency-checking tools during manuscript preparation; the experimental GPT-4o model use is described in the Methods. All scientific claims, references, analyses, figures, tables, and conclusions were verified by the human authors, who take full responsibility for the work. LLM tools are not listed as authors and did not generate patient-level data or final statistical results.

Acknowledgments

Omitted for double-blind review.

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