Fine-tuning Zero-shot Large Language Models for Patient-reported Outcomes (Student Abstract)

Yang Yan[1], **Matthew W. Chen**[2], **Jiayi Lyu**[3], **Chen Zhao**[4], **Hao Gao**[5], **Zhong Chen**[1]
[1]School of Computing, Southern Illinois University
[2]Department of Radiation Oncology, University of Kansas Medical Center
[3]Electrical and Computer Engineering, Southern Illinois University
[4]Department of Computer Science, Baylor University
[5]Department of Radiation Oncology, UT Southwestern Medical Center
{yang.yan, jiayi.lyu, zhong.chen}@siu.edu, matthewepic360@gmail.com, chen_zhao@baylor.edu, hao.gao.2012@gmail.com

Abstract

Radiotherapy (RT) is a cornerstone of cancer treatment. Following RT, patient-reported outcomes (PROs) collected via standardized questionnaires are crucial for monitoring patients' quality of life and side effects. However, traditional statistical and machine learning methods, which rely on structured numerical data, often fail to capture semantic meaning within patients' health status. To address this, we developed a novel framework using zero- and few-shot large language models (LLMs) to identify patients experiencing mild to severe depression. Furthermore, classification performance is enhanced through parameter-efficient fine-tuning. Experiments on a prostate cancer PRO dataset for depression have demonstrated that our fine-tuned LLMs consistently outperformed other baseline methods across key evaluation metrics.

Introduction

Prostate cancer is one of the most frequently diagnosed cancers in men, for which radiotherapy (RT) is a cornerstone of curative treatment (Chen et al. 2021; Ahire et al. 2023; Chen et al. 2025). Despite its efficacy in tumor control, RT often induces significant treatment-associated toxicities that adversely affect patients' quality of life. Patient-reported outcomes (PROs) (Yan et al. 2025a,b), systematically captured by standardized questionnaires, have become an essential tool for toxicity monitoring, enabling clinicians to track recovery trajectories and proactively identify those at risk.

While machine learning (ML) and deep learning (DL) hold significant potential for tracking PROs in prostate cancer (Verma, Bach, and Mork 2021; Yan et al. 2024, 2025c), their conventional application faces a critical limitation. These methods (Verma, Bach, and Mork 2021) often treat textual PRO data as structured numerical variables, thereby overlooking the rich semantic and emotional nuances embedded in patients' natural language. Consequently, they frequently fail to detect subtle yet clinically vital signals of distress or nuanced shifts in symptoms. This shortcoming is especially critical for identifying depressive symptoms, where the accurate and early detection for patients with severe depression is essential (Yan et al. 2024, 2025a).

With the development of large language models (LLMs) (Li et al. 2025; Lin et al. 2025), they offer a powerful solution to overcome the limitations of traditional ML/DL in RT toxicity tracking. Trained on massive corpora, LLMs are able to interpret semantics, sentiment, and context directly from text-based PROs, reducing reliance on handcrafted features. Zero-shot and few-shot prompting enable LLMs to adapt to clinical tasks without extensive task-specific annotation, while parameter-efficient fine-tuning strategies, such as Low-Rank Adaptation (LoRA) (Hu et al. 2022), allow for specialization in model domains under limited computational budgets. In this work, we develop an LLM-based anomaly detection framework to identify prostate cancer patients who experience mild to severe depression during RT.

Methods

Zero-shot AD-LLM Prompts

We implement two distinct experimental settings using the AD-LLM framework (Yang et al. 2024) based on Meta's Llama 3.1 8B Instruct model. The first, termed Llama-3.1-8B Instruct Mild, is primed solely on the characteristics of mild depressive symptoms, and the second setting is Llama-3.1-8B Instruct Mild+Severe, where both mild and severe depressive symptoms are described in the prompts. Our zero-shot AD-LLM prompt is architecturally structured around four integral components to ensure robust anomaly detection. First, it provides a clear definition of PROs, developed in collaboration with radiation oncologists to delineate scenarios indicative of severe depression, thereby grounding the LLM's understanding of textual anomalies. The second component is a strict scoring mechanism that supplies the LLM with explicit criteria for quantifying an anomaly score. Third, a step-by-step reasoning process, facilitated by a Chain-of-Thought (CoT) approach (Wei et al. 2022), is embedded to guide the LLM in logically deriving its rationale before producing a final score. Finally, a structured response is saved into a JSON file for further analysis.

Few-shot AD-LLM Prompts

Building upon the zero-shot AD-LLM prompt architecture, we further developed a few-shot prompting strategy for the Meta Llama 3.1 8B Instruct model to enhance its performance. Motivated by the zero-shot model, we collaborated

Model	Prompt Strategy	Depression				
		AUC	AUCPR	Precision	Recall	F1
Zero-shot AD-LLM	Mild	0.6555	0.4993	0.4774	0.8571	0.6132
	Mild + Severe	0.7274	0.5799	0.5362	0.8346	0.6529
Few-shot AD-LLM	Mild	0.7235	0.5783	0.5562	0.7652	0.6441
	Mild + Severe	0.7260	0.5792	0.5487	0.7682	0.6402
Normal-shot AD-LLM	Mild	0.7369	0.6002	0.5522	0.7852	0.6484
	Mild + Severe	0.7461	0.6095	0.5229	0.8485	0.6470
Anomaly-shot AD-LLM	Mild	0.7073	0.5633	0.4640	**0.8810**	0.6079
	Mild + Severe	0.7222	0.5898	0.5326	0.7702	0.6297
Fine-tuning Zero-shot AD-LLM	Mild + Severe	**0.7497**	**0.6111**	**0.5665**	0.8008	**0.6809**

Table 1: Overall comparison of different prompting strategies. The best results are highlighted in bold.

with radiation oncologists to augment the existing prompt structure, comprising definition, scoring, CoT reasoning, and response format, with a critical new example section. We designed and evaluated three distinct exemplar strategies: Few-shot, which provides five examples each of mild and severe depression PRO texts complete with their reasoning processes; Normal-shot, which provides only five examples of mild depression; and Anomaly-shot, which provides five examples exclusively of severe depression.

Fine-tuning Zero-shot AD-LLM Prompts

We further employ LoRA (Hu et al. 2022) to fine-tune the Meta Llama 3.1 8B Instruct model. This approach efficiently specializes the model for the nuances of prostate PRO data while crucially preserving the vast general linguistic knowledge encoded in the original pre-trained Llama model. LoRA is a parameter-efficient fine-tuning method that avoids the computational expense of updating the parameters. Instead, it freezes the pre-trained weights and injects trainable, low-rank decomposition matrices into the attention and feed-forward layers. By constraining weight updates to the product of two smaller matrices, LoRA reduces the number of trainable parameters while retaining the expressive power needed for the PRO task adaptation.

Experiments

PRO Dataset

The prostate PRO dataset comprises 22 instruments measuring urinary-related functions, bowel-related functions, and sexual-related functions, featuring depression feelings after each RT as the target evaluation. Responses were classified as mild depression feelings ("No problem", "Very small problem", "Small problem") or severe depression feelings ("Moderate problem", "Big problem"). In total, this dataset consists of 10,931 PRO records, including 10,266 cases of mild depression and 665 cases of severe depression. For fine-tuning LLMs, we split data into three partitions with 50% for training, 10% for validation, and 40% for testing.

Overall Performance Comparison

Table 1 compares the overall performance across AUC, AUCPR, Precision, Recall, and F1 score. The fine-tuned zero-shot AD-LLM model consistently outperforms all other prompting strategies, including standard zero-shot, few-shot, and normal-shot. This superior performance validates the significant effectiveness of parameter-efficient LoRA fine-tuning, demonstrating its critical advantage in mitigating class imbalance and achieving greater robustness beyond what is possible with prompting strategies alone.

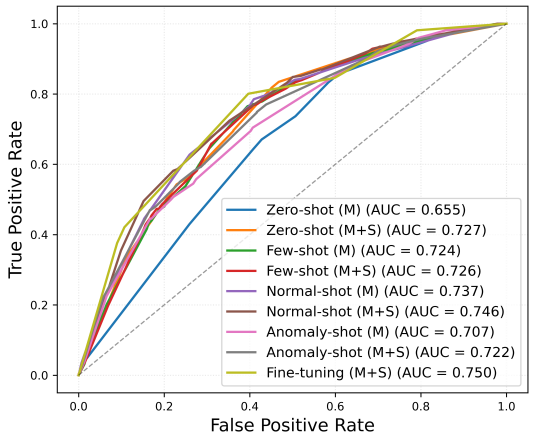

Figure 1. ROC curves of prompting LLMs.

ROC Curves and AUC Comparison

Figure 1 presents the ROC curves for five prompt strategies. The fine-tuned zero-shot Llama 3.1 8B Instruct model with a LoRA adapter achieved the best overall performance, attaining the highest AUC score of 0.750. Among the non-fine-tuned models, the normal-shot strategies (using either mild-only or mild+severe definitions) yielded competitive results with scores of 0.737 and 0.746, generally outperforming the few-shot and zero-shot approaches.

Conclusion

This study has proposed various prompting strategies to identify severe depression for prostate cancer patients undergoing radiotherapy. Our results demonstrate that fine-tuned LLMs can capture nuanced semantic and emotional context in PRO data, enabling early detection of severe symptoms.

Acknowledgments

This work has been supported in part by an Illinois Innovation Network (IIN) sustaining Illinois seed funding grant.

References

Ahire, V.; Ahmadi Bidakhvidi, N.; Boterberg, T.; Chaudhary, P.; Chevalier, F.; Daems, N.; Delbart, W.; Baatout, S.; Deroose, C. M.; Fernandez-Palomo, C.; et al. 2023. Radiobiology of combining radiotherapy with other cancer treatment modalities. In *Radiobiology Textbook*, 311–386.

Chen, Z.; Li, W.; Shen, X.; Chen, R. C.; Lin, Y.; and Gao, H. 2025. A few-shot u-net learning framework for fast and accurate three-dimensional dose prediction in radiotherapy. *Physica Medica*, 139: 105184.

Chen, Z.; Zhang, W.; Deng, H.; and Zhang, K. 2021. Effective cancer subtype and stage prediction via dropfeature-DNNs. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 19(1): 107–120.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Li, D.; Zhao, X.; Yu, L.; Liu, Y.; Cheng, W.; Chen, Z.; Chen, Z.; Chen, F.; Zhao, C.; and Chen, H. 2025. SolverLLM: Leveraging Test-Time Scaling for Optimization Problem via LLM-Guided Search. *arXiv preprint arXiv:2510.16916*.

Lin, Y.; Li, D.; Wu, X.; Shao, M.; Zhao, X.; Chen, Z.; and Zhao, C. 2025. Face4FairShifts: A Large Image Benchmark for Fairness and Robust Learning across Visual Domains. *arXiv preprint arXiv:2509.00658*.

Verma, D.; Bach, K.; and Mork, P. J. 2021. Application of machine learning methods on patient reported outcome measurements for predicting outcomes: a literature review. In *Informatics*, volume 8, 56.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*, 35: 24824–24837.

Yan, Y.; Chen, Z.; Shen, X.; Chen, R. C.; and Gao, H. 2025a. Short-and long-term weekly patient-reported outcomes prediction undergoing radiotherapy: single-patient time series model vs. transformer-based multi-patient time series model. *BioData Mining*, 18(1): 53.

Yan, Y.; Chen, Z.; Xu, C.; Shen, X.; Shiao, J.; Einck, J.; Chen, R. C.; and Gao, H. 2025b. An oversampling-enhanced multi-class imbalanced classification framework for patient health status prediction using patient-reported outcomes. *IEEE Access*.

Yan, Y.; Lominska, C.; Gan, G. N.; Gao, H.; and Chen, Z. 2024. Accounting for Cancer Patients with Severe Outcomes: An Anomaly Detection Perspective. In *IEEE BigData*, 8253–8255.

Yan, Y.; Lominska, C.; Gan, G. N.; Gao, H.; and Chen, Z. 2025c. Advanced Anomaly Detection Framework for Enhancing Prediction of Severe Health Outcomes in Cancer Patients Undergoing Radiotherapy. In *IEEE BigDataService*, 73–80.

Yang, T.; Nian, Y.; Li, S.; Xu, R.; Li, Y.; Li, J.; Xiao, Z.; Hu, X.; Rossi, R.; Ding, K.; et al. 2024. Ad-llm: Benchmarking large language models for anomaly detection. *arXiv preprint arXiv:2412.11142*.