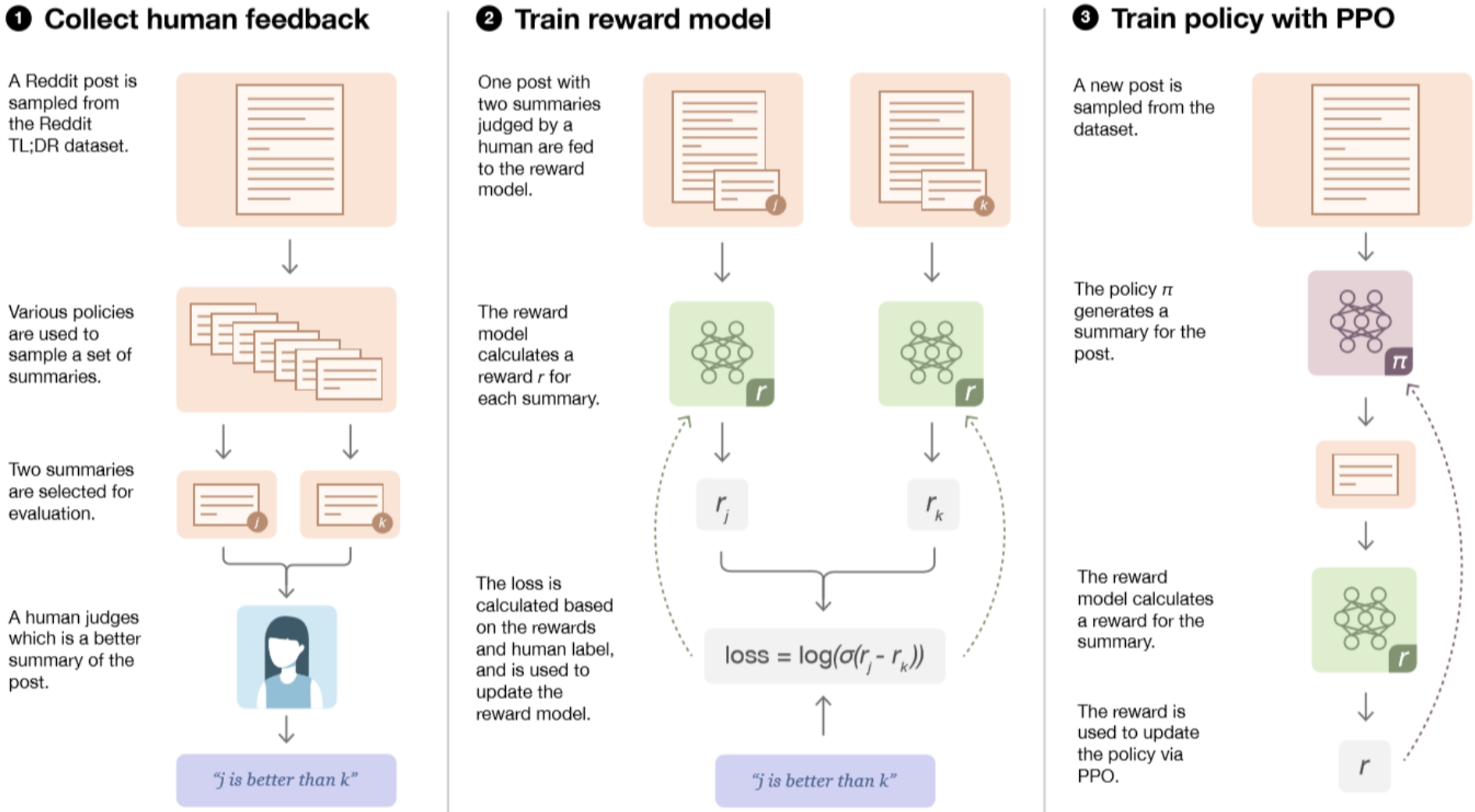
New ML Technologies in OpenAI's ChatGPT

12/14/2022

Learning to Summarize with Human Feedback



Proximal Policy Optimization (PPO)

PPO computes an update at each step that minimizes the cost function while ensuring the deviation from the previous policy is relatively small. In short, **小步快跑**. For instance, PPO uses an adaptive KL penalty to control the change of the policy at each iteration. The loss function for ChatGPT is

$$L^{CLIP}(\theta) = \hat{E}_t[min(r_t(\theta)\hat{A}_t, clip(r_t(\theta), 1-\varepsilon, 1+\varepsilon)\hat{A}_t)\ ]$$

- θ is the policy parameter
- $\hat{E}_t$ denotes the empirical expectation over timesteps
- r_t is the ratio of the probability under the new and old policies, respectively
- $\hat{A}_t$is the estimated advantage at time t
- ε is a hyperparameter, usually 0.1 or 0.2

Reinforcement Learning with Human Feedback

1. Collect the data of human feedback/reaction.
2. Train the reward model/function that can simulate the human's evaluation, say Eval. Remind RL model of AlaphZero.
3. Train the RL model, updating the policy by means of the reward model Eval.

The main idea is from Turing's "child machine", and the teacher is simulated to evaluate the performance of ML task (for instance, text summarization). "**文无第一，武无第二**", what's the difference?

- 武：As a comparison, the supervised learning is suitable to the data with explicit labels, like object identification.
- 文：RLHF is suitable to QA, machine translation, etc.

ChatGPT: RLHF

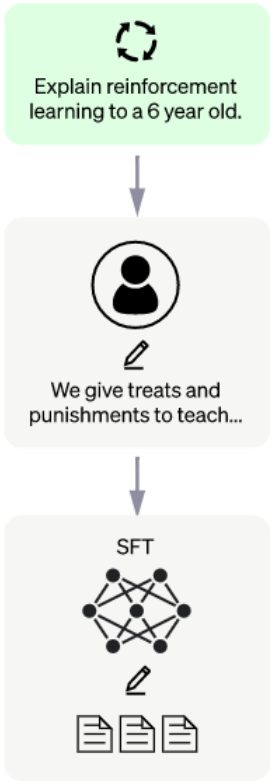
Step 1

Collect demonstration data and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3.5 with supervised learning.



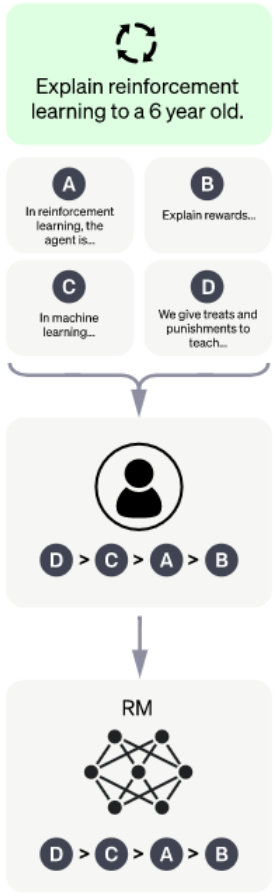
Step 2

Collect comparison data and train a reward model.

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.



Step 3

Optimize a policy against the reward model using the PPO reinforcement learning algorithm.

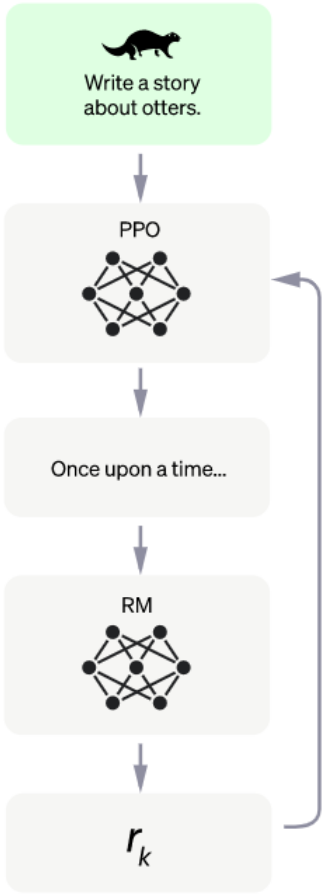
A new prompt is sampled from the dataset.

The PPO model is initialized from the supervised policy.

The policy generates an output.

The reward model calculates a reward for the output.

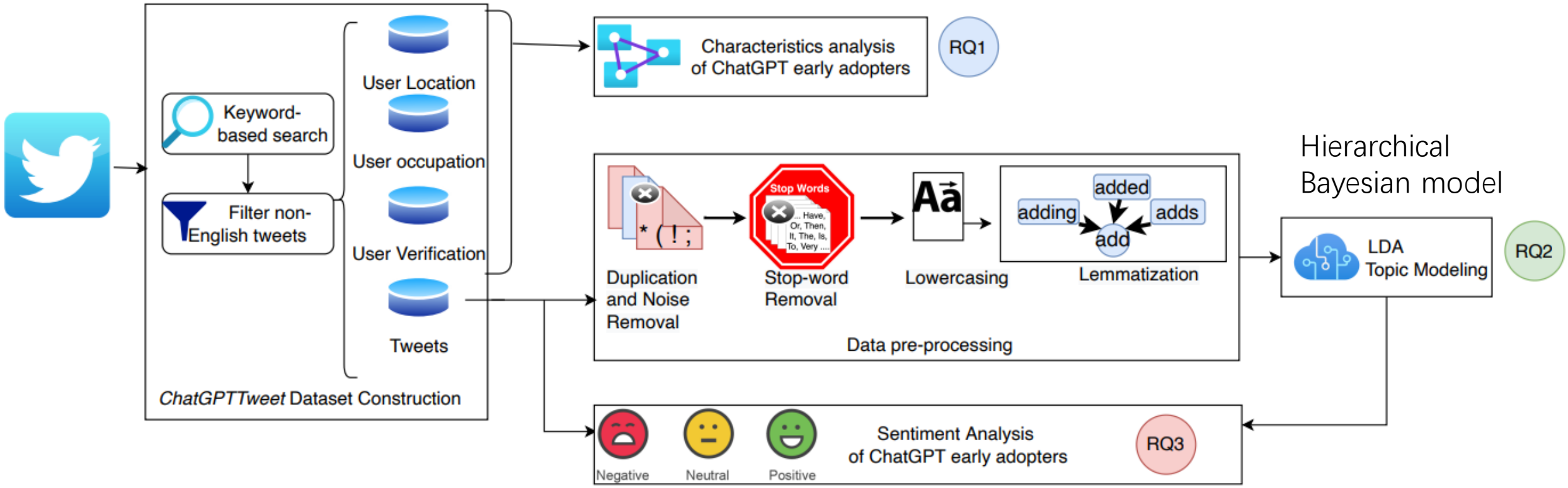
The reward is used to update the policy using PPO.



Challenges of ChatGPT

1. **Knowledge/fact/experience is necessary**: During RL training, there's currently no source of truth. So, sometimes ChatGPT generates ridiculous answers.
2. **AI ethics**: Training the model to be more cautious causes it to decline questions that it can answer correctly. Thus, ChatGPT can avoid some sensitive questions.
3. **More adaptive learning is needed**: Supervised training misleads the model because the ideal answer depends on what the model knows, rather than what the human demonstrator knows. It will be great to let the machine find the truth itself.
4. **Consistency in logic/semantics**: Among all the candidate answers, how to guarantee the logic/semantics is consistent?

Further Application: Topic Model for Twitter



What possible applications to our business? For instance, customer service, autonomous driving, GTS, automatic modeling for cloud computing, etc.