Centre for Frontier AI Research
A*STAR CFAR

FedRLHF: A Convergence-Guaranteed Federated Framework for **Privacy-Preserving** and **Personalized RLHF**

Flint Xiaofeng Fan, Cheston Tan, Yew-Soon Ong, Roger Wattenhofer, and Wei Tsang Ooi

AAMAS 2025

Flint Xiaofeng Fan
Postdoc, ETH Zurich
Scientist, A*STAR CFAR
A*STAR International Fellow

Background: RL from Human Feedback (RLHF)

- RL from Human Feedback

-> to align with human preferencc

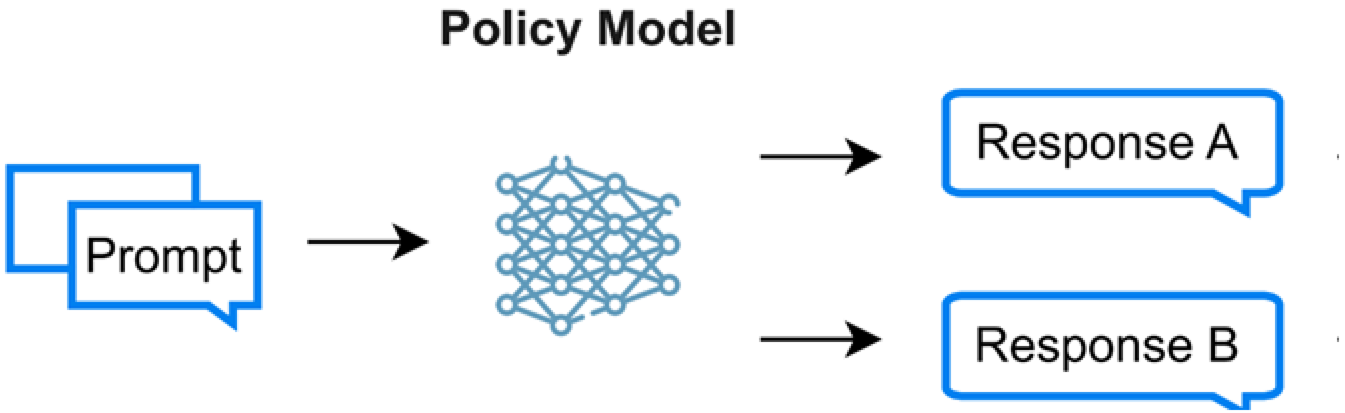
- Key Applications
  - Robotics
  - Recommendation
  - Language Models

-> e.g, ChatGPT

Background: RL from Human Feedback (RLHF)

- RL from Human Feedback -> to align with human preferencc

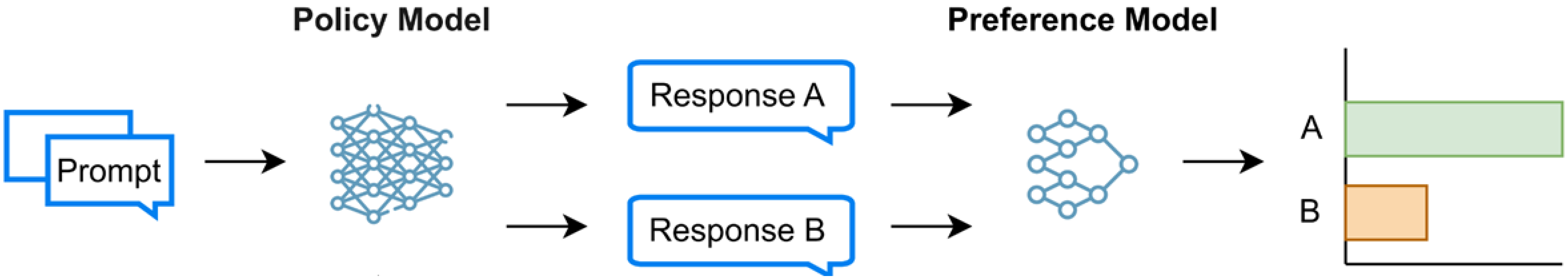
- Key Applications
  - Robotics
  - Recommendation
  - Language Models -> e.g, ChatGPT

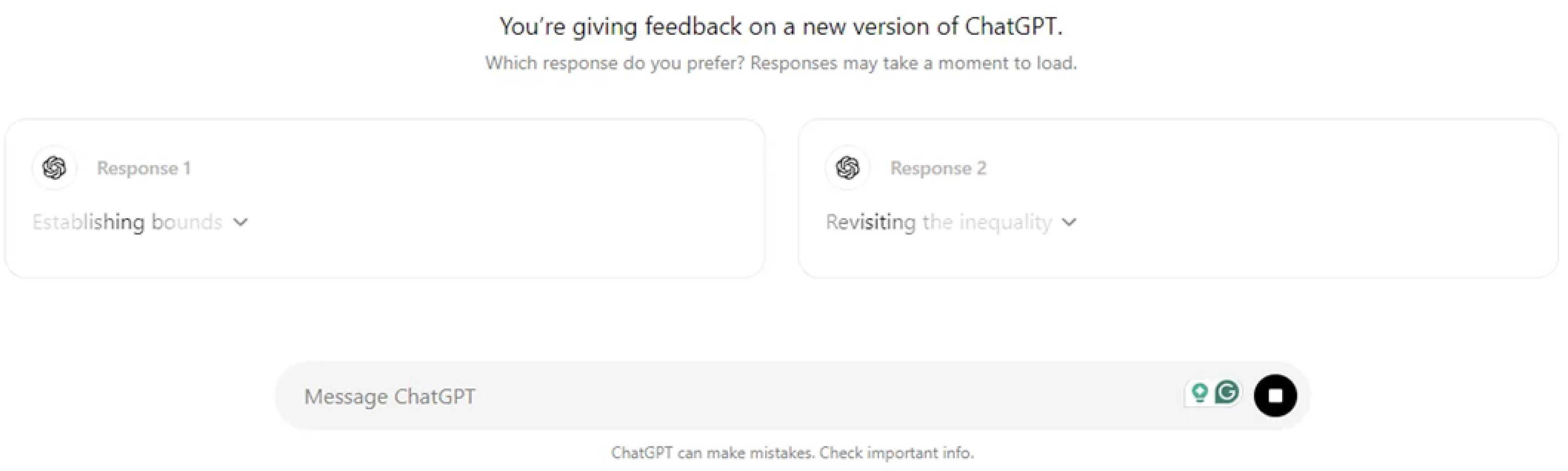


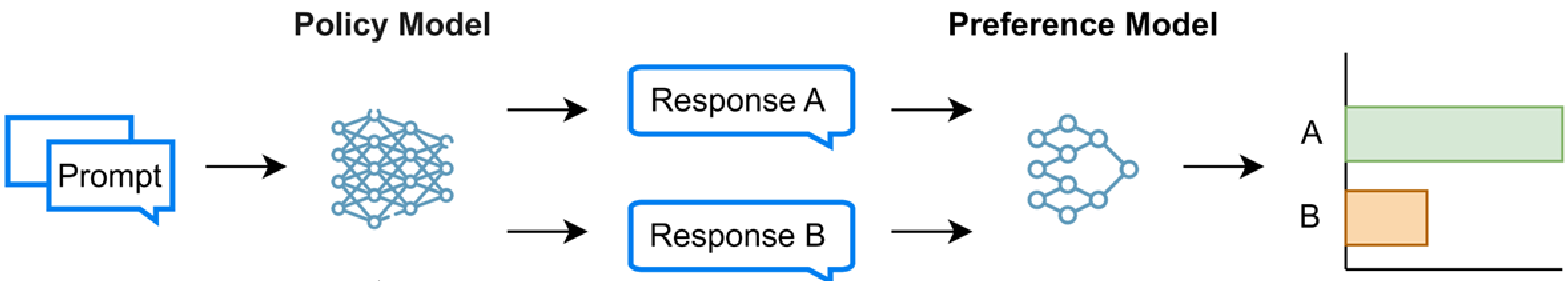
Background: RL from Human Feedback (RLHF)

- RL from Human Feedback -> to align with human preferencc

- Key Applications
  - Robotics
  - Recommendation
  - Language Models -> e.g, ChatGPT

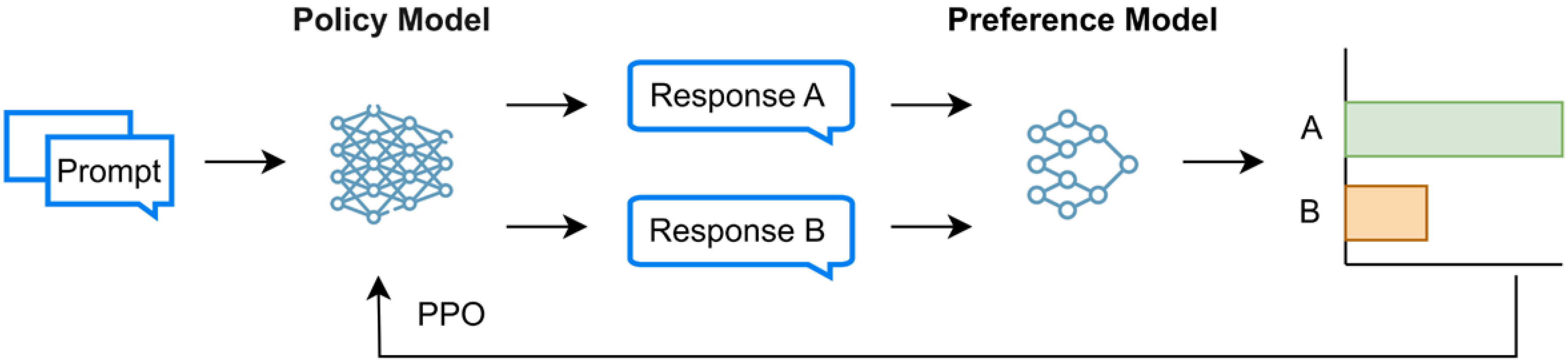



You're giving feedback on a new version of ChatGPT.
Which response do you prefer? Responses may take a moment to load.
Response 1
Establishing bounds
Response 2
Revisiting the inequality
Message ChatGPT
ChatGPT can make mistakes. Check important info.


Policy Model
Preference Model
Prompt
Response A
Response B
A
B

Background: RL from Human Feedback (RLHF)

- RL from Human Feedback -> to align with human preferencc
- Key Applications
  - Robotics
  - Recommendation
  - Language Models -> e.g, ChatGPT

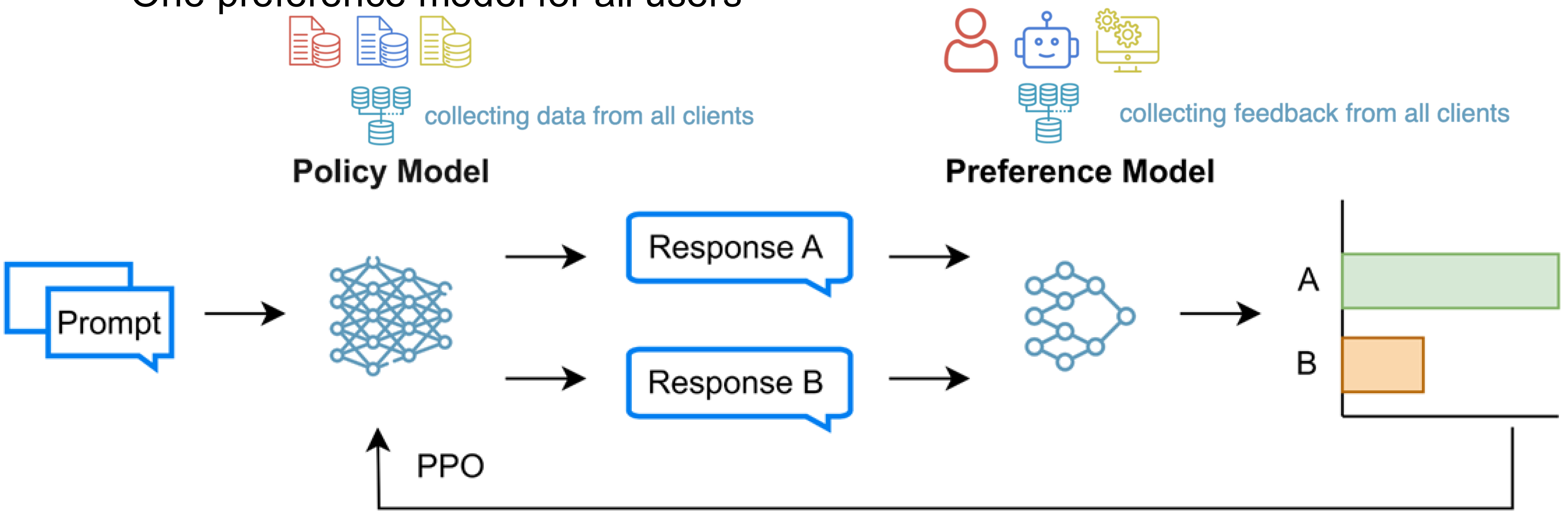


Challenge: Centralization

- Privacy

    - All user data and human preference are collected (HIPAA, GDPR, PDPA, etc.)
- Personalization

    - One preference model for all users



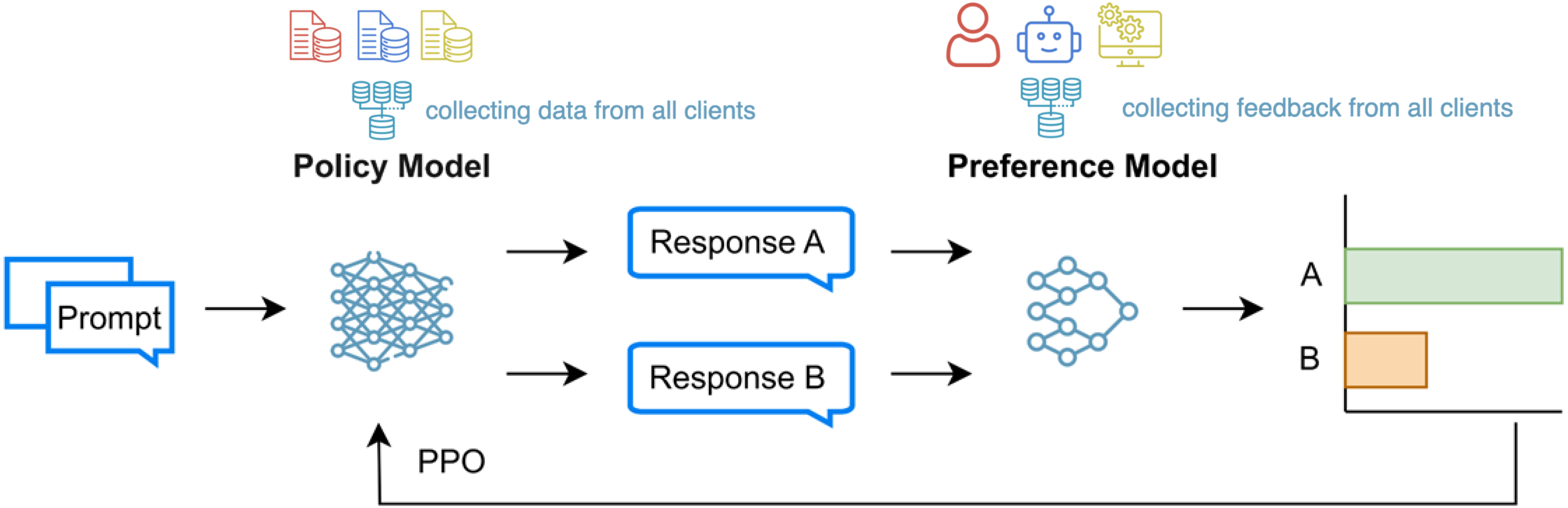
Challenge: Centralization

- Privacy

  - All user data and human preference are collected (HIPAA, GDPR, PDPA, etc.)

- Personalization

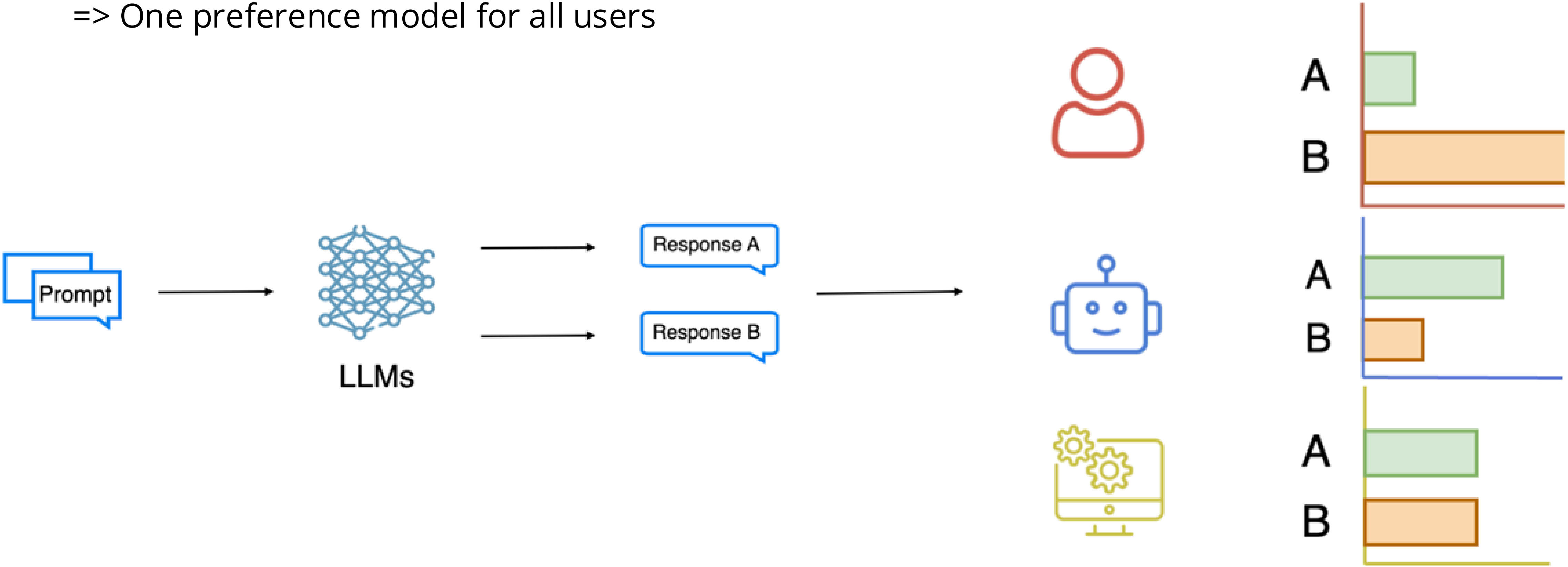
  - One preference model for all users



Challenge: Centralization

- Privacy
  => All user data and human preference are collected (HIPAA, GDPR, etc.)

- Personalization
  => One preference model for all users



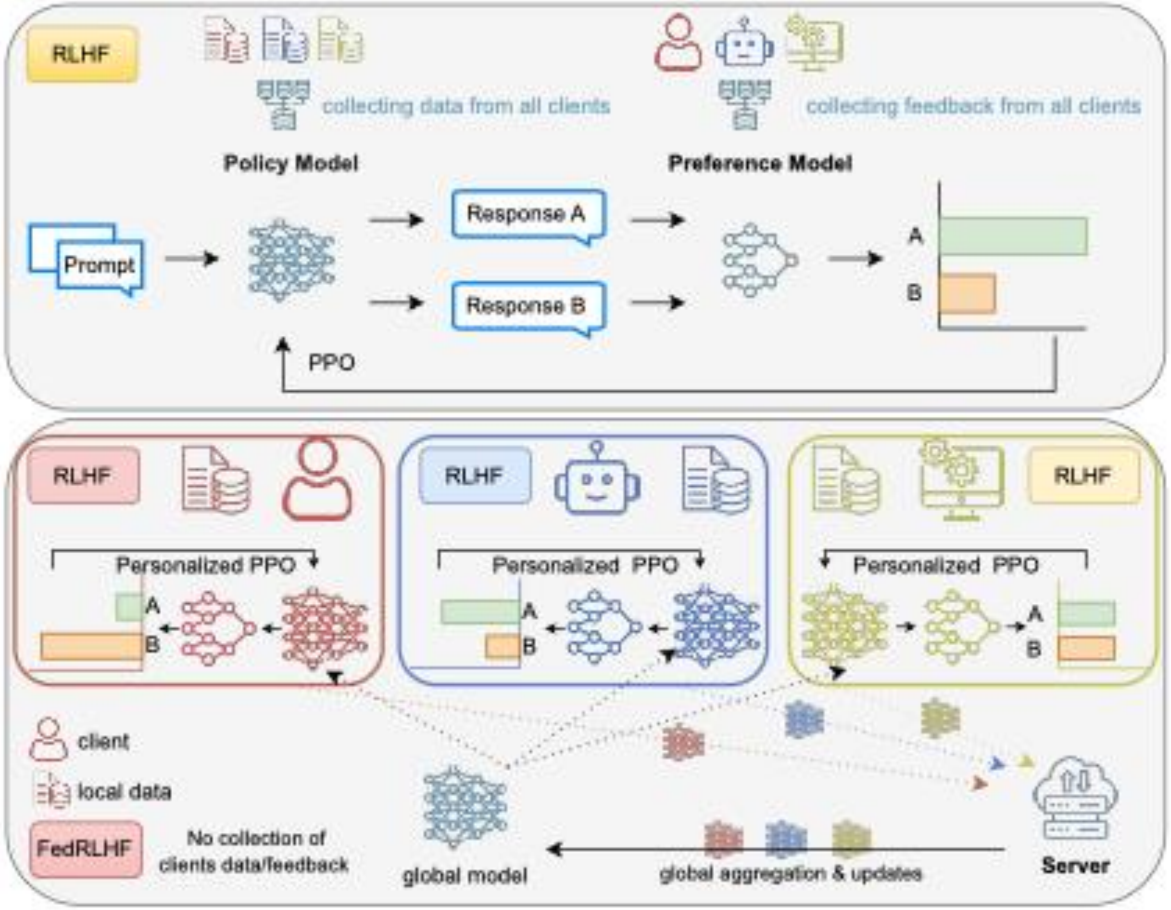
Solution: FedRLHF

RLHF

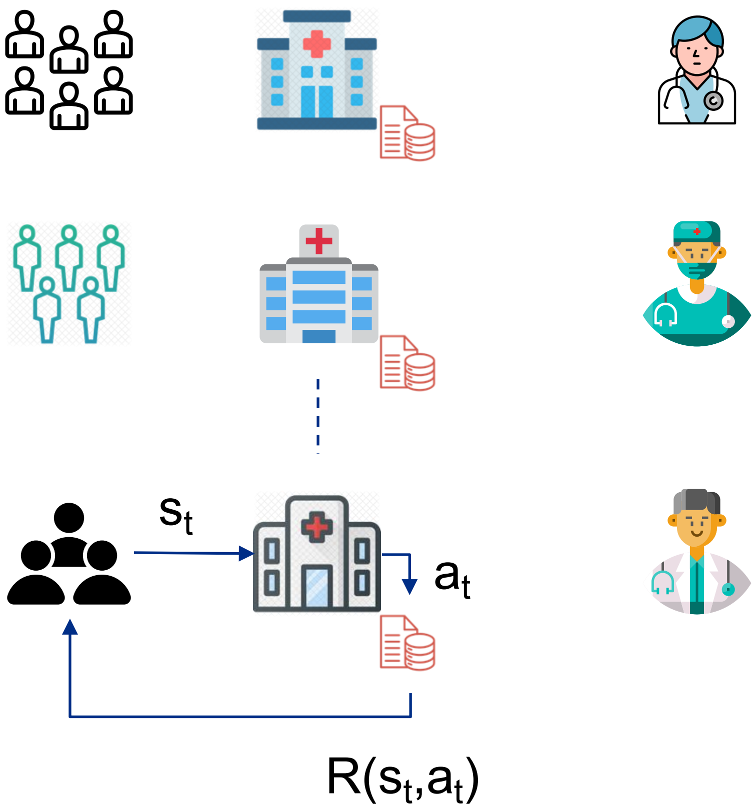
- ❌ Centralized data/feedback collection
- ❌ Single global preference model

FedRLHF

- ✅ Local training on local data
- ✅ Local feedback models
- ✅ Only model weights shared
- ✅ Personalized policies



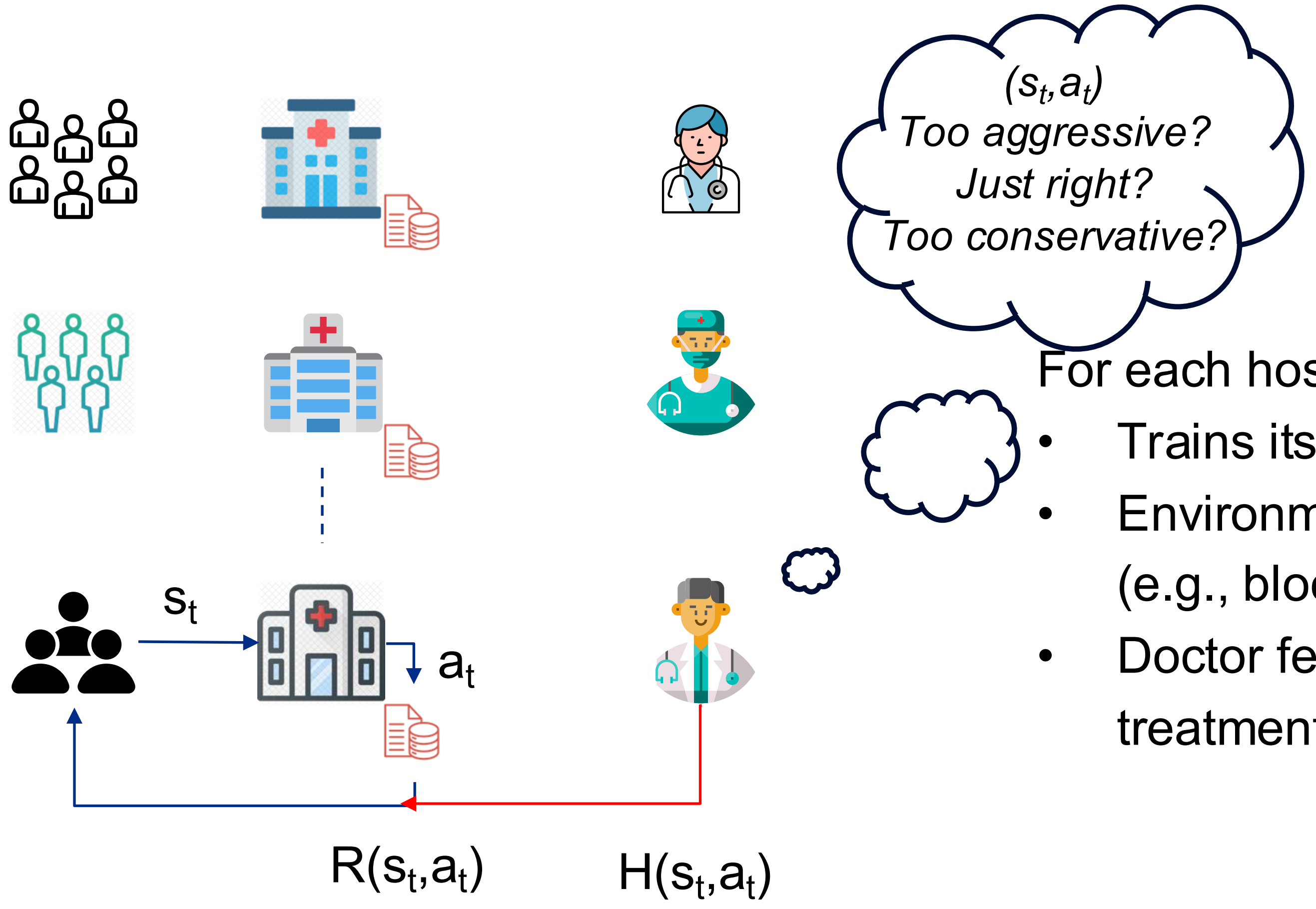
FedRLHF – Practical Example



For each hospital

- Trains its RL model locally using patient data
- Environmental rewards: Objective health metrics (e.g., blood sugar levels, recovery time)

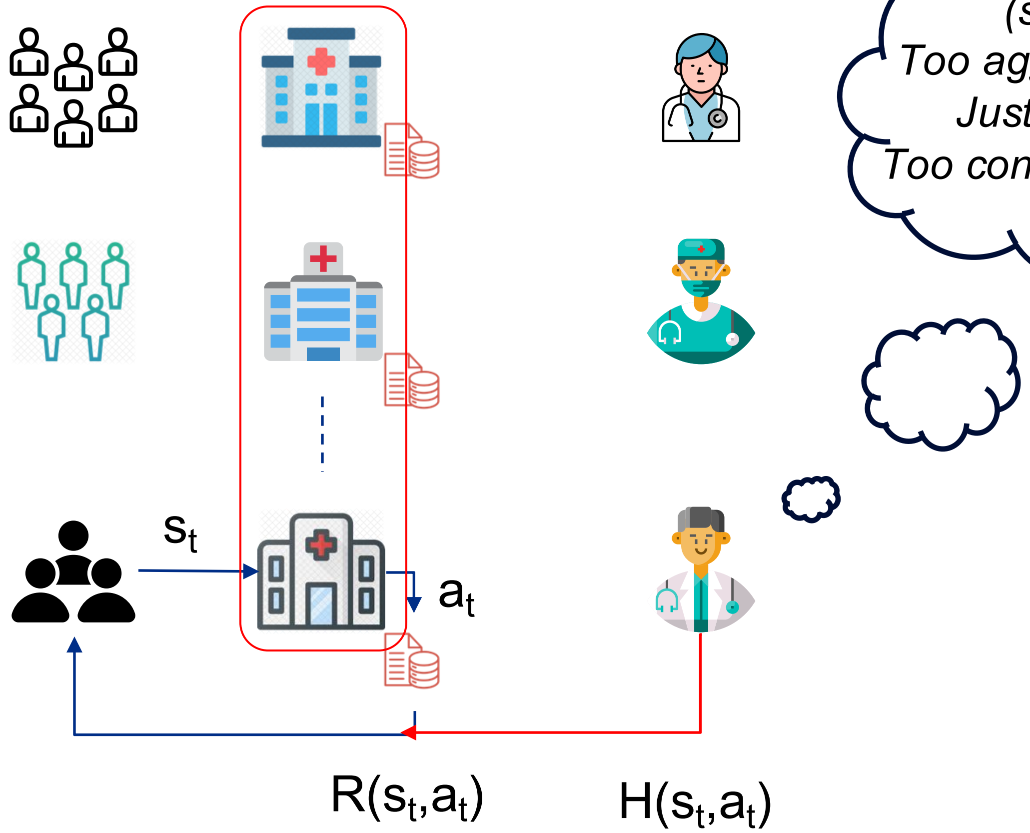
FedRLHF – Practical Example

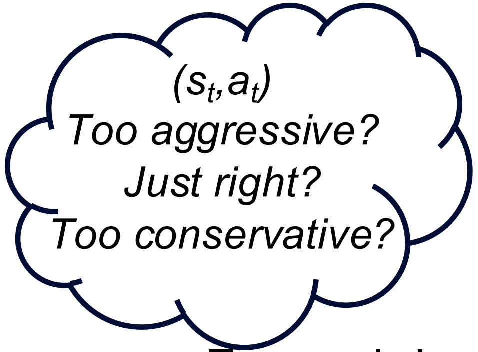


For each hospital

- Trains its RL model locally using patient data
- Environmental rewards: Objective health metrics (e.g., blood sugar levels, recovery time)
- Doctor feedback provides qualitative signals (e.g., treatment appropriateness)

FedRLHF – Practical Example





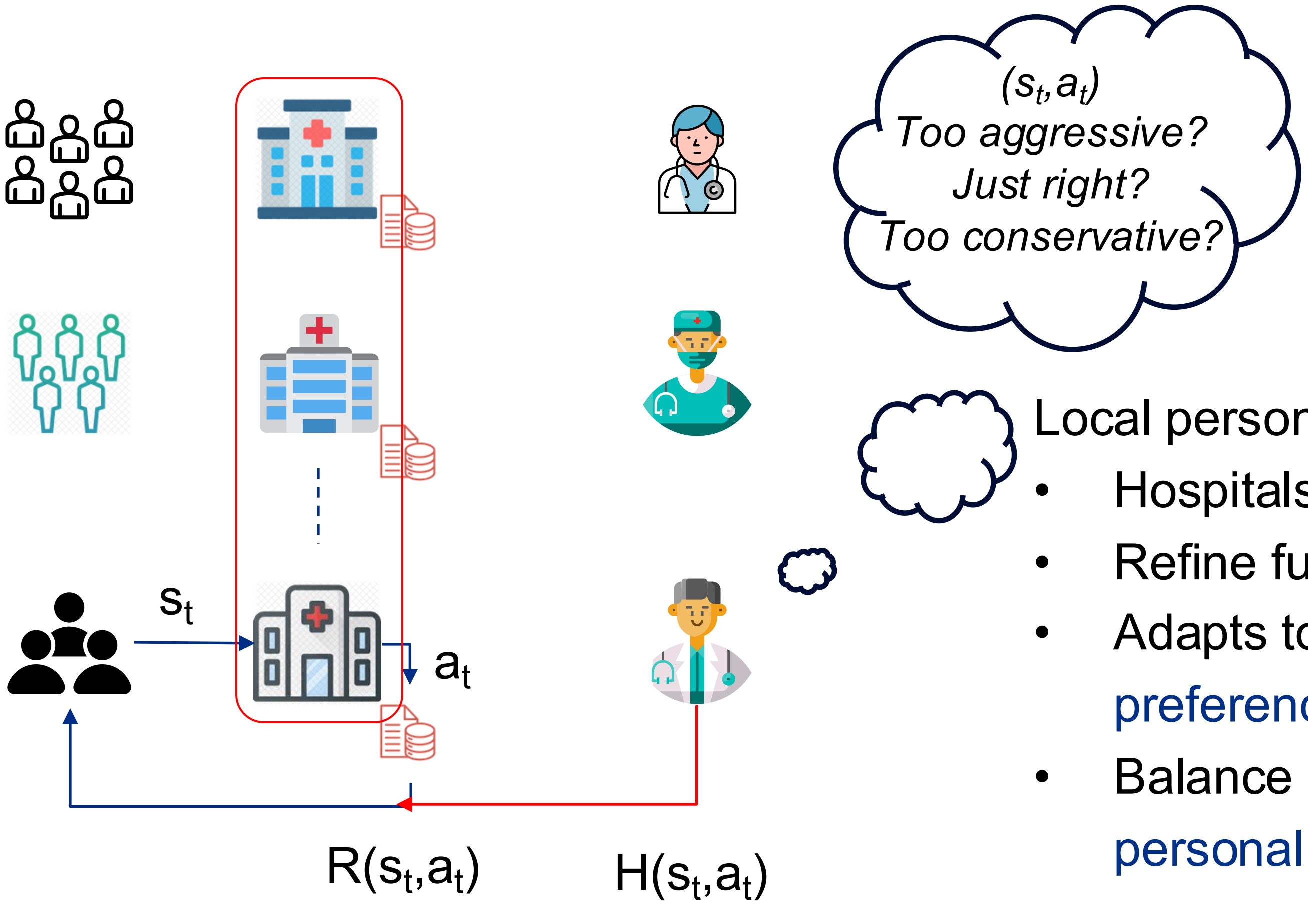
For each hospital

- Trains its RL model locally using patient data
- Environmental rewards: Objective health metrics (e.g., blood sugar levels, recovery time)
- Doctor feedback provides qualitative signals (e.g., treatment appropriateness)

A central party (central hospital, government, etc.)

- Aggregates model updates from all hospitals
- Captures shared knowledge

FedRLHF – Practical Example



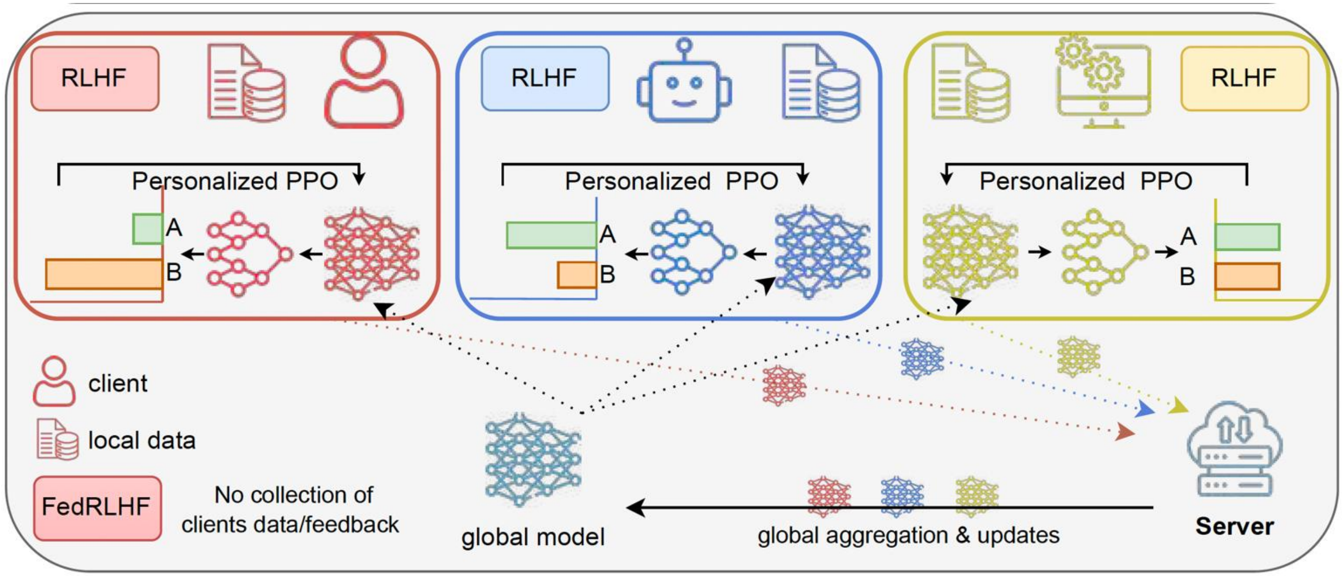
Local personalization

- Hospitals receive the global model and
- Refine further using local data and feedback
- Adapts to local patient demographics and doctor preferences
- Balance between global collaboration and highly personalized care

FedRLHF – Problem Formulation

System setup: K clients, each client $k \in 1, , 2, \ldots, K$

- Local MDP $M_k = (\mathcal{S}, \mathcal{A}, P_k, R_k^0, \rho_0(s), \gamma)$
- Local feedback $H_k(s, a)$
- Shaped reward $R_k(s, a) = R_k^0(s, a) + \lambda H_k(s, a),$
- Local objective $J_k(\theta) = \mathbb{E}_{\tau \sim \pi_\theta}[\sum_{t=0}^{\infty} \gamma^t R_k(s_t, a_t)]$



FedRLHF – Problem Formulation

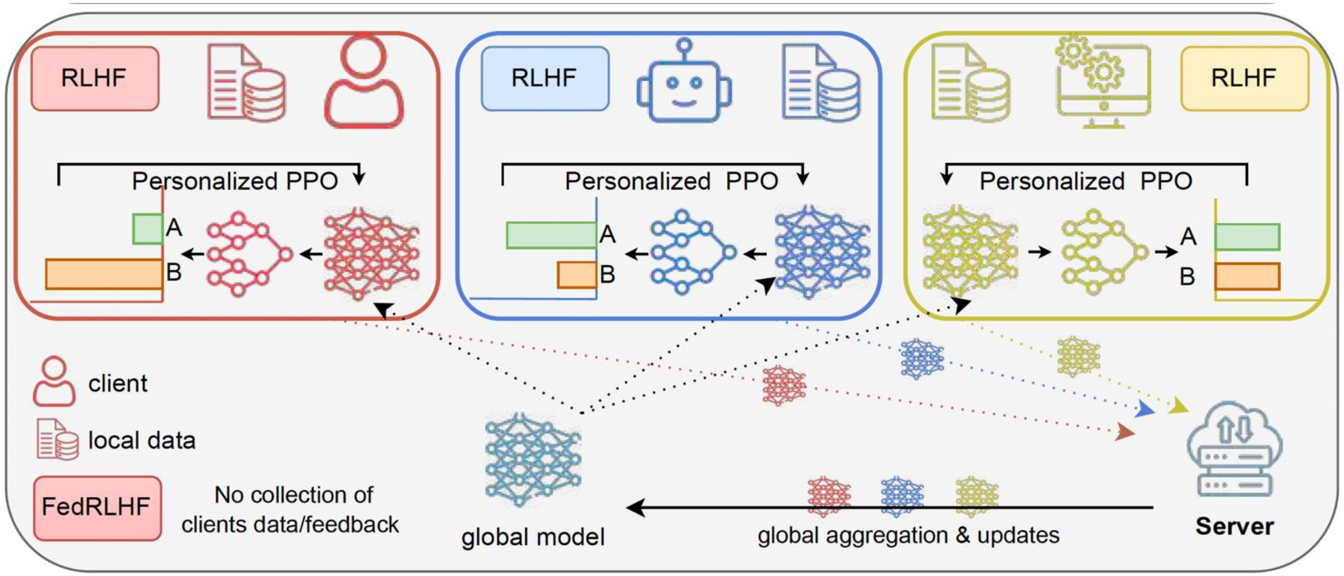
System setup: K clients, each client $k \in 1, , 2, \dots, K$

- Local MDP $M_k = (\mathcal{S}, \mathcal{A}, P_k, R_k^0, \rho_0(s), \gamma)$
- Local feedback $H_k(s, a)$
- Shaped reward $R_k(s, a) = R_k^0(s, a) + \lambda H_k(s, a),$
- Local objective $J_k(\theta) = \mathbb{E}_{\tau \sim \pi_\theta}[\sum_{t=0}^{\infty} \gamma^t R_k(s_t, a_t)]$

Server-clients communication FedAvg

- Server broadcasts global model
- Clients send parameters

Global objective $$J(\theta) = \frac{1}{K} \sum_{k=1}^{K} J_k(\theta)$$



FedRLHF - Convergence

Assumption 6 (Bounded Human Feedback). *For all $s \in \mathcal{S}$, $a \in \mathcal{A}$, and $k \in [K]$, the human feedback is bounded:*

$$|H_k(s,a)| \leq H_{\max}.$$

Remark. *Assumption 6 limits the variance introduced by human feedback in the learning process. In our experiments with the MovieLens task, we implement this by bounding feedback values and options (Section 6.1.2), similar to practical systems like ChatGPT that curate feedback for consistency.*

Theorem 4.1 (Convergence of FedRLHF). The output of Algorithm FedRLHF satisfies:

$$\mathbb{E}[J(\theta^*) - J(\theta_{\text{avg}})] \leq \boxed{\frac{L}{\mu T}(J(\theta^*) - J(\theta_0))} + \boxed{\frac{1}{2\mu K}(G^2 + \sigma^2)} + \boxed{\frac{L}{\mu}\lambda H_{\max}.}$$

O(1/T):
linear convergence rate
w.r.t. T

O(1/K):
scalability
w.r.t. K

O(1):
bounded impact
from feedback

FedRLHF - Sample Complexity

Theorem 4.2 (Sample Complexity of FedRLHF). To achieve an expected optimality gap of $\mathbb{E}[J(\theta^*) - J(\theta_{\text{avg}})] \leq \epsilon$, the total number of samples required across all clients is:

$$N = O\left(\frac{\boxed{L}(G^2 + \sigma^2)}{\boxed{\mu}^2\epsilon^2}\right)$$

Task-dependent properties

resulting in per-client sample complexity:

$$N_c = \frac{N}{K} = O\left(\frac{L\boxed{(G^2 + \sigma^2)}}{\boxed{K}\mu^2\epsilon^2}\right)$$

combined effect of gradient bound and variance

decreases proportionally with K

FedRLHF - Personalization vs Performance

A fundamental trade-off in collaborative learning:

Global Learning		Local Adaptation
- Shared knowledge	vs	- Personal preferences
- Collaborative gain		- Client-specific behavior
- Model consistency		- Divergent policies

Definition 5.2 (Personalization Score).

$$P_k(\theta) = \mathbb{E}_{s\sim\rho}[D_{\mathrm{KL}}(\pi_k(\cdot|s,\theta) \parallel \pi(\cdot|s,\theta))]$$

Definition 5.3 (Global Performance Metric).

$$J_g(\theta) = \frac{1}{K}\sum_{k=1}^{K} J_k^0(\pi)$$

FedRLHF - Personalization vs Performance

Theorem 5.1 (Personalization-Performance Trade-off). Using the FedRLHF algorithm, the global performance metric (Definition 5.3) satisfies:

$$J_g(\theta) \geq \frac{1}{K}\sum_{k=1}^{K} J_k^0(\pi_k) - C \cdot \left(\frac{1}{K}\sum_{k=1}^{K}\sqrt{P_k(\theta)}\right)$$

Theorem 5.2 (Impact of Human Feedback):

a) The average personalization scores increases at $O(\lambda^2)$
b) The global performance decreases at $O(\lambda)$
c) The system sample complexity increases at $O(\lambda)$

FedRLHF - Personalization vs Performance

Takeaway:

1. Trade-off Highlight: Improving personalization comes at the cost of global performance.

2. $$R_k(s,a) = R_k^0(s,a) + \lambda H_k(s,a),$$

λ	↑	Local Adaptation	↑	Global Consistency	↓
λ	↓	Local Adaptation	↓	Global Consistency	↑

Experiment: Sentiment-Controlled Text Generation

Dataset & Task:

- 50,000 labelled reviews from IMDb dataset (positive/negative sentiment)
- Partitioned among K=5 clients, each with ~10k reviews
- Objective: Fine-tune a GPT-like model to generate text that matches a *desired sentiment style*. Different clients want different positivity levels.

Setup:

- Transformer RL (TRL)[2] library from hugging face
- Flower API[3] for simulating realistic distributed network, governed by a server

[2]https://github.com/huggingface/trl/tree/main
[3]https://flower.ai/docs/framework/index.html
Please refer to our paper for more details and results.

Implementation & Local Feedback Mechanism

Model *pretrained from HuggingFace*

- GPT-2 (125M parameters + PPO overhead for RL)

Intrinsic rewards $R_k^0 \in [0, 1]$

- Based on log-likelihood of coherent text (fluency)

Feedback/Sentiment $R_{\text{sentiment}} \in [0, 1]$

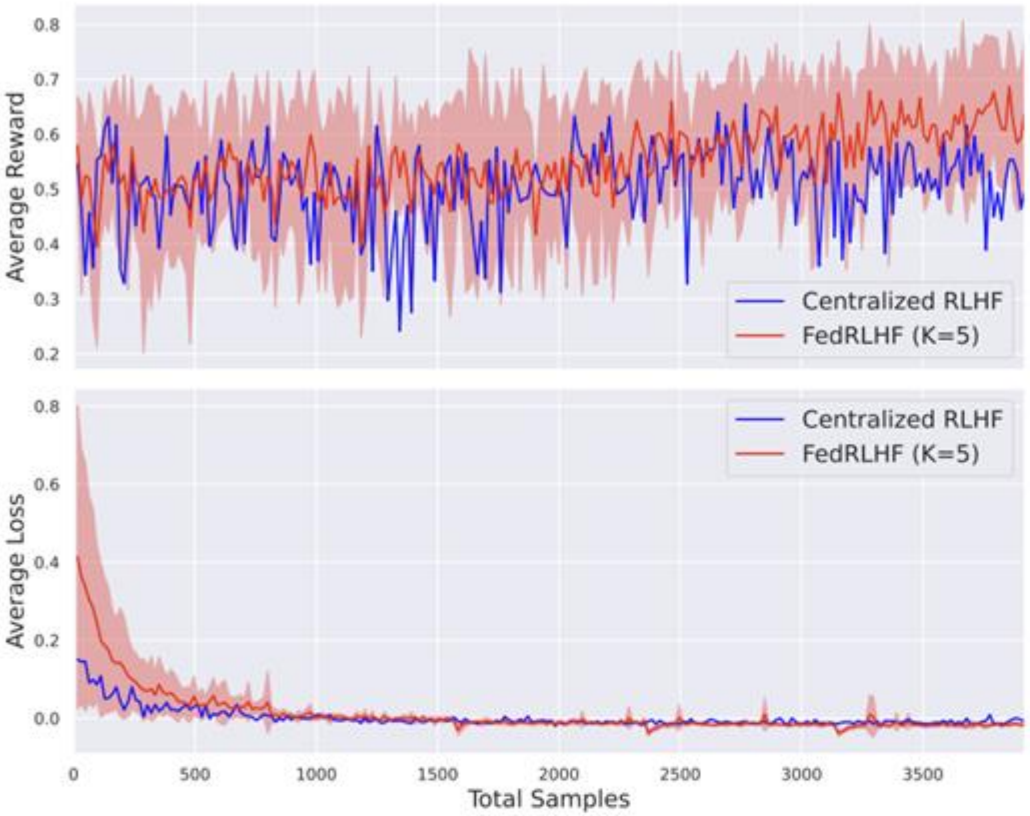
- Outputs from pretrained DistBERT sentiment classifier

Combined reward $R_k = \lambda_k \cdot R_{\text{sentiment}} + (1 - \lambda_k) \cdot R_k^0$

Results – FedRLHF vs Centralized RLHF

- FedRLHF catches up in average reward, eventually surpassing centralized RLHF after ~1500–2000 samples.
- FedRLHF starts with higher loss in early rounds, due to variance among clients, but quickly drops due to the exploration of more clients.

- FedRLHF achieves comparable, if not better, performance than centralized RLHF, while preserving privacy



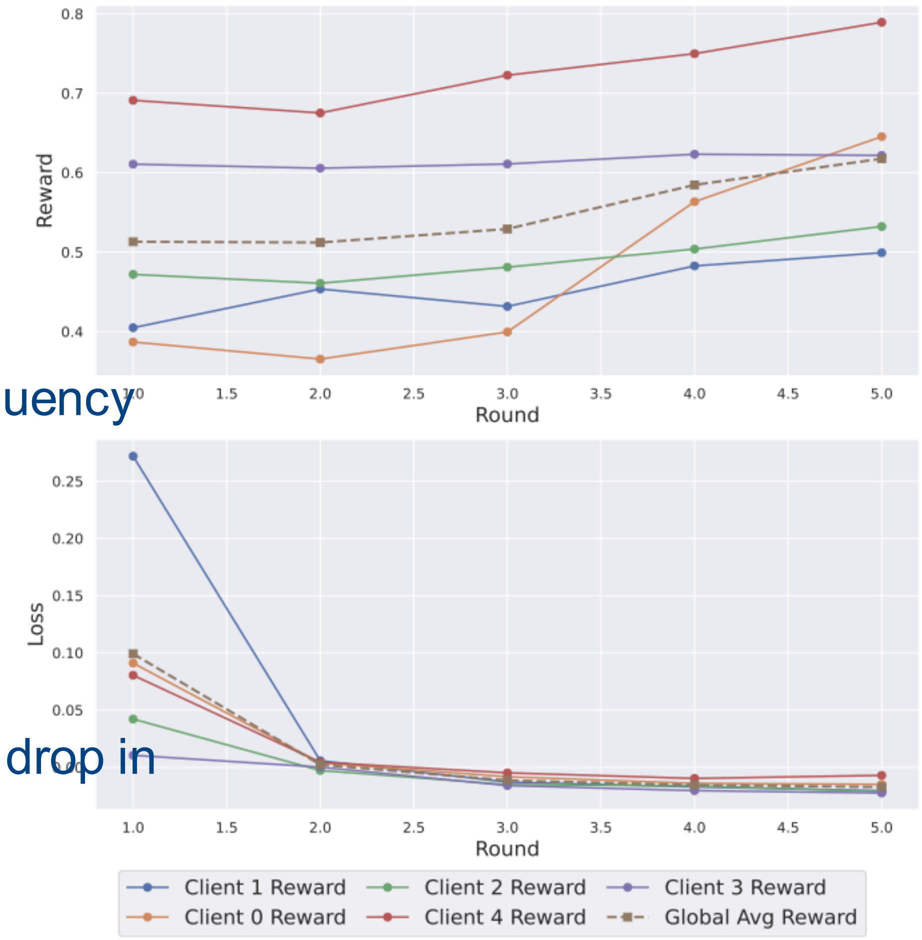
Results – Personalization & λ Tuning

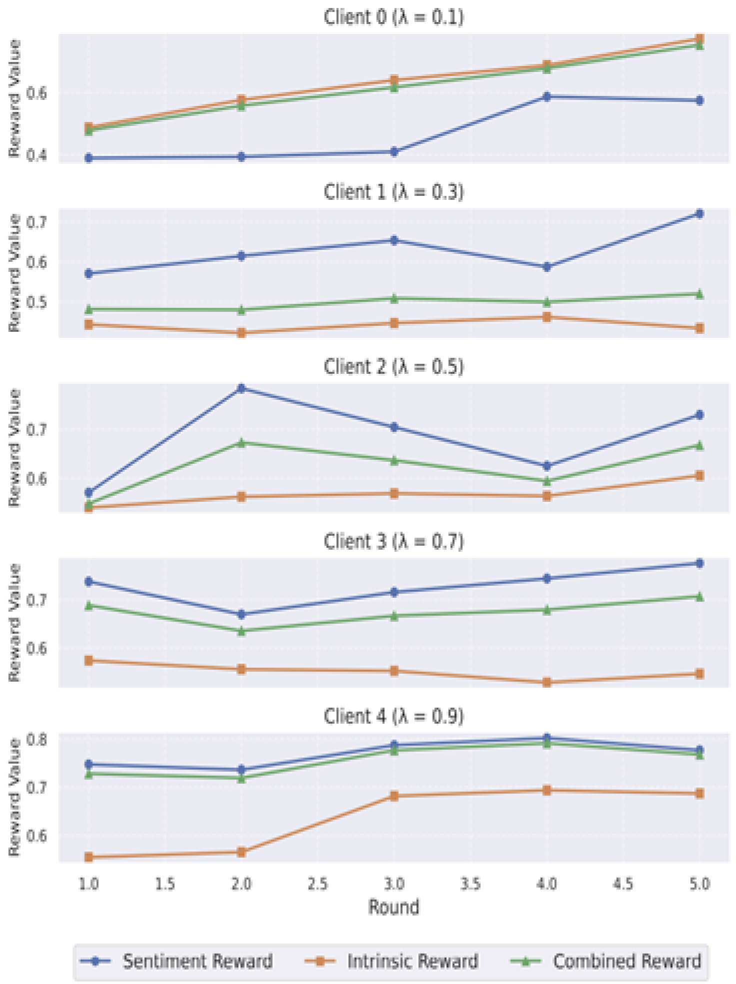
Different λ across clients:

$$\lambda_k \in [0.1, 0.9]$$

Observations:

- Client 0 (λ=0.1): better fluency but weaker sentiment alignment.
- Client 4 (λ=0.9): strong sentiment control, slight drop in fluency.





Conclusion

1. FedRLHF decentralizes RLHF in FL
    - Personalized preference/reward model
2. Theoretical Foundations
    - Convergence guarantees
    - Sample complexity analysis
    - Personalization-performance tradeoff
3. Looking ahead
    - Relaxed assumptions
    - Further privacy enhancements
    - Robust aggregation
    - Principled implementation of feedback

Let's Connect

https://flint-xf-fan.github.io

Project Code

flint-xf-fan/Federated-RLHF

fxf@u.nus.edu

Agency for
Science, Technology
and Research

SINGAPORE

THANK YOU

ASTARSG ASTAR-SG