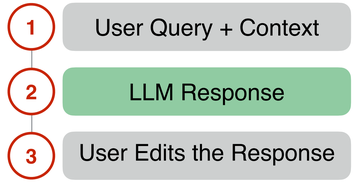
Principled Fine-tuning of LLMs from User-Edits: A Medley of Preference, Supervision, and Reward

Dipendra Misra*, Aldo Pacchiano*, Ta-Chung Chi, and Ge Gao

Post-Training LLMs with User-Edits Deployment Data

- Post-training LLMs on an in-distribution labeled data can lead to improvement in the performance.
- **Challenge:** where to get the training data for post-training?
  - Get annotations from third-party. **Cons:** expensive, not in-distribution, need to ensure diversity.
  - What if we could use deployment data? **Pros:** in-distribution, abundant.
- We focus on **user-edits** — a common type of deployment data:
  - Common in coding and writing AI assistants.
  - Provides naturally occurring user feedback for improving LLMs!

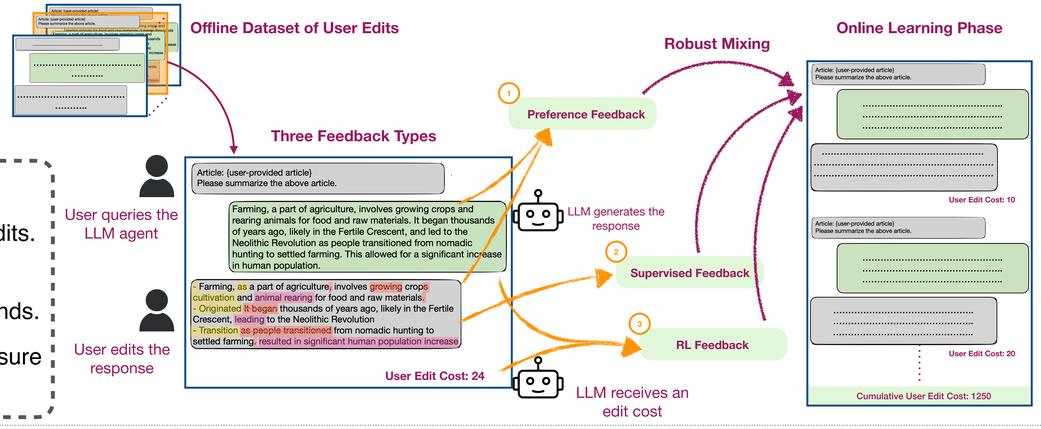


Principled Learning from User-Edits

- Initiate principled algorithm development in learning from user-edits setting.
- Rich algorithm design landscape: three different learning feedback type available!

Learning Setup

- **Offline Learning:** Learn from a dataset of generations and user-edits.
- **Online Learning:** Evaluate with a user over a small number of T rounds.
- **Edit Distance Cost:** Used to measure performance (lower → better).



Theoretical Results & Main Algorithm

Theoretical Results: Derive bounds on sub-optimality (SubOpt) of algorithms that use the three individual feedback types:

1. SFT (labeled)
2. DPO (preference data)
3. Offline RL (cost)

Informal Theorem: Sub-optimality of DPO policy $\hat{\pi}_{\mathrm{DPO}}$ on user-edits preference data satisfies:

$$\texttt{SubOpt}(\hat{\pi}_{\mathrm{DPO}}) \leq \epsilon, \qquad \text{if } n \geq \Omega\left(\frac{C_{PREF}\log\frac{|\Pi|}{\delta}}{\beta^2\sigma'(-V_{\max})^2\epsilon^2}\right)$$

- DPO is more affected by coverage while SFT by quality of feedback.

Main Algorithm: Combines different user feedback types in two ways:

- Early Ensemble: train a policy to optimize the loss of different types jointly.
- Late Ensemble: run a bandit algorithm in the online phase on the set of policies trained with various feedback types.

Results

- Evaluate on two domains from Gao et al. 2024, using simulated LLM user.
- A weak user converges more slowly $\pi^\star$ to than a strong user.

Method	Summarization		Email Writing		
	Strong User	Weak User	Strong User	Weak User	Max SubOpt
Base	$0.9455_{\pm 0.01}$	$0.9445_{\pm 0.02}$	$0.5108_{\pm 0.03}$	$0.4923_{\pm 0.01}$	$0.7364_{\pm 0.10}$
SFT	$0.5377_{\pm 0.02}$	$0.9304_{\pm 0.19}$	$0.4159_{\pm 0.05}$	$0.4539_{\pm 0.03}$	$0.5772_{\pm 0.19}$
DPO	$1.0790_{\pm 0.06}$	$0.8267_{\pm 0.06}$	$\mathbf{0.3365}_{\pm 0.00}$	$\mathbf{0.3368}_{\pm 0.01}$	$0.8698_{\pm 0.11}$
EarlyEnsemble	$\mathbf{0.2092}_{\pm 0.09}$	$\mathbf{0.3586}_{\pm 0.01}$	$0.3438_{\pm 0.06}$	$0.4864_{\pm 0.01}$	$0.1612_{\pm 0.01}$
LateEnsemble	$0.2768_{\pm 0.13}$	$0.4403_{\pm 0.03}$	$0.4202_{\pm 0.11}$	$0.3739_{\pm 0.04}$	$\mathbf{0.1586}_{\pm 0.04}$