Poster P07 | ASM-ABS1298

SAMSUNG MEDICAL CENTER

Fine-Tuning a Small Language Model on Limited Synthetic Data for LI-RADS Report Feature Extraction

Jeong Hyun Lee, M.D.

Department of Radiology, Samsung Medical Center

Key Findings

- → A parameter-efficiently fine-tuned **8B on-prem LLM** can extract LI-RADS features from liver imaging reports with **88.8% accuracy (213/240 features correct)**, close to GPT-4o while preserving data privacy.
- → Trained on **100 synthetic bilingual (Ko–En) reports** from a set of 120 (100 training, 20 validation), using an **exact-match reward**; finished within **<24 hours on a single 48-GB VRAM workstation**.
- → **No statistically significant difference** vs GPT-4o on accuracy/MCC under cluster bootstrap (Acc Δ=2.5 pp, 95% CI −0.83–5.83; p=0.2006; MCC Δ=0.040, 95% CI −0.019–0.099; p=0.1894).

Background

Structured data extraction from narrative liver imaging reports is essential for clinical workflows and research, yet manual extraction is labor-intensive and error-prone. While commercial large language models (LLMs) have demonstrated high accuracy in clinical information extraction tasks[1], privacy concerns and infrastructure constraints often preclude their use in secure healthcare environments.

The Liver Imaging Reporting and Data System (LI-RADS) provides a standardized framework for reporting liver observations[2,3], but extracting structured LI-RADS features from free-text radiology reports remains challenging. Large proprietary models demonstrate strong performance in data labeling tasks but cannot protect data privacy[1]. An on-premises solution that maintains comparable performance while ensuring data privacy would address a critical clinical need.

Objective

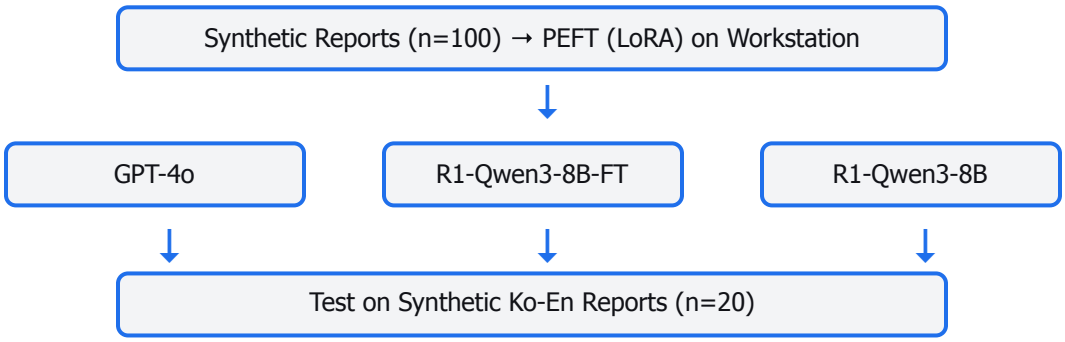
To develop and evaluate a compact, locally deployable language model capable of extracting structured features from CT/MRI radiology reports—including **size, location, features of chronic liver disease (CLD), and LI-RADS imaging features**—with performance comparable to a proprietary frontier model while operating under strict privacy constraints and resource limitations.

Methods

Study Design & Data

Training dataset: 120 synthetic mixed Korean–English liver imaging reports were created by institutional radiologists to emulate routine clinical reporting while preserving confidentiality. Reports were randomly split into 100 for training and 20 for validation, each paired with corresponding feature labels.

Task definition: Extract 12 features from the index lesion (defined as the largest suspicious lesion; ties resolved by first occurrence): size, location, features of CLD, LR-M targetoid features, LR-M non-targetoid features, tumor-in-vein, arterial phase hyperenhancement, non-peripheral washout, enhancing capsule, threshold growth, ancillary features favoring benignity, and ancillary features favoring malignancy.



Model & Training

Base model: DeepSeek-R1-0528-Qwen3-8B (abbreviated as R1-Qwen3-8B)

Fine-tuning: Parameter-efficient fine-tuning (PEFT) using LoRA (Low-Rank Adaptation)[4] with reinforcement learning via Group Relative Policy Optimization (GRPO)[5]. The reward function rewards accurate feature extraction while penalizing incorrect output formatting and excessive length.

Hardware: Workstation with one NVIDIA A6000 (48GB VRAM)

Training time: 17 hours

Evaluation

Test set: 20 validation reports (240 observations total)

Comparators: Fine-tuned R1-Qwen3-8B (LoRA) vs GPT-4o

Metrics: Overall accuracy and Matthews correlation coefficient (MCC)

Statistical analysis: 95% confidence intervals via cluster bootstrap at report level (b=10,000 iterations)

Significance testing: Paired cluster bootstrap for model comparison

Output Format

Models were prompted to return structured output as CSV with a header row followed by 12 feature rows. Each row contains a feature name and its value (numeric for size/location, binary 0/1 for categorical features).

Results

88.8%
R1-Qwen3-8B-FT Accuracy

91.3%
GPT-4o Accuracy

p=0.20
Not Significant

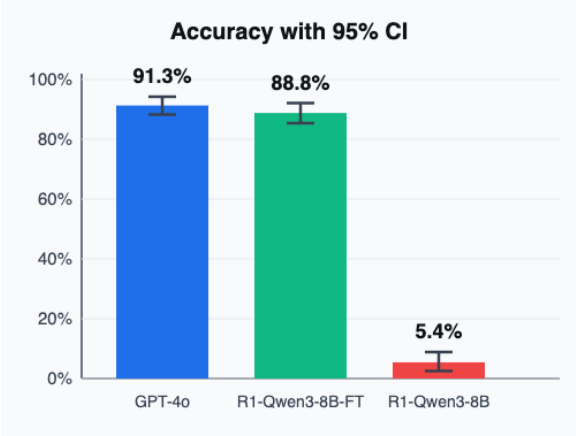


Figure 1. Model accuracy with 95% confidence intervals computed via report-level cluster bootstrap (b=10,000). Error bars represent 95% CI.

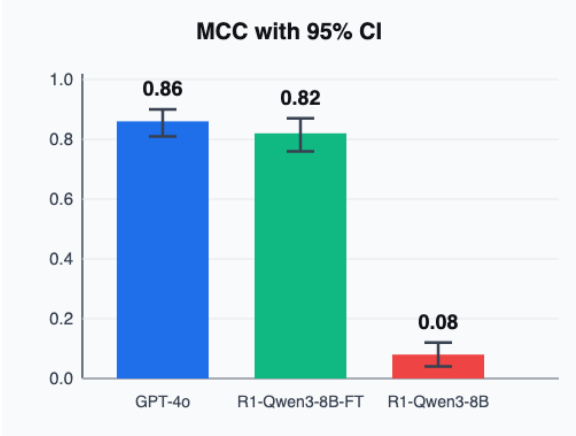


Figure 2. Matthews correlation coefficient (MCC) with 95% confidence intervals computed via report-level cluster bootstrap (b=10,000). Error bars represent 95% CI.

Overall Performance (n=240 observations / 20 reports)

Model	Accuracy (%)	95% CI	MCC	95% CI
GPT-4o	91.3	88.3–94.2	0.86	0.81–0.90
R1-Qwen3-8B-FT	88.8	85.4–92.1	0.82	0.76–0.87
R1-Qwen3-8B	5.4	2.5–8.8	0.08	0.04–0.12

Paired Comparison (GPT-4o − R1-Qwen3-8B-FT)

Metric	Difference	95% CI	p-value
Accuracy (pp)	+2.5	−0.8 to +5.8	0.20
MCC	+0.04	−0.02 to +0.10	0.19

Discussion

This study demonstrates that an 8-billion-parameter language model (R1-Qwen3-8B) fine-tuned with parameter-efficient methods on a limited synthetic dataset can achieve performance for structured feature extraction that is statistically indistinguishable from GPT-4o.

The fine-tuned model achieved 88.8% accuracy (95% CI 85.4–92.1) and MCC of 0.82 (95% CI 0.76–0.87), compared to GPT-4o's 91.3% accuracy (88.3–94.2) and MCC of 0.86 (0.81–0.90). Paired cluster bootstrap analysis revealed no statistically significant differences between the models (p=0.20 for accuracy, p=0.19 for MCC).

Key advantages of this approach:

- **Privacy preservation:** All data remain on-premises; no PHI transmitted to external services
- **Resource efficiency:** Training completed in <24 hours on a single consumer-grade workstation
- **Data efficiency:** High performance achieved with only 100 synthetic training examples

The use of synthetic data generated by institutional radiologists enabled realistic clinical scenarios while completely avoiding PHI concerns. Reinforcement learning via Group Relative Policy Optimization (GRPO) proved effective for this task, with the reward function guiding the model to extract features accurately while maintaining proper output formatting. Reinforcement learning approaches are more resistant to overfitting compared to supervised fine-tuning, as RL generalizes rather than memorizes training patterns and can effectively reuse the same training prompts through multiple stochastic rollouts that provide diverse learning signals[6,7].

Limitations

Several limitations warrant consideration. First, the small validation set of only 20 reports is the most significant limitation, as it may not provide sufficient statistical power to detect meaningful performance differences or capture the full range of clinical variation. Second, validation was performed on synthetic reports from the same institutional source; prospective testing on real-world clinical reports is needed to confirm performance. Third, the synthetic training dataset, while clinically realistic, comprised only 100 reports, which may limit generalization to the full diversity of clinical reporting styles.

Conclusion

A compact 8-billion-parameter language model fine-tuned with parameter-efficient methods on a small synthetic bilingual dataset **matches GPT-4o performance** for feature extraction while remaining **on-premises and privacy-preserving**. This approach is feasible on modest hardware for healthcare institutions requiring secure, local deployment of AI-assisted clinical information extraction tools.

References

1. Gu, H., Lee, J. H., Kim, Y. J., Park, S. H., & Kim, S. Y. (2024). Using GPT-4 for LI-RADS feature extraction and categorization with multilingual free-text reports. *Liver International*, *44*(6), 1391–1402. https://doi.org/10.1111/liv.15891

2. American College of Radiology. (n.d.). *Liver Imaging Reporting and Data System (LI-RADS)*. https://www.acr.org/Clinical-Resources/Reporting-and-Data-Systems/LI-RADS

3. Chernyak, V., Fowler, K. J., Kamaya, A., Kielar, A. Z., Elsayes, K. M., Bashir, M. R., Kono, Y., Do, R. K., Mitchell, D. G., Singal, A. G., Tang, A., & Sirlin, C. B. (2018). Liver Imaging Reporting and Data System (LI-RADS) version 2018: Imaging of hepatocellular carcinoma in at-risk patients. *Radiology*, *289*(3), 738–753. https://doi.org/10.1148/radiol.2018181494

4. Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*. https://arxiv.org/abs/2106.09685

5. Guo, D., Yang, D., Zhang, H., Hu, S., Huang, Z., Zhou, A., Zhang, W., Wu, S., Liu, T., Liao, Y., Liang, C., Wang, R., Fan, L., Gao, Y., Dong, Y., Sun, R., Liu, A., Zhao, Y., Hong, S., ... Wu, Y. (2025). DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, *645*, 633–638. https://doi.org/10.1038/s41586-025-09422-z

6. Chen, J., Liu, H., Zheng, C., Gao, C., Peng, J., Huang, S., Lin, Z., Wang, W., Sun, D., Zan, D., & Gao, G. (2024). Unlock the correlation between supervised fine-tuning and reinforcement learning in training code large language models. *arXiv preprint arXiv:2406.10305*. https://arxiv.org/abs/2406.10305

7. Chu, T., Zhai, Y., Yang, J., Tong, S., Xie, S., Schuurmans, D., Le, Q. V., Levine, S., & Ma, Y. (2025). SFT memorizes, RL generalizes: A comparative study of foundation model post-training. *arXiv preprint arXiv:2501.17161*. https://arxiv.org/abs/2501.17161