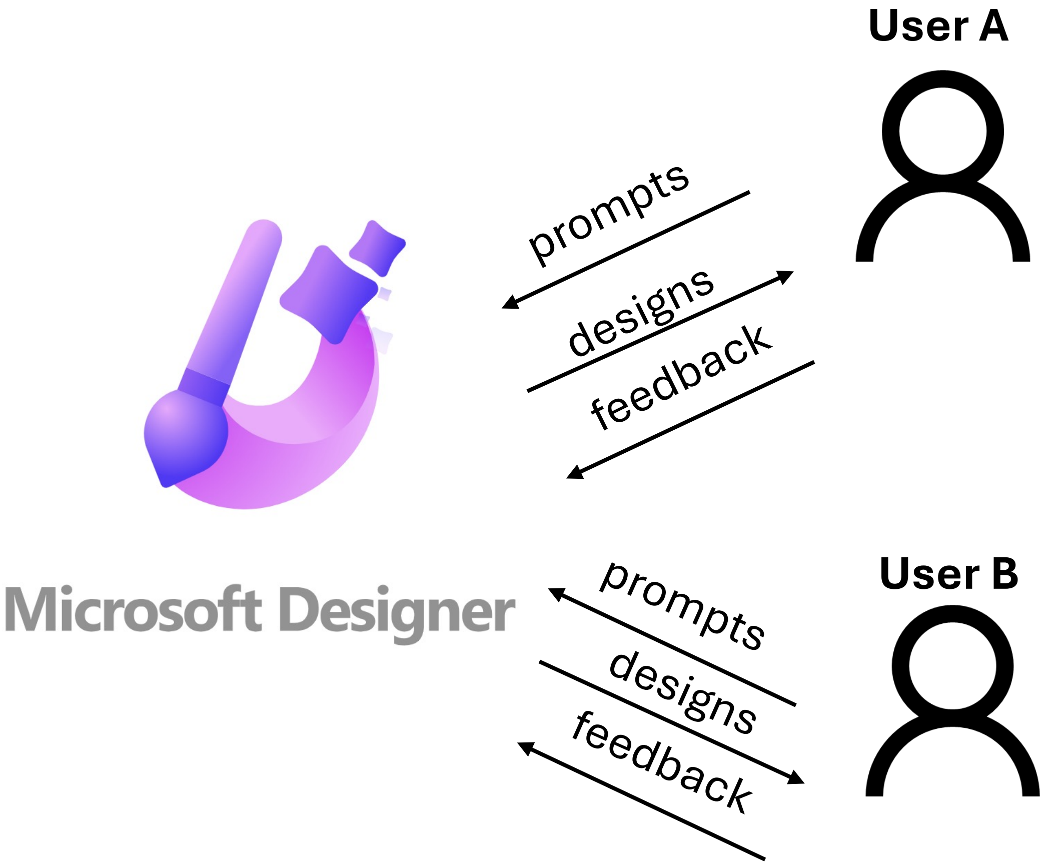
Privately Aligning Language Models with Reinforcement Learning

Fan Wu[1], Huseyin A. Inan[2], Arturs Backurs[3],
Varun Chandrasekaran[1], Janardhan Kulkarni[3], Robert Sim[2]

[1]University of Illinois Urbana-Champaign, [2]M365 Research, [3]Microsoft Research

Motivation – the designer app



User A

prompts

designs

feedback

Microsoft Designer

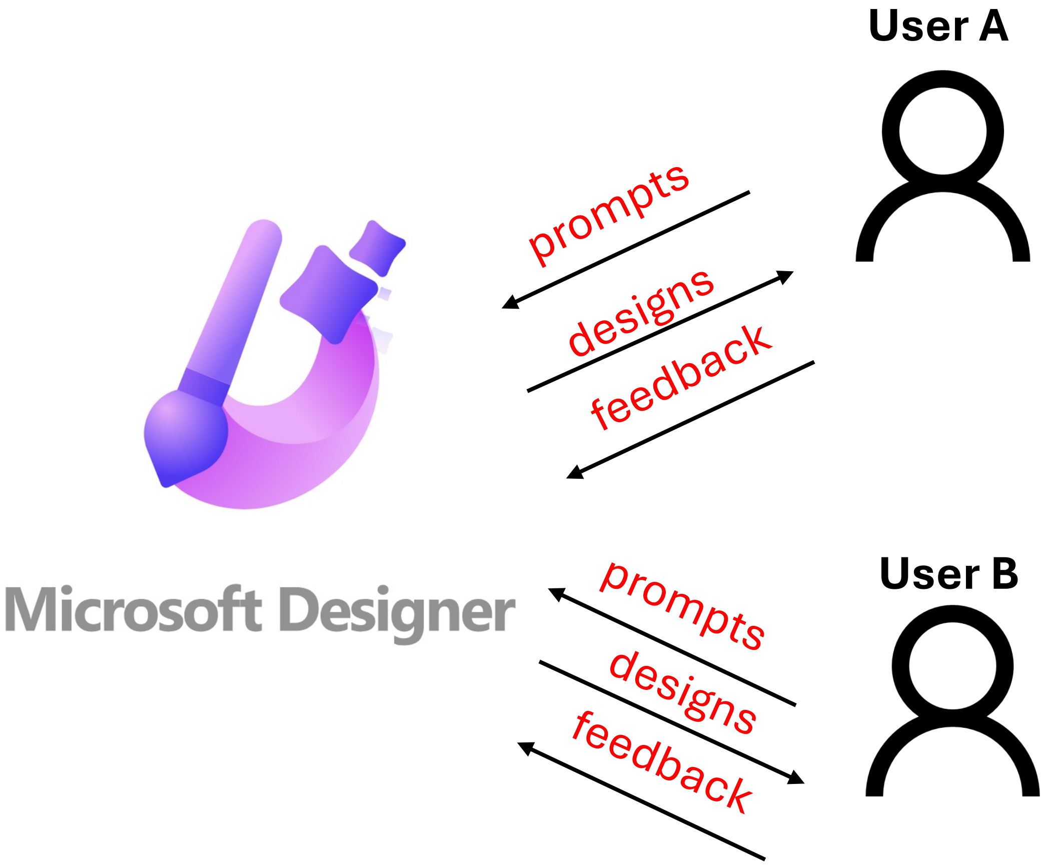
User B

prompts

designs

feedback

Motivation – the designer app



Microsoft Designer

User A

prompts

designs

feedback

User B

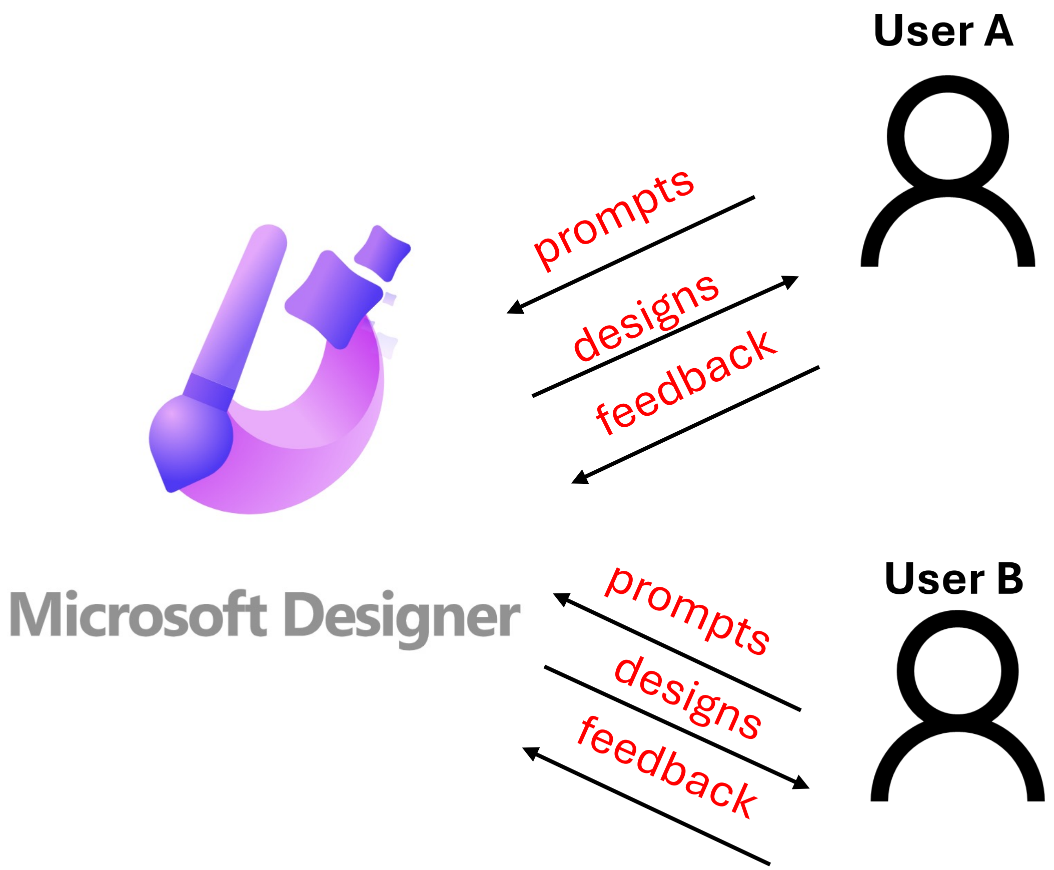
prompts

designs

feedback

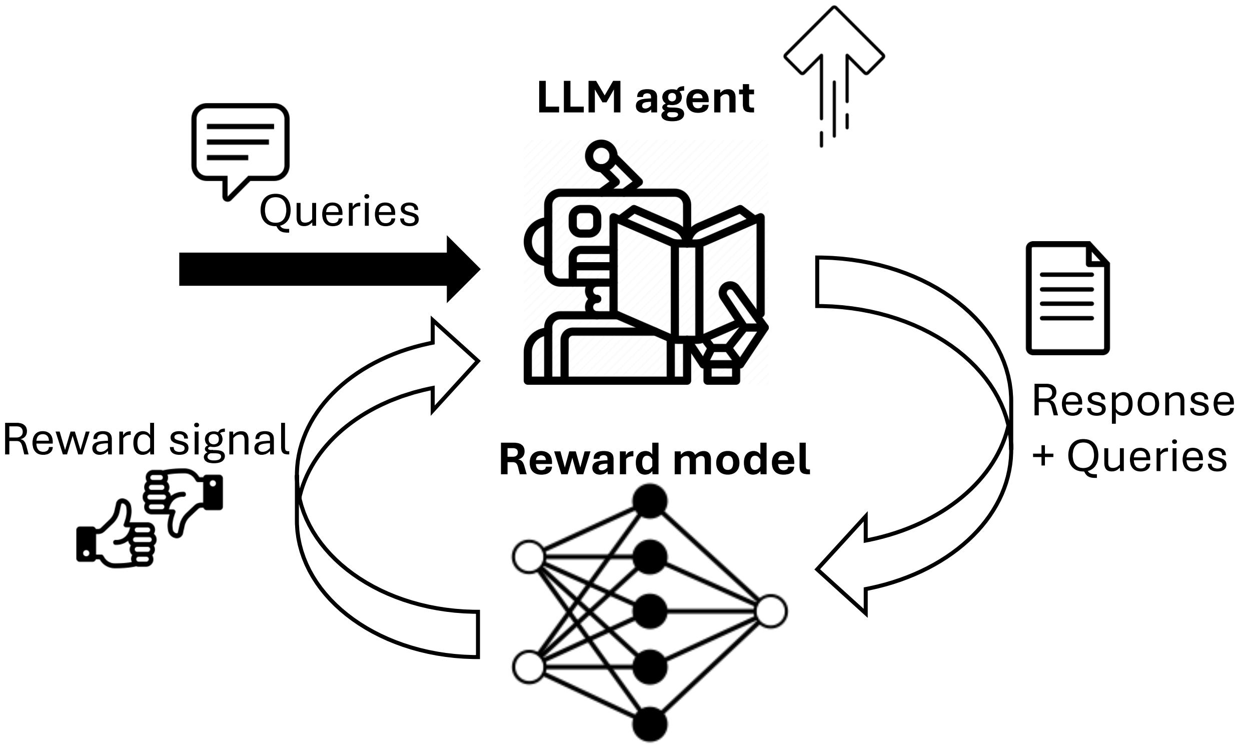
Eyes-off private data!

Motivation – the designer app



PRIVATELY align with HUMAN PREFERENCE

Aligning LLM via RL



LLM agent

Queries

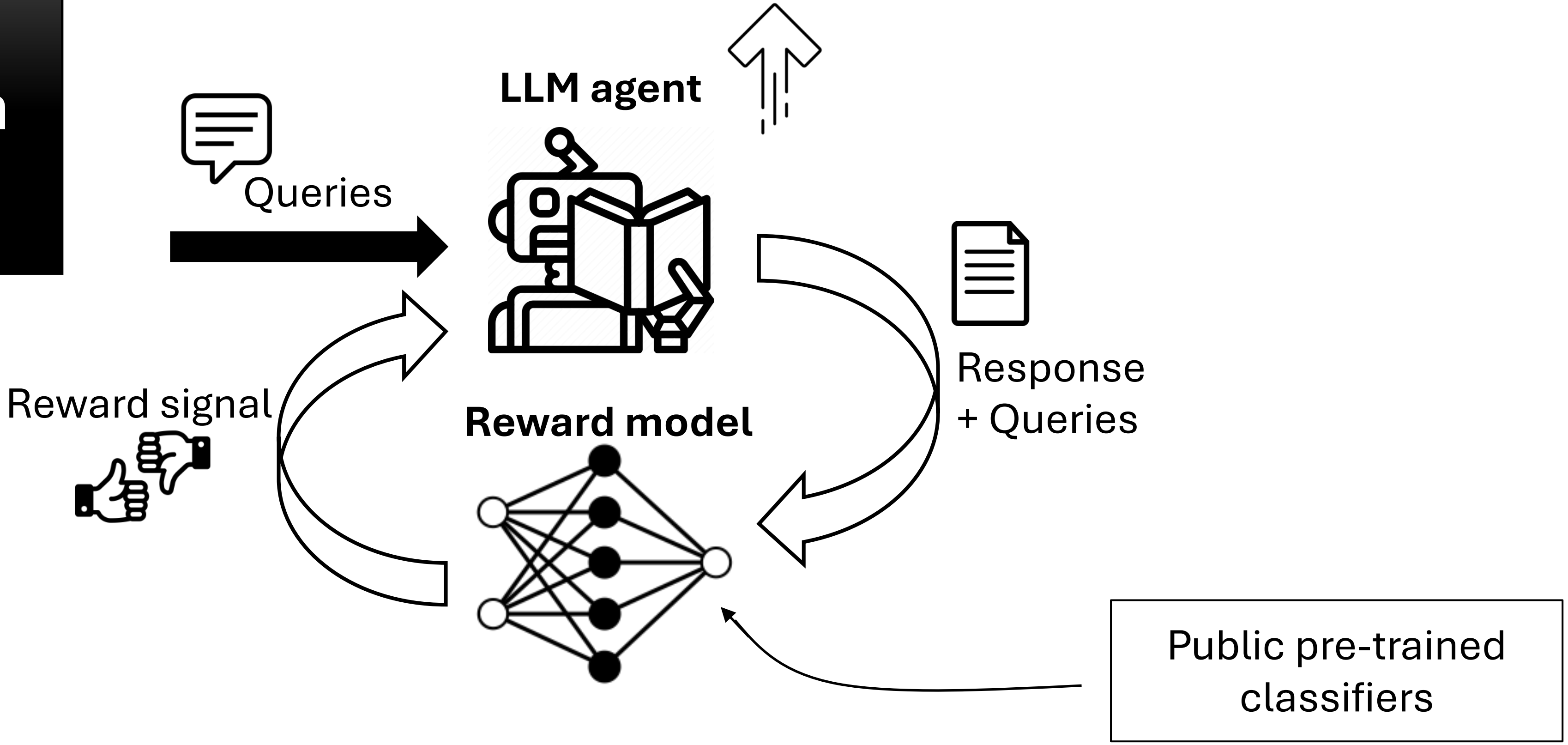
Response
+ Queries

Reward signal

Reward model

Two paradigms

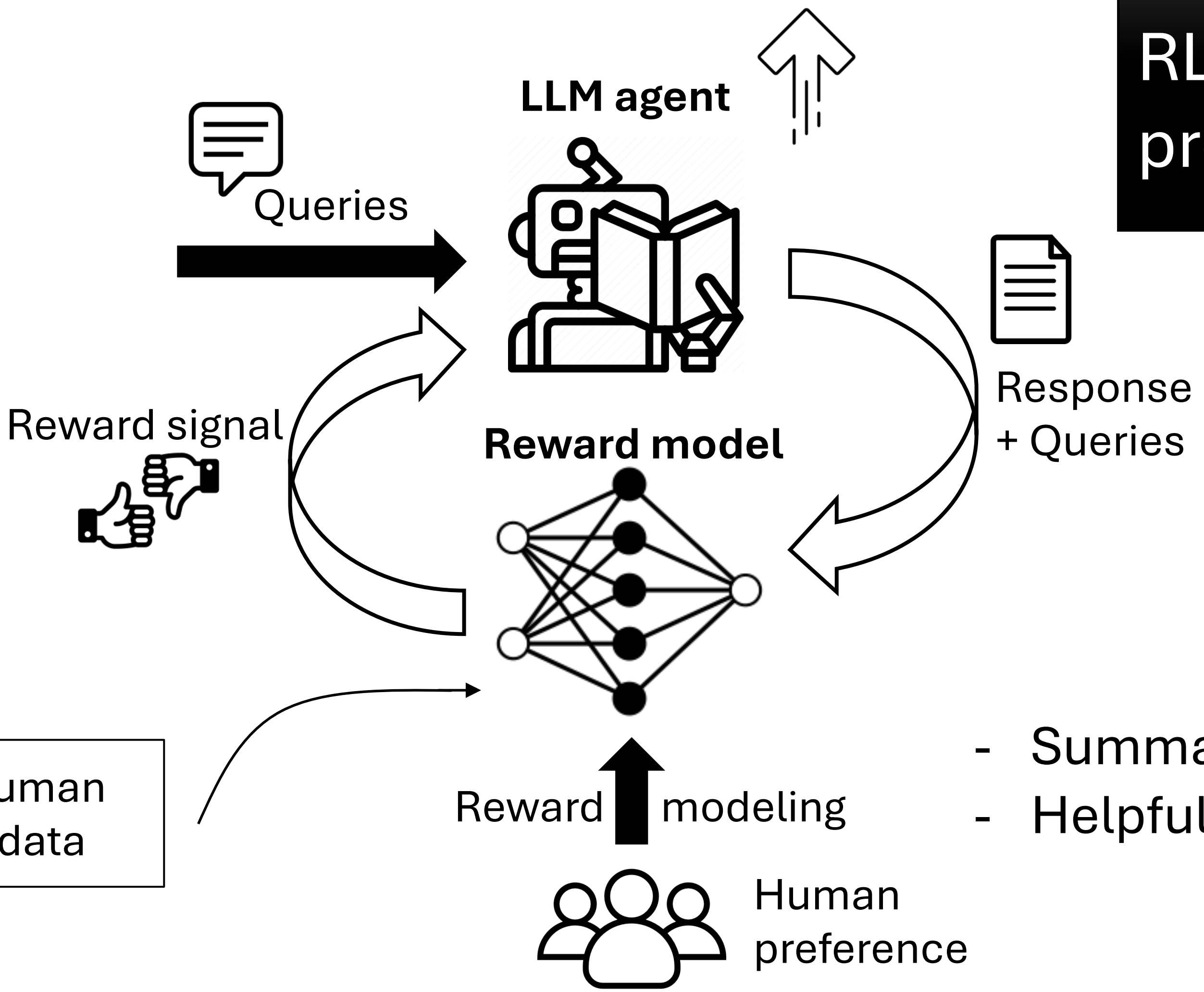
Paradigm 1:
RL without human in the loop



- Sentiment tuning
- Toxicity reduction

Two paradigms

Paradigm 2:
RL with human preference



LLM agent

Queries

Response
+ Queries

Reward signal

Reward model

Reward modeling

Human preference

Trained on human preference data

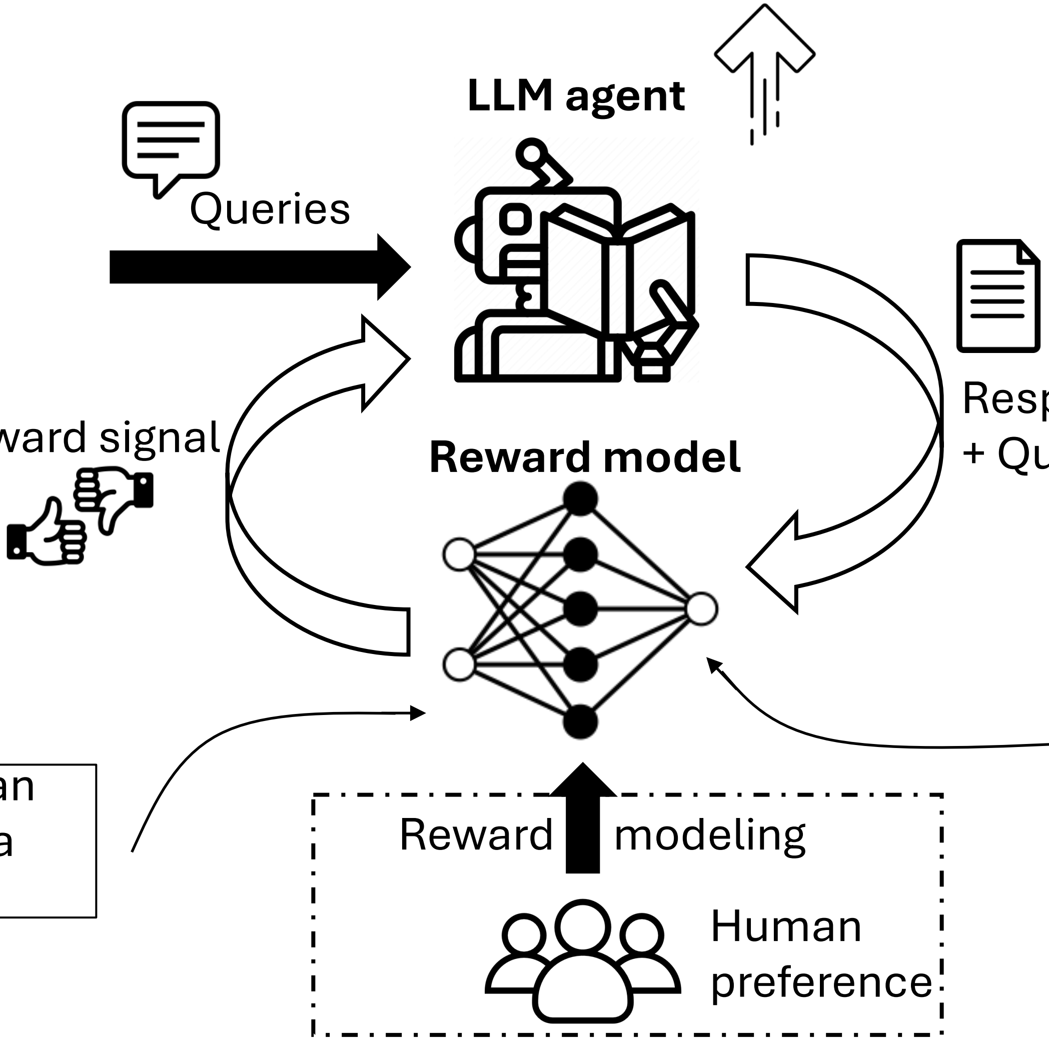
- Summarization
- Helpful and harmless assistants

Two paradigms

Differ in whether the **alignment objective** is easy to **characterize**

Paradigm 1:
RL without human in the loop

Paradigm 2:
RL with human preference



Queries

LLM agent

Response
+ Queries

Reward signal

Reward model

Reward modeling

Human preference

Trained on human preference data (paradigm 2)

Public pre-trained classifiers (paradigm 1)

Differential privacy

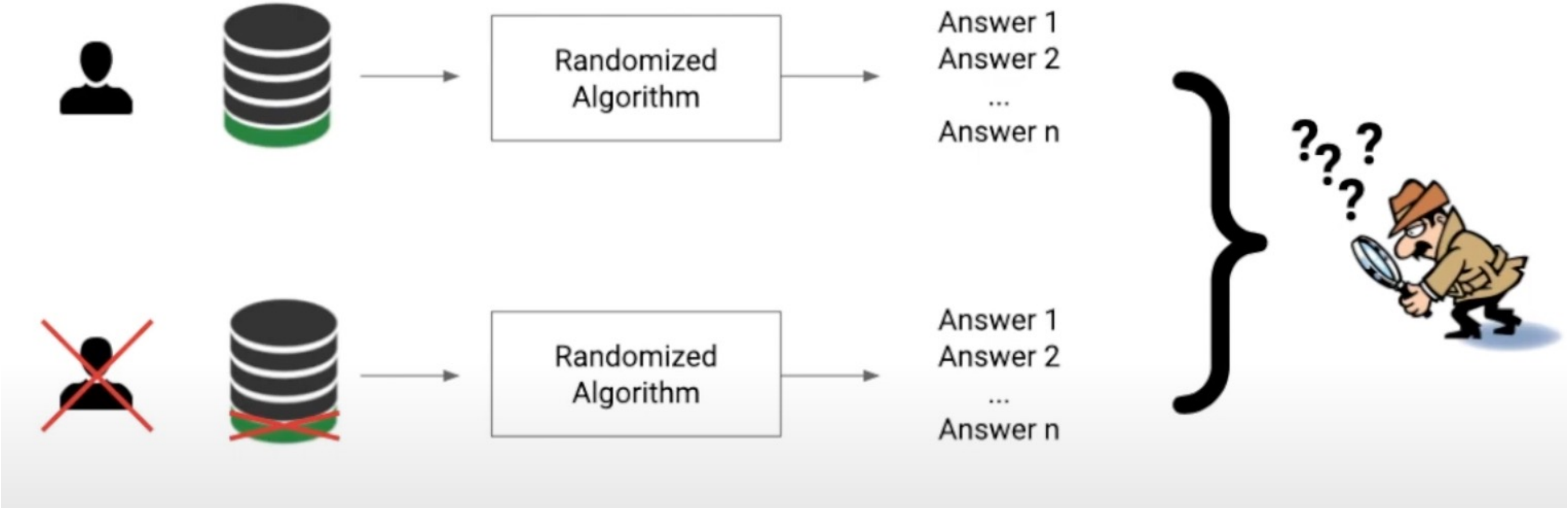


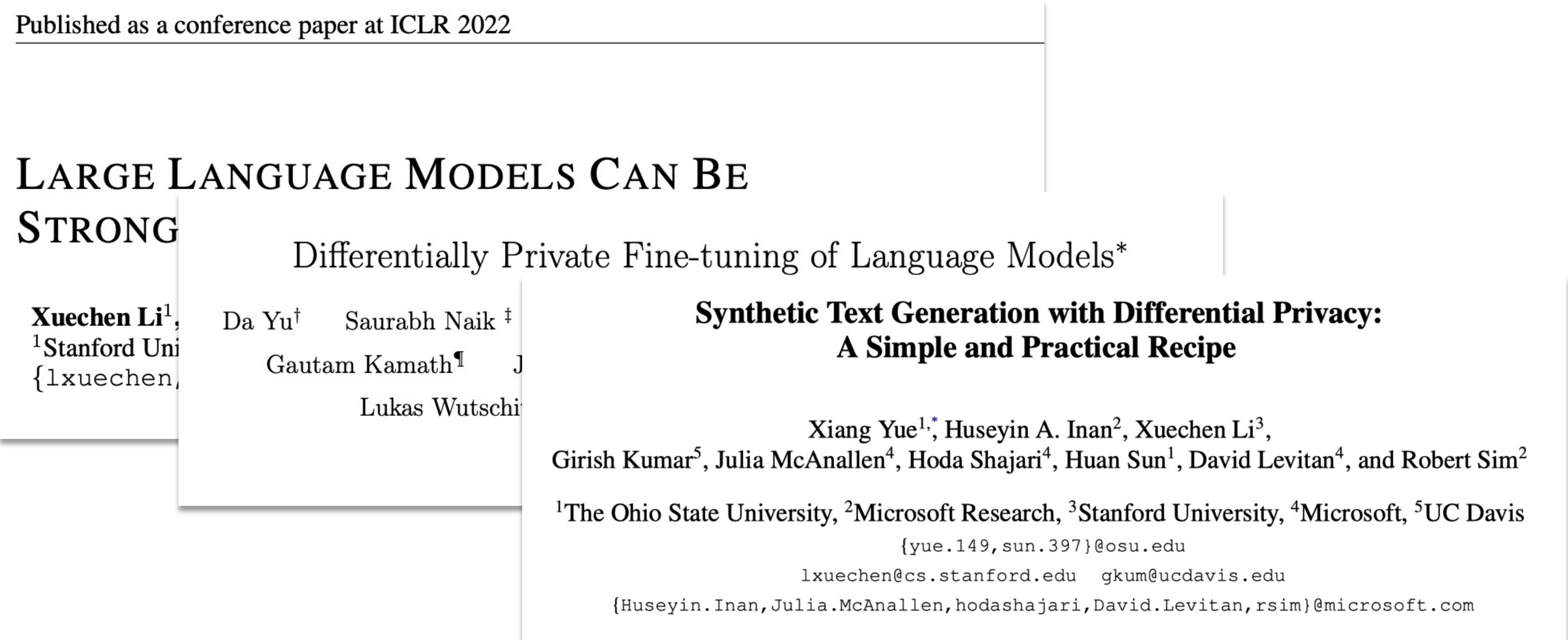
Fig. an illustration of differential privacy. Image from https://youtu.be/YRVBSx0mpO8?si=aHO_sDlZFLHmS9k0

Definition 1 ((ϵ, δ)-DP (Dwork & Roth, 2014)). *A randomized algorithm $\mathcal{M}$ achieves (ϵ, δ)-DP, if for any neighboring datasets D_1 and D_2 (differing in at most one entry) and for any $S \in Range(\mathcal{M})$,*

$$\Pr(\mathcal{M}(D_1) \in S) \leq e^{\epsilon} \Pr(\mathcal{M}(D_2) \in S) + \delta. \tag{3}$$

Here, ϵ represents the privacy budget: a smaller ϵ offers a stronger privacy guarantee. δ accounts for the probability that $\mathcal{M}$ violates ϵ-DP.

Differential privacy in language models



Published as a conference paper at ICLR 2022

LARGE LANGUAGE MODELS CAN BE
STRONG

Xuechen Li[1],
[1]Stanford Uni
{lxuechen,

Differentially Private Fine-tuning of Language Models*

Da Yu† Saurabh Naik ‡
Gautam Kamath¶ J
Lukas Wutschi

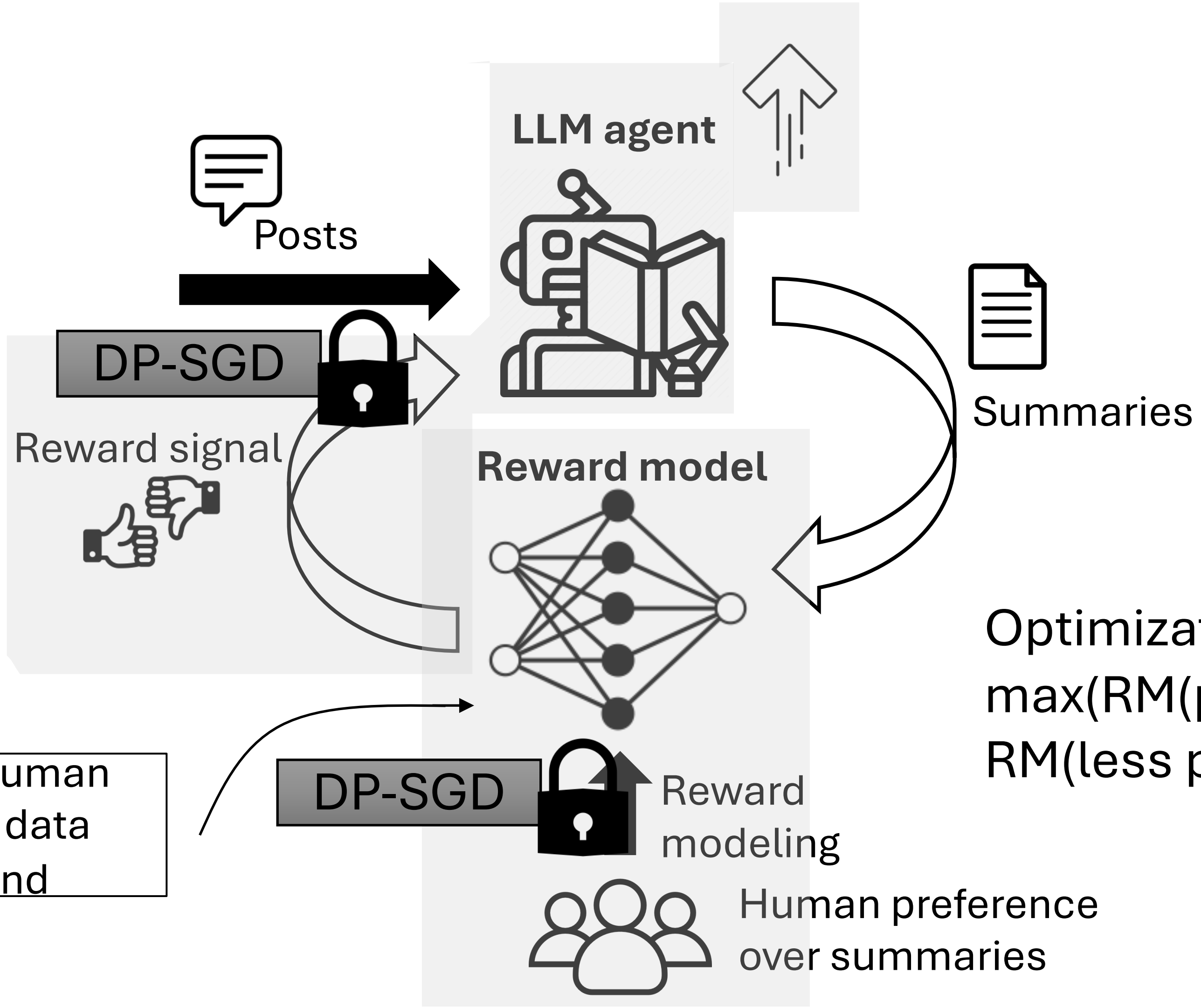
Synthetic Text Generation with Differential Privacy:
A Simple and Practical Recipe

Xiang Yue[1,*], Huseyin A. Inan[2], Xuechen Li[3],
Girish Kumar[5], Julia McAnallen[4], Hoda Shajari[4], Huan Sun[1], David Levitan[4], and Robert Sim[2]

[1]The Ohio State University, [2]Microsoft Research, [3]Stanford University, [4]Microsoft, [5]UC Davis
{yue.149,sun.397}@osu.edu
lxuechen@cs.stanford.edu gkum@ucdavis.edu
{Huseyin.Inan,Julia.McAnallen,hodashajari,David.Levitan,rsim}@microsoft.com

Scenario 2: summarization

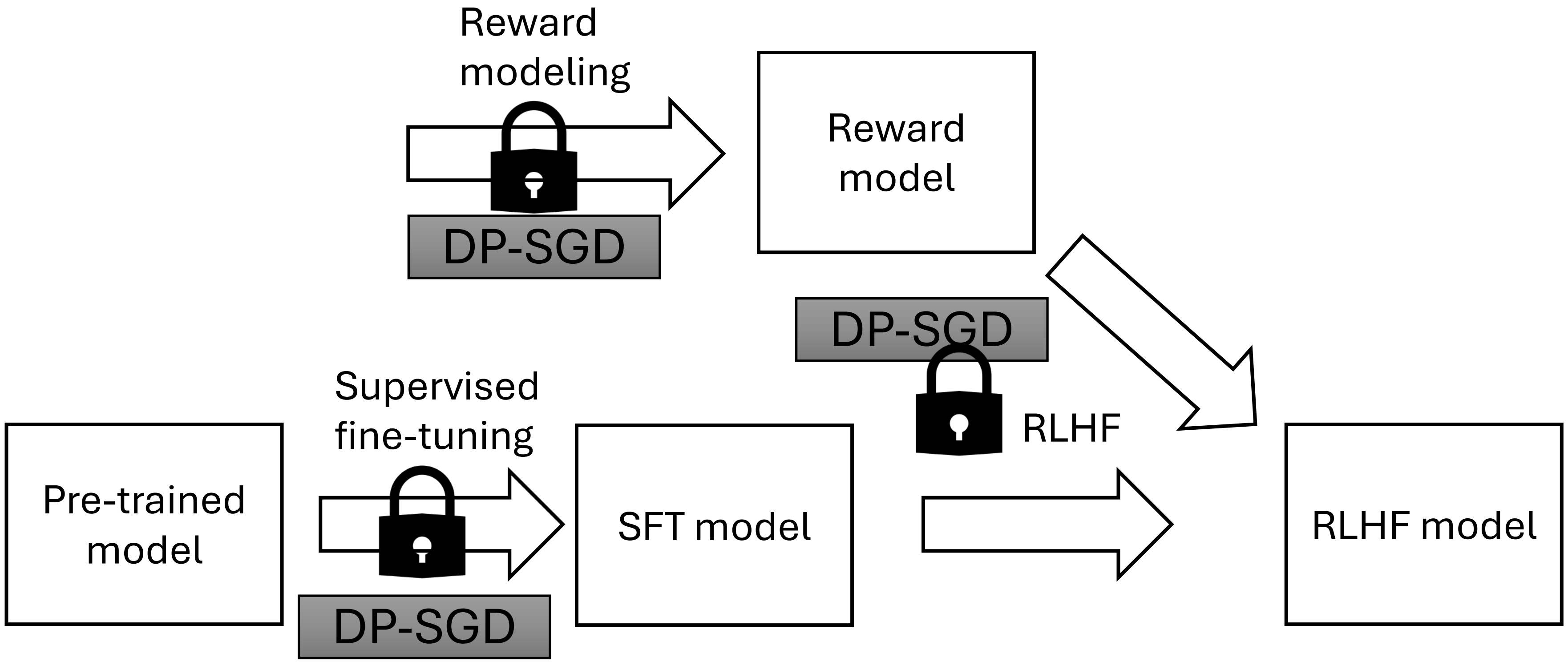
Optimization using
RL algorithms
(e.g., PPO)



Trained on human
preference data
beforehand

Optimization objective:
max(RM(preferred completion) –
RM(less preferred completion))

Scenario 2: summarization – detailed procedures



Privacy analysis:

(For simplicity we take $\varepsilon_1 = \varepsilon_2 = \varepsilon_3 = 4$)

- disjoints datasets for different stages
- $\varepsilon_1, \varepsilon_2, \varepsilon_3$ for the three stages
- Overall consumption: $\max(\varepsilon_1, \varepsilon_2, \varepsilon_3)$ by parallel composition

Scenario 2: summarization – pipeline

Stage I: SFT

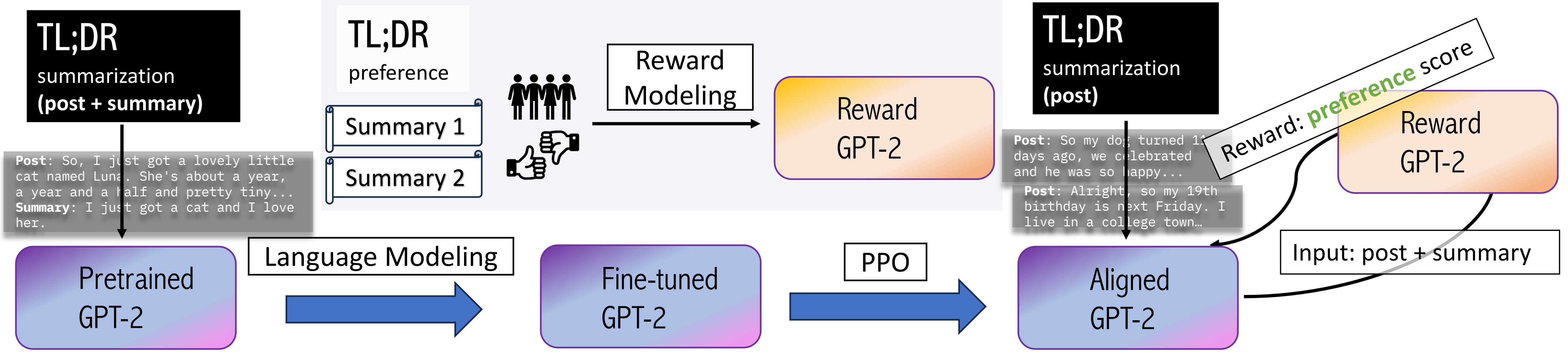
Obj: learn to summarize

Stage II: Reward Modeling

Obj: learn to model human preference

Stage III: RL Fine-Tuning

Obj: learn to summarize with human preference



TL;DR
summarization
(post + summary)

Post: So, I just got a lovely little cat named Luna. She's about a year, a year and a half and pretty tiny...
Summary: I just got a cat and I love her.

TL;DR
preference

TL;DR
summarization
(post)

Post: So my dog turned 11 days ago, we celebrated and he was so happy...

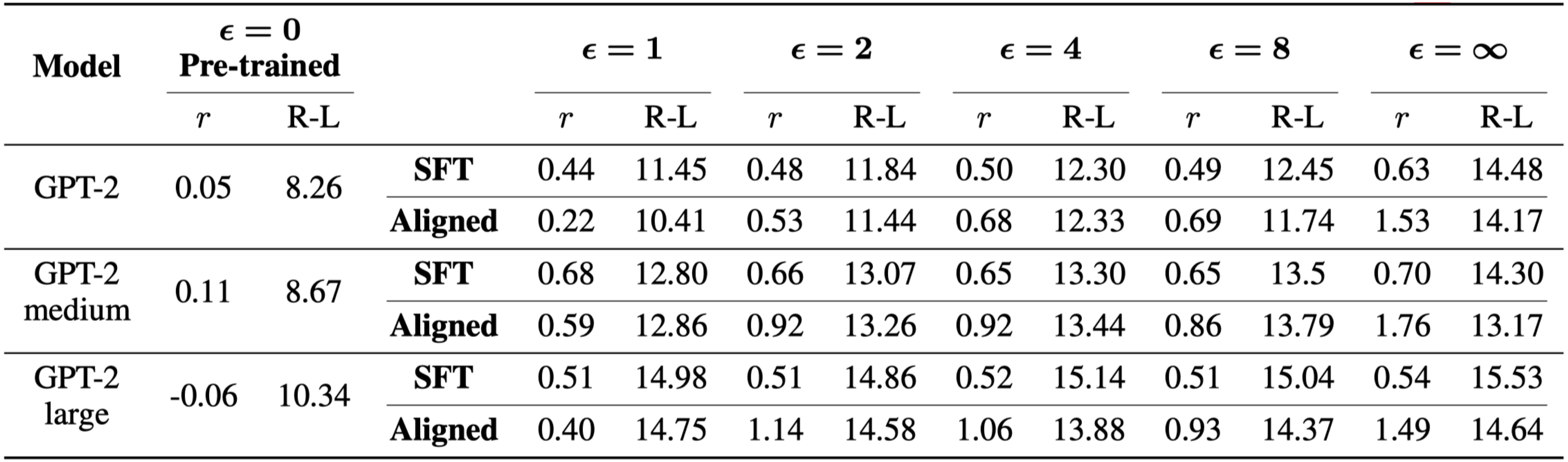
Post: Alright, so my 19th birthday is next Friday. I live in a college town…

Input: post + summary

Scenario 2: summarization – results

- Generation model:
    - gpt2 model family
- Reward model:
    - gpt2

Table 2: The average reward score (r) and ROUGE-L score (R-L)



Model	**$\epsilon = 0$ Pre-trained**			**$\epsilon = 1$**		**$\epsilon = 2$**		**$\epsilon = 4$**		**$\epsilon = 8$**		**$\epsilon = \infty$**	
	r	R-L		r	R-L	r	R-L	r	R-L	r	R-L	r	R-L
GPT-2	0.05	8.26	**SFT**	0.44	11.45	0.48	11.84	0.50	12.30	0.49	12.45	0.63	14.48
			Aligned	0.22	10.41	0.53	11.44	0.68	12.33	0.69	11.74	1.53	14.17
GPT-2 medium	0.11	8.67	**SFT**	0.68	12.80	0.66	13.07	0.65	13.30	0.65	13.5	0.70	14.30
			Aligned	0.59	12.86	0.92	13.26	0.92	13.44	0.86	13.79	1.76	13.17
GPT-2 large	-0.06	10.34	**SFT**	0.51	14.98	0.51	14.86	0.52	15.14	0.51	15.04	0.54	15.53
			Aligned	0.40	14.75	1.14	14.58	1.06	13.88	0.93	14.37	1.49	14.64

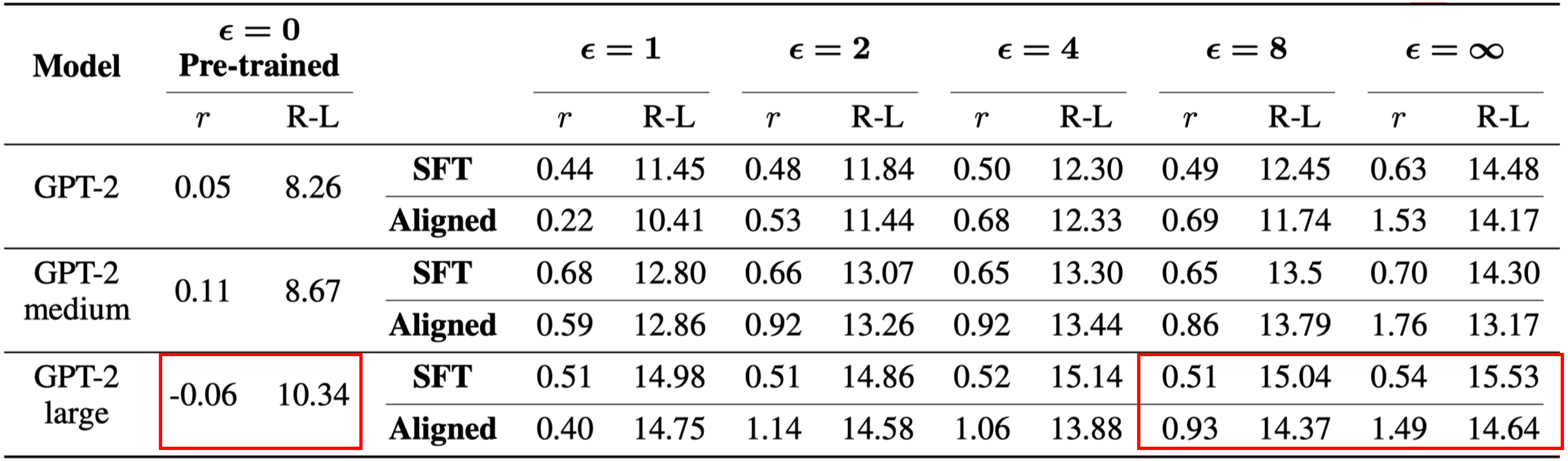
Scenario 2: summarization – results

- Generation model:
  - gpt2 model family
- Reward model:
  - gpt2

Observation 2:
Non-private RLHF > dp RLHF >> PT

DP leads to only mild degradation

Table 2: The average reward score (r) and ROUGE-L score (R-L)



Model	$\epsilon = 0$ Pre-trained			$\epsilon = 1$		$\epsilon = 2$		$\epsilon = 4$		$\epsilon = 8$		$\epsilon = \infty$	
	r	R-L		r	R-L	r	R-L	r	R-L	r	R-L	r	R-L
GPT-2	0.05	8.26	**SFT**	0.44	11.45	0.48	11.84	0.50	12.30	0.49	12.45	0.63	14.48
			Aligned	0.22	10.41	0.53	11.44	0.68	12.33	0.69	11.74	1.53	14.17
GPT-2 medium	0.11	8.67	**SFT**	0.68	12.80	0.66	13.07	0.65	13.30	0.65	13.5	0.70	14.30
			Aligned	0.59	12.86	0.92	13.26	0.92	13.44	0.86	13.79	1.76	13.17
GPT-2 large	-0.06	10.34	**SFT**	0.51	14.98	0.51	14.86	0.52	15.14	0.51	15.04	0.54	15.53
			Aligned	0.40	14.75	1.14	14.58	1.06	13.88	0.93	14.37	1.49	14.64

Scenario 2: summarization – case studies

Post:
I'll keep this short. I'm teaching abroad and met this girl in teaching training. We have similar geeky interests so we meet up every couple months for a weekend of city fun together. She IS very fun except for an ongoing habit negative commentary for EVERYTHING. I don't think she realizes she does it and it's her way of making conversation. For example if I asked how her food was she'd tell me with a smile that it was fine but this place is too crowded food tastes weird too expensiveeven if it was a restaurant she recommended. Now imagine that with EVERYTHING. I've literally never heard her talk about something without having to bring up a flaw it has . NOTHING can just be GOOD. After a weekend it drives me up a wall. A few times I've almost joked You ever realize you're kind of a debbie downer but I think this is too harsh since we're not super close. Should I just suck it up as an annoying character trait It really pisses me off when she happily shits on something I suggested but again I think she's totally unaware of her constant criticizing. We're both pretty blunt people you can probably tell she is so I think I canbe a bit direct. How do I hint to cut back on the bitching Or do I grin and bear it every few months

PT model generated summary:
TIP: Stress listening to everyone. Hinting clothing that's going to make you look like shit. Don't run away or whine. Varried someone else before and they just lectured me with their take on mangamedoes. Learn to admit when you're wrong. Make outlandish claims as long as you use innuendo and personal attacks. Tricking your way into conversation is way more efficient than acknowledging and waiting for someone to understand you. Get your friend's web password

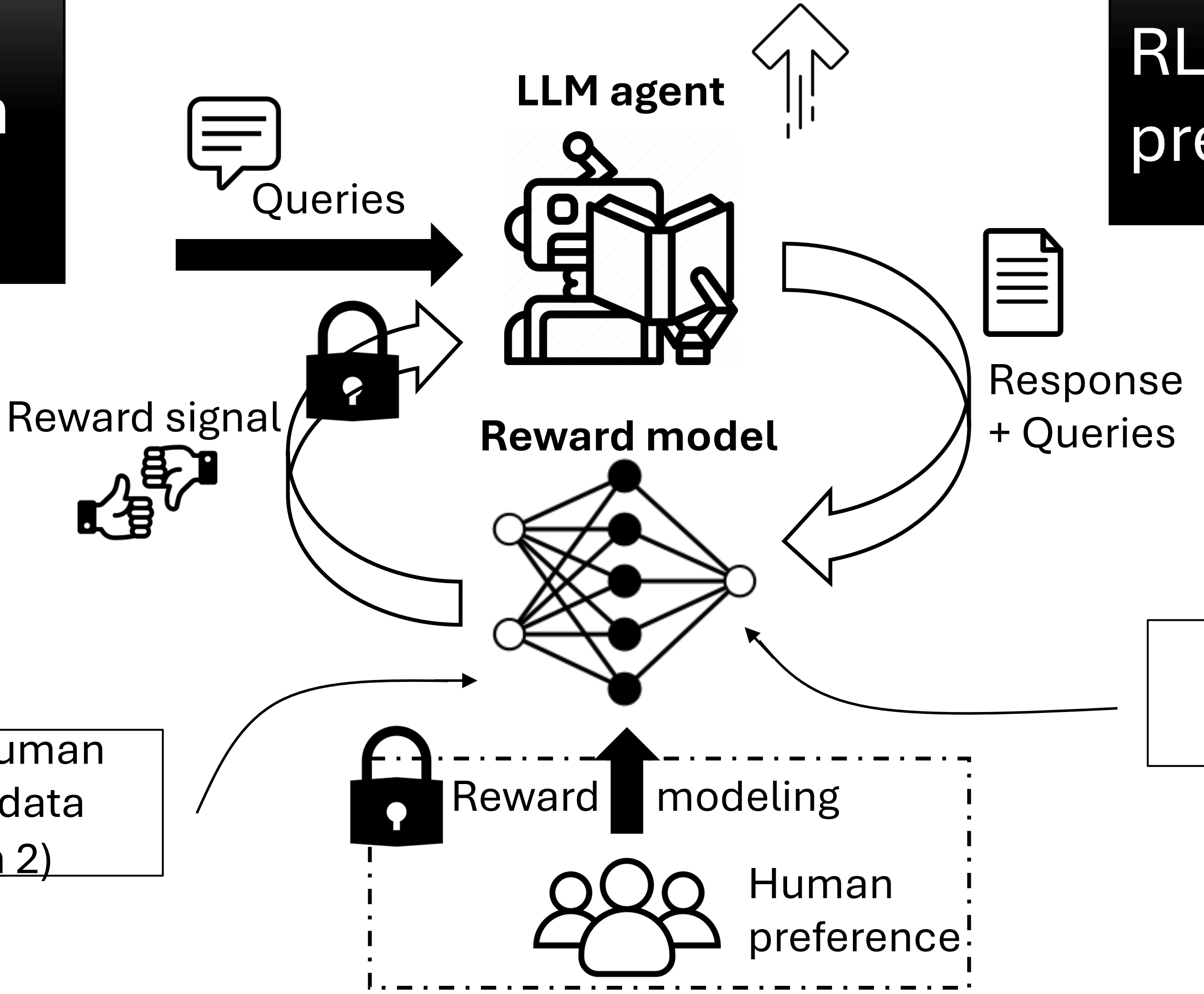
SFT model generated summary:
I know I'm annoying but I fear she's enjoying the same criticism that I do. How do I stop

RLHF model generated summary:
Girl is bad at communicating and constantly shits on everything I suggest. How do I gently hint to cut her down without prompting?

Conclusions

Paradigm 1:
RL without human in the loop

Paradigm 2:
RL with human preference



Trained on human preference data (paradigm 2)

Public pre-trained classifiers (paradigm 1)

Thank you!

Scan to visit the paper!

FLOWCODE

PRIVACY.FLOWCODE.COM