This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

FedPref: Federated Preference Learning for Structured Radiology Report Extraction

Flint Xiaofeng Fan[1,2], Cheston Tan[2], Yew-Soon Ong[2], and Roger Wattenhofer[1]

[1] ETH Zurich, Zurich, Switzerland
{xiafan,wattenhofer}@ethz.ch
[2] Agency for Science, Technology and Research (A*STAR), Singapore, Singapore
{fanx,cheston-tan,ong_yew_soon}@a-star.edu.sg

Abstract. Radiology reports describe findings and locations in free text, but downstream search and analysis require these relations in a fixed schema. Learning this extraction requires labels that are unevenly distributed across institutions: smaller hospitals have less local evidence, and pooling data may be infeasible. We introduce FedPref: frozen public language models propose alternative JSON extractions, local annotations rank them, and sites collaboratively train compact Qwen3-8B adapters while sharing only model updates. A heterogeneous teacher pool provides cross-model contrast when repeated single-model samples collapse. On development data from six simulated hospitals with unequal data volume and disease prevalence, FedPref improves client-mean F1 by 2.49 points and worst-site F1 by 9.10 points compared with training each site in isolation, with the largest gains at the sites holding the least data. Central training on the pooled preference-pair union is 2.66 points higher on client-mean F1. On a locked, 400-report manually validated gold test set, FedPref reaches 68.68 F1 and pooled training 71.67, preserving that same ordering. FedPref thus lets institutions with unequal, unpooled data benefit from collaboration without ever sharing reports or annotations.

Keywords: Federated learning · Preference learning · Radiology reports · Structured extraction

1 Introduction

Radiology reports encode two linked facts: which findings are present and where they occur. A report describing a left pleural effusion should yield the disease and its location; assigning the disease to the wrong anatomy remains an extraction error. Structuring these relations makes reports queryable and reusable as supervision. We study nine diseases and their locations in the centralized benchmark built from MIMIC-CXR and Chest ImaGenome [6, 14, 12].

Institutions observe unequal numbers and combinations of cases. Centralized training broadens coverage by pooling data; isolated training preserves institutional boundaries but leaves smaller hospitals with less evidence. Federated learning offers a third path by aggregating locally optimized model states [9,

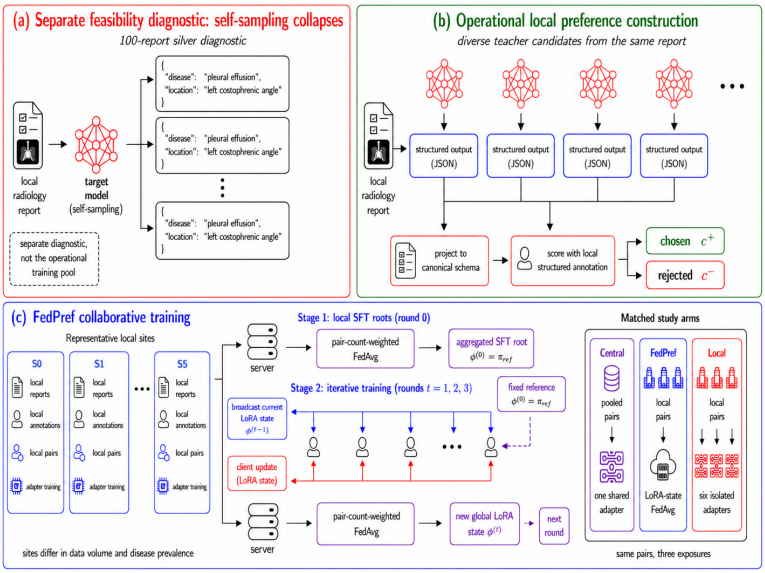

Fig. 1. FedPref intuition and study overview. (a) A separate feasibility diagnostic shows that repeated self-sampling from the target model can collapse to nearly identical structured outputs, leaving little preference contrast. (b) Operationally, heterogeneous frozen teachers generate alternative JSON extractions for the same local report. The site projects them into a shared schema and scores them with its structured annotation to form chosen c^+ and rejected c^-. (c) Sites train LoRA adapters on their own preference pairs and exchange only adapter states. Pair-count-weighted FedAvg forms the initial SFT root and aggregates subsequent fixed-reference DPO updates; reports, annotations, and preference pairs remain local. Central, FedPref, and Local denote pooled, federated, and isolated training under matched pair exposure.

11, 4, 1]. The schema and annotation rule stay fixed while sites differ in patient volume and disease prevalence.

FedPref unifies preference construction and cross-site learning (Fig. 1). A separate feasibility diagnostic first tests naive self-sampling: repeated target-model samples often collapse to one structure, leaving little rankable contrast. Operationally, four frozen teachers instead propose alternative JSON extractions for the same report. Each site projects them into a common schema and uses its structured annotation to form a chosen response c^+ and rejected response c^-. This tests our first question: whether cross-model candidates supply useful preference contrast.

Each site performs a short supervised fine-tuning (SFT) stage on its chosen responses, then uses Direct Preference Optimization (DPO) to teach a Qwen3-8B adapter to favor c^+ over c^- [10]. Sites exchange only compact low-rank adaptation (LoRA) states, which pair-count-weighted Federated Averaging (FedAvg)

combines into the next shared adapter. Reports, annotations, candidates, and preference pairs remain local.

Our second question is whether aggregating these local preferences across sites improves client-mean and worst-site performance over isolated training, and how far it remains below pooled Central training. We evaluate three matched arms: Local trains each site in isolation, FedPref aggregates site adapters, and Central pools all preference pairs. On development data across three rounds, FedPref improves client-mean and worst-site extraction over Local, with the largest gains at the two smallest sites. Central measures the observed pooled–federated gap under matched exposure, and a locked manually validated gold test checks generalization beyond automatically derived labels. The study therefore tests both the prerequisite for preference learning and the value of cross-site aggregation.

2 Related Work

MIMIC-CXR and Chest ImaGenome provide report-level and anatomy-linked supervision, while Sabour et al. formulate the centralized structured-extraction task used here [6, 14, 12].

Federated learning coordinates models without centralizing raw institutional records, making it attractive for medical data silos [9, 11]. Data locality alone does not guarantee privacy: formal protection requires a threat model and mechanisms such as secure aggregation or differential privacy [3]. FedAvg combines local states, while LoRA reduces the number of trainable parameters communicated [9, 5]. Factorwise LoRA averaging can nevertheless introduce an optimization gap [13], and unequal client utility motivates reporting both average and lower-tail performance [7].

DPO learns directly from response rankings, and federated variants distribute feedback-driven, personalized, or medical-dialogue alignment [10, 2, 8]. FedPref connects these strands by using local disease-location annotations to rank heterogeneous teacher outputs and aggregating only the resulting LoRA states.

3 Method

Overview. Site k holds report-annotation pairs $\mathcal{S}_k = \{(x_i, y_i)\}$. For report x_i, teachers propose structures and annotation y_i designates the best and worst eligible outputs as chosen response c_i^+ and rejected response c_i^-. Supervised fine-tuning (SFT) on chosen responses establishes the schema and fixed reference; DPO then refines which valid structures the model favors. FedPref aggregates LoRA states, Central trains on the pair union, and Local trains independently.

Local preference construction. Four frozen teachers, Qwen3-14B, Qwen2.5-14B-Instruct, Llama-3.1-8B-Instruct, and Mistral-7B-Instruct-v0.3, each generate one completion; the Qwen3-8B target is excluded [15]. Teachers receive the same task messages through native chat templates and use deterministic greedy decoding

with reasoning disabled. Each completion is projected into the ontology and canonical JSON; failed projections are discarded. Eligible candidates are scored by

$$s(c, y) = 0.70F_1^{\text{loc+dis}}(c, y) + 0.25F_1^{\text{dis}}(c, y) + 0.05\mathbb{1}[c \text{ valid}]. \quad (1)$$

For an empty target set, we replace F1 by $1/(1 + m)$ for m predicted positives. The larger weight on Location+Disease F1 reflects the primary task: a disease assigned to the wrong anatomy should not be preferred. Because only successfully projected candidates are eligible, the validity term records schema compliance without changing their ordering. The highest- and lowest-scoring candidates form (c_i^+, c_i^-) when their structures differ, their margin is at least 0.10, and c_i^+ has higher $F_1^{\text{loc+dis}}$. The annotation determines ordering but never enters a teacher prompt or candidate.

Fixed-reference preference optimization. Round 0 performs one chosen-response SFT epoch. Central trains on the exact pair union, Local retains one root per site, and pair-weighted aggregation of those same local roots initializes FedPref. In DPO rounds $t = 1, 2, 3$, each policy continues from its preceding checkpoint while its round-0 root remains the fixed reference π_{ref}. Writing $h_\theta(c, x) = \log \pi_\theta(c \mid x) - \log \pi_{\text{ref}}(c \mid x)$, site k minimizes

$$\mathcal{L}_k(\theta) = -\mathbb{E}_{\mathcal{D}_k} \log \sigma\{\beta[h_\theta(c^+, x) - h_\theta(c^-, x)]\}. \quad (2)$$

The objective raises the chosen response's relative likelihood while the fixed SFT reference keeps that change comparable across rounds.

Aggregation and matched controls. At round t, every FedPref site receives the preceding global LoRA state $\phi^{(t-1)}$ and trains for one local epoch. With $n_k = |\mathcal{D}_k|$, the server computes

$$\phi^{(t)} = \sum_k \frac{n_k}{\sum_j n_j} \phi_k^{(t)}. \quad (3)$$

Pair weighting makes each site's influence proportional to its retained training evidence. Central trains one adapter for three epochs on $\cup_k \mathcal{D}_k$; Local trains six isolated adapters for three local epochs, matching pair exposure across arms. In FedPref, reports, labels, candidates, and preference pairs remain at their simulated sites; only LoRA states cross the boundary.

4 Experimental Setup

Task and cohorts. We align MIMIC-CXR reports with Chest ImaGenome annotations [6, 14]. Automatically derived labels provide 4,800 training and 600 development reports; the manually validated Chest ImaGenome gold standard provides 100 diagnostic and 400 locked-test reports. The schema contains nine diseases and 18 locations, 15 with positive support. All cohorts are patient-disjoint; gold-standard patients were excluded before constructing the automatically labeled cohorts. Code and reproduction instructions are available at https://github.com/flint-xf-fan/FedPref.

Simulated federation. We partition the data into six sites spanning 384 to 1,344 training reports and 297 to 717 retained pairs (Table 2). A capacity-constrained multilabel Dirichlet allocator ($\alpha = 0.5$) varies disease prevalence while preserving site sizes. Prevalence divergence exceeds a capacity-matched IID allocation: 8.83 versus 1.89 points for training pairs and 10.14 versus 4.14 for development (both $p = 0.0005$). The federation therefore combines unequal evidence with verified case-mix heterogeneity.

Preference data and optimization. The teachers produce 19,200 training and 2,400 development responses, yielding 2,961 training and 376 development pairs. All arms adapt Qwen3-8B with LoRA rank 16 and $\alpha = 32$. SFT uses learning rate 2×10^{-4}; DPO uses 5×10^{-6}, $\beta = 0.1$, BF16, cosine scheduling, and length 2,048. Each of three seeds independently trains the SFT root and all three DPO rounds, with three DPO exposures per pair in every arm.

Evaluation. Predictions are canonically projected before scoring; projection success records whether they contain a valid projectable JSON object. The primary endpoint is micro-F1 over positive disease-location labels (Location+Disease F1), which requires both finding and anatomy to be correct. Disease F1, projection success, false positives on reports with no targets, and held-out preference accuracy are secondary. Global F1 pools reports and weights larger sites more; client mean weights sites equally; worst-site F1 records the minimum. Central and FedPref use one shared policy, while home-routed Local sends each site's reports to its adapter. A paired patient-cluster bootstrap computes mean and worst-site effects over 10,000 draws and averages them across three seeds. Before gold-standard access, FedPref client-mean F1 selects one common round, breaking ties by worst-site F1 and then earlier round. Selected Central and FedPref policies are evaluated once on the locked gold-standard annotations.

5 Results

Cross-model candidates provide preference contrast. We measure candidate contrast by oracle headroom, the F1 gain from selecting the best candidate per report over the target baseline, and usable-pair coverage, the fraction of reports satisfying the pair margin. Across 100 automatically labeled reports, repeated Qwen3-8B sampling produced 71 identical parsed sets, 3.05 points of headroom, and 20.0% coverage. A separate five-model diagnostic (four teachers plus the target) on 100 gold-standard reports produced 14.30 points and 76.0%; its oracle exceeded the best individual model by 6.21 points (95% CI, 3.70 to 9.01). The first diagnostic identifies target-model collapse; the second establishes cross-model contrast and motivates excluding the target from the operational four-teacher pool.

Federated aggregation improves isolated sites. FedPref improves across communication rounds (Fig. 2). Its client-mean Location+Disease F1 increases from

Table 1. Three-seed development performance at selected round 3. Values are mean ± sample standard deviation; paired effects are reported in text with patient-cluster confidence intervals.

System	Global $F_1^{\text{loc+dis}}$	Mean $F_1^{\text{loc+dis}}$	Worst $F_1^{\text{loc+dis}}$	Pref. acc.
Central	71.77 ± 0.67	71.77 ± 0.65	70.03 ± 1.15	94.95 ± 0.27
FedPref	69.04 ± 0.22	69.11 ± 0.17	68.12 ± 0.22	91.49 ± 0.27
Local	67.32 ± 0.28	66.62 ± 0.32	59.02 ± 1.45	91.67 ± 0.85

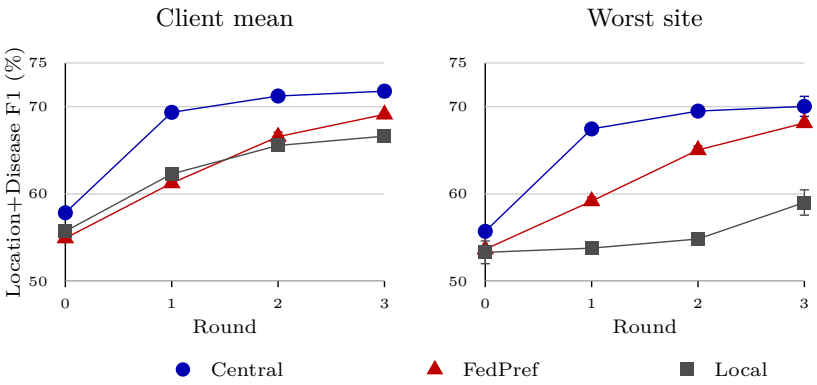

Fig. 2. Performance across communication rounds under matched preference-pair exposure. Error bars show sample standard deviation over three seeds. FedPref overtakes home-routed Local on client mean at round 2 and retains a nine-point worst-site advantage at the selected third round.

54.93 ± 0.47 to 69.11 ± 0.17, while worst-site F1 increases from 53.67 ± 0.18 to 68.12 ± 0.22. It overtakes Local on client mean at round 2, and round 3 maximizes the prespecified FedPref client-mean selection criterion.

At the selected round (Table 1), FedPref reaches 69.04 ± 0.22 global, 69.11 ± 0.17 client-mean, and 68.12 ± 0.22 worst-site F1. Home-routed Local reaches 66.62 ± 0.32 and 59.02 ± 1.45 on client mean and worst site. FedPref therefore improves the client mean by +2.49 points (95% CI, +1.24 to +3.79) and the worst site by +9.10 points (95% CI, +1.01 to +11.87).

Central reaches 71.77 ± 0.67, 71.77 ± 0.65, and 70.03 ± 1.15; Central exceeds FedPref by 2.66 points on client mean. FedPref and Local attain similar held-out preference accuracy, 91.49 ± 0.27 and 91.67 ± 0.85, despite their different extraction F1. Preference accuracy therefore does not track the extraction gap between the two systems. FedPref achieves $99.89 \pm 0.10\%$ projection success and has a $3.42 \pm 0.00\%$ false-positive rate on reports with no target findings.

Benefits concentrate at smaller sites. FedPref improves sites S1, S3, S4, and S5 (Table 2). The largest gains, +9.46 and +4.47 points, occur at S4 and S5, the

two smallest training cohorts; S0 and S2 favor Local. In this partition, gains concentrate at sites with less local evidence.

Table 2. Per-site support (training reports / retained pairs) and selected-round FedPref–Local effect.

Site	S0	S1	S2	S3	S4	S5
Support	1,344/717	1,056/591	864/556	672/486	480/297	384/314
$\Delta F_1^{\text{loc+dis}}$ (pt)	−1.87	+0.87	−1.72	+3.73	+9.46	+4.47

Locked gold-standard generalization. On the 400-report locked gold test, Central records Location+Disease F1 of 71.67±0.23, Disease F1 of 90.12±0.17, projection success of 99.58±0.29%, and a false-positive rate of 7.58±0.00%. FedPref records 68.68 ± 0.35, 86.36 ± 0.16, $100.00 \pm 0.00\%$, and $8.33 \pm 0.76\%$, respectively. The primary-endpoint comparison yields +2.98 points for Central minus FedPref (95% patient-cluster CI, +1.53 to +4.54). Scores are mean ± sample standard deviation over three seeds, and the interval uses 10,000 patient-cluster resamples conditional on those seeds. Outputs from both policies project successfully for nearly every report and generalize to the gold-standard labels while preserving the Central–FedPref extraction ordering; the home-routed development results above provide the site-level evidence.

6 Discussion and Conclusion

FedPref reframes structured clinical annotations as a local preference interface. Each institution uses its existing annotations to rank candidate structures under a shared extraction rule, keeping reports and task-specific judgments under local control. Cross-model candidates supply local supervision that adapter aggregation carries into a shared model. This separation accommodates institutions that share a target schema but differ in data volume, case mix, and governance, and suggests a general strategy for federated adaptation of schema-constrained language models. The boundaries of the present evidence and the requirements for clinical deployment are detailed in Supplementary Sec. A.1.

In this controlled federation, aggregating adapters trained from local preference signals improves extraction most at the sites with the least evidence. Central training's remaining advantage measures the observed gap under matched exposure, and a locked gold test preserves the same ordering. FedPref thus offers a practical middle ground between isolated and pooled training for structured radiology extraction.

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

References

1. Dai, Z., Fan, F.X., Tan, C., Hoang, T.N., Low, B.K.H., Jaillet, P.: Federated sequential decision making: Bayesian optimization, reinforcement learning, and beyond. In: Federated Learning, pp. 257–279. Elsevier (2024)
2. Fan, F.X., Tan, C., Ong, Y.S., Wattenhofer, R., Ooi, W.T.: FedRLHF: A convergence-guaranteed federated framework for privacy-preserving and personalized RLHF. In: Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS) (2025)
3. Fan, F.X., Tan, C., Wattenhofer, R., Ong, Y.S.: Position paper: Rethinking privacy in rl for sequential decision-making in the age of llms. In: 2025 International Joint Conference on Neural Networks (IJCNN). pp. 1–8. IEEE (2025)
4. Fan, X., Ma, Y., Dai, Z., Jing, W., Tan, C., Low, B.K.H.: Fault-tolerant federated reinforcement learning with theoretical guarantee. In: Advances in Neural Information Processing Systems. pp. 1007–1021 (2021)
5. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022)
6. Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G., Horng, S.: MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data (2019)
7. Li, T., Sanjabi, M., Beirami, A., Smith, V.: Fair resource allocation in federated learning. In: International Conference on Learning Representations (2020)
8. Ma, T., Luo, X., Tan, R., Gao, H.: Privacy and fairness-guaranteed federated preference optimization for large language models in internet of medical things. IEEE Transactions on Consumer Electronics (2025)
9. McMahan, B., Moore, E., Ramage, D., Hampson, S., Agüera y Arcas, B.: Communication-efficient learning of deep networks from decentralized data. In: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 54, pp. 1273–1282. PMLR (2017)
10. Rafailov, R., Sharma, A., Mitchell, E., Manning, C.D., Ermon, S., Finn, C.: Direct preference optimization: Your language model is secretly a reward model. In: Advances in neural information processing systems (2023)
11. Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H.R., Albarqouni, S., Bakas, S., Galtier, M.N., Landman, B.A., Maier-Hein, K., et al.: The future of digital health with federated learning. NPJ digital medicine **3**(1), 119 (2020)
12. Sabour, A., Chu, K., Dehaghani, M.E., Moradi, M.: Mind the evaluation gap: Large language models for structured data extraction from radiology reports. In: International Workshop on Emerging LLM/LMM Applications in Medical Imaging. pp. 19–27. Springer (2025)
13. Singhal, R., Ponkshe, K., Vepakomma, P.: FedEx-LoRA: Exact aggregation for federated and efficient fine-tuning of foundation models. arXiv preprint arXiv:2410.09432 (2024)
14. Wu, J.T., Agu, N.N., Lourentzou, I., Sharma, A., Paguio, J.A., Yao, J.S., Dee, E.C., Mitchell, W., Kashyap, S., Giovannini, A., Celi, L.A., Moradi, M.: Chest ImaGenome dataset for clinical reasoning. In: NeurIPS DB Track (2021)
15. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)