Privacy-Preserving LLM Alignment via Private Reward Modeling: A Holistic and Data-Efficient Framework

Young Hyun Cho[1], Will Wei Sun[2]

[1]*Department of Statistics, Purdue University*

[2]*Mitch Daniels School of Business, Purdue University*

ABSTRACT

Adapting Large Language Models (LLMs) to specific domains via preference alignment is essential for capturing nuanced human expectations, yet it introduces significant privacy risks when sensitive data is involved. While Differential Privacy (DP) is the standard solution, applying it to the preference fine-tuning stage presents a critical dilemma. Standard multi-stage pipelines necessitate inefficient data partitioning that limits utility, whereas applying DP directly to methods like Direct Preference Optimization (DPO) suffers from severe gradient instability. Furthermore, common Label-DP approaches fail to protect user prompts, leaving sensitive interaction details exposed.

In this work, we introduce a framework that decouples the privacy mechanism from policy optimization to achieve stability and holistic, Tuple-level DP. Our approach learns a Private Reward Model using the entire dataset—resolving partitioning inefficiencies—and derives the final policy via a deterministic post-processing step, thereby circumventing unstable gradient updates. This guarantees DP for the entire interaction tuple (prompt, response, and label), offering significantly stronger protection than Label-DP. Theoretically, we establish sample complexity bounds matching the non-private rate up to an additive privacy cost. Empirically, our method significantly outper- forms private DPO and PPO baselines on the Gemma-2b-it task, highlighting improved stability and privacy–utility trade-offs relative to contemporary baselines.

Keywords: Large Language Models, Reinforcement Learning with Human Feedback, Differential Privacy, Reward Modeling, Sample Complexity

Bridging Spatial Transcriptomics and Histopathology through AI

Wei Chen[1,*], Chongyue Zhao[1], Tianhao Liu[1]

(Author order follows that of the original publication)

Department of Pediatrics, University of Pittsburgh,

UPMC Children's Hospital of Pittsburgh, Pittsburgh, PA, USA

ABSTRACT

Spatial transcriptomics (ST) technologies from multicellular platforms such as Visium and Curio to subcellular platforms including Visium HD and Stereo-seq remain limited in resolving gene expression at true single-cell resolution. Existing computational methods operate largely at the spot level and cannot accurately reconstruct full transcriptomes for individual cells across ST resolutions. We introduce a unified framework that integrates high-resolution histology images with spot-level ST data using a Vision Transformer model and contrastive learning. It reconstructs single-cell gene expression profiles from both multicellular and subcellular platforms. Using in-house mouse lung datasets generated on Visium, Visium HD, and Stereo-seq, along with public datasets, our method consistently improves biological signal recovery: (1) at the tissue level, it delineates structural domains and immune regions (2) at the cellular level, it identifies major immune populations, resolves CAF and macrophage subtypes including rare $SPP1^{+}$ macrophages, and detects tertiary lymphoid structures in colorectal cancer, and (3) at the molecular level, it enhances cell-type separation and differential expression accuracy. Beyond ST reconstruction, our method further extends to histology-only cell annotation, leveraging learned multimodal representations to classify individual cells on routine H&E images, including those from complex disease tissues. In addition, a user-friendly interactive interface enables real-time visualization, expert refinement, and scalable annotation, supporting broad translational applications in spatial genomics and digital pathology.

Keywords: Spatial Transcriptomics; Vision Transformer; Histopathology; scRNA-seq.

Fair Graph Learning Without Complete Demographics

Zichong Wang[1], **Fang Liu[*2]**, Shimei Pan[3], Jun Liu[4], Fahad Saeed[1], Meikang Qiu[5], Wenbin Zhang[1]

[1]*Florida International University, FL, USA*

[2]*University of Notre Dame, IN, USA*

[3]*University of Maryland Baltimore County, MD, USA*

[4]*Northeastern University, MA, USA*

[5]*Augusta University, GA, USA*

ABSTRACT

Graph Neural Networks (GNNs) have excelled in diverse applications due to their outstanding predictive performance, yet they often overlook fairness considerations, prompting numerous recent efforts to address this societal concern. However, most fair GNNs assume complete demographics by design, which is impractical in most real-world socially sensitive applications due to privacy, legal, or regulatory restrictions. For example, the Consumer Financial Protection Bureau mandates that creditors ensure fairness without requesting or collecting information about an applicant's race, religion, nationality, sex, or other demographics. We propose fairGNN-WOD, a first-of-its-kind framework that considers mitigating unfairness in graph learning without using demographic information. We analyze bias in node representations and establish the relationship between utility and fairness objectives. Experiments on three real-world graph datasets illustrate that fairGNN-WOD outperforms state-of-the-art baselines in achieving fairness but also maintains comparable prediction performance.

Keywords: AI Ethics, Trust, Fairness, Graph Neural Networks