# Self-Rewarding Language Models

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li,
Sainbayar Sukhbaatar, Jing Xu, Jason Weston

# Motivation

- Standard alignment approach

# Motivation

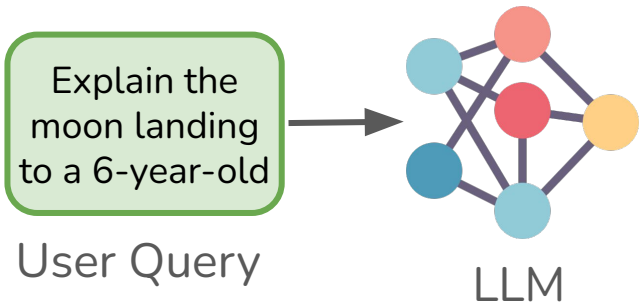
- Standard alignment approach

Explain the moon landing to a 6-year-old

User Query

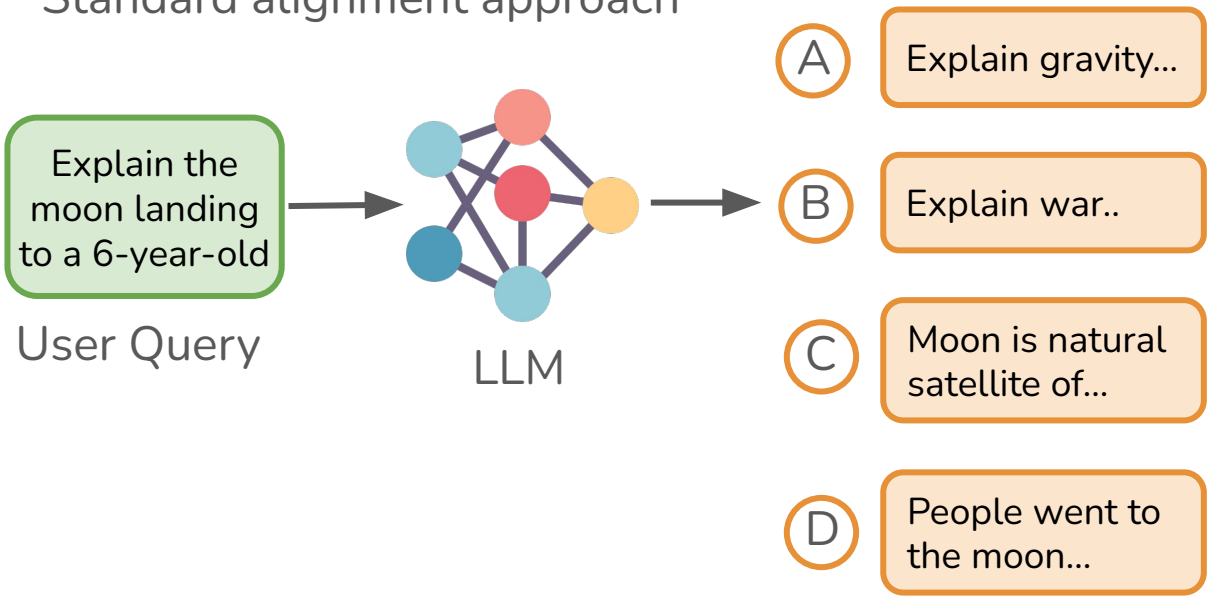
# Motivation

- Standard alignment approach



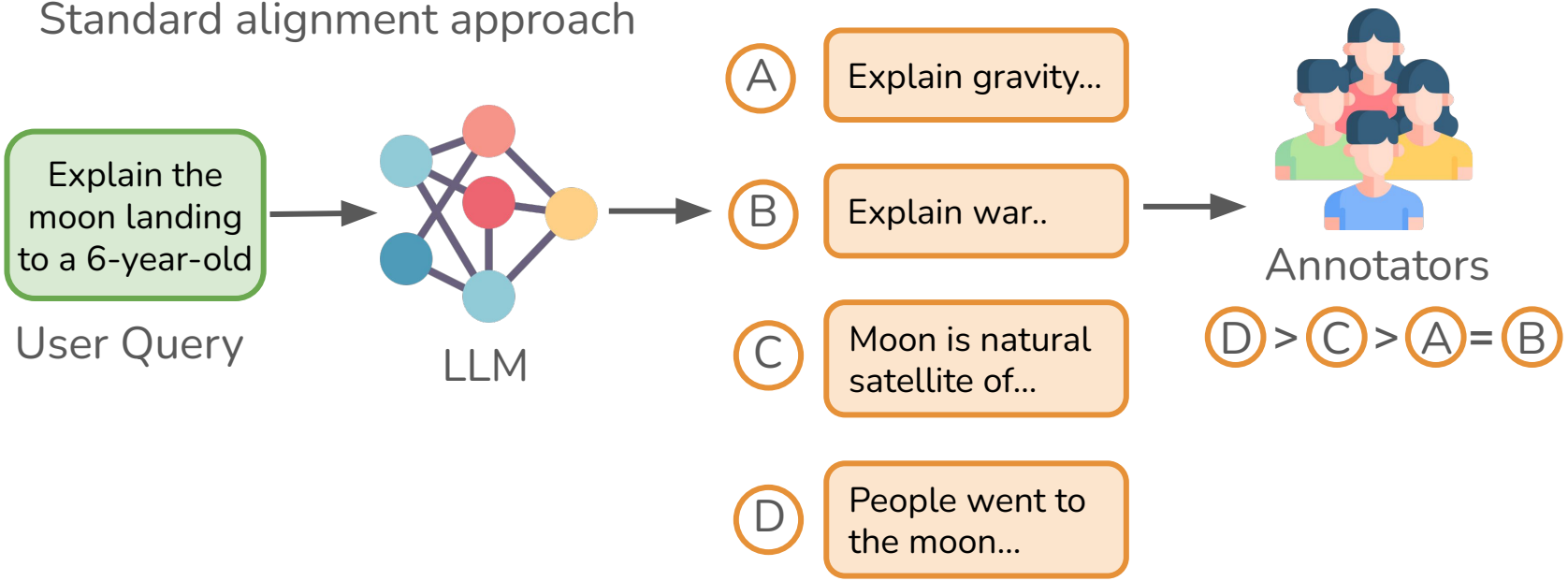
# Motivation

- Standard alignment approach



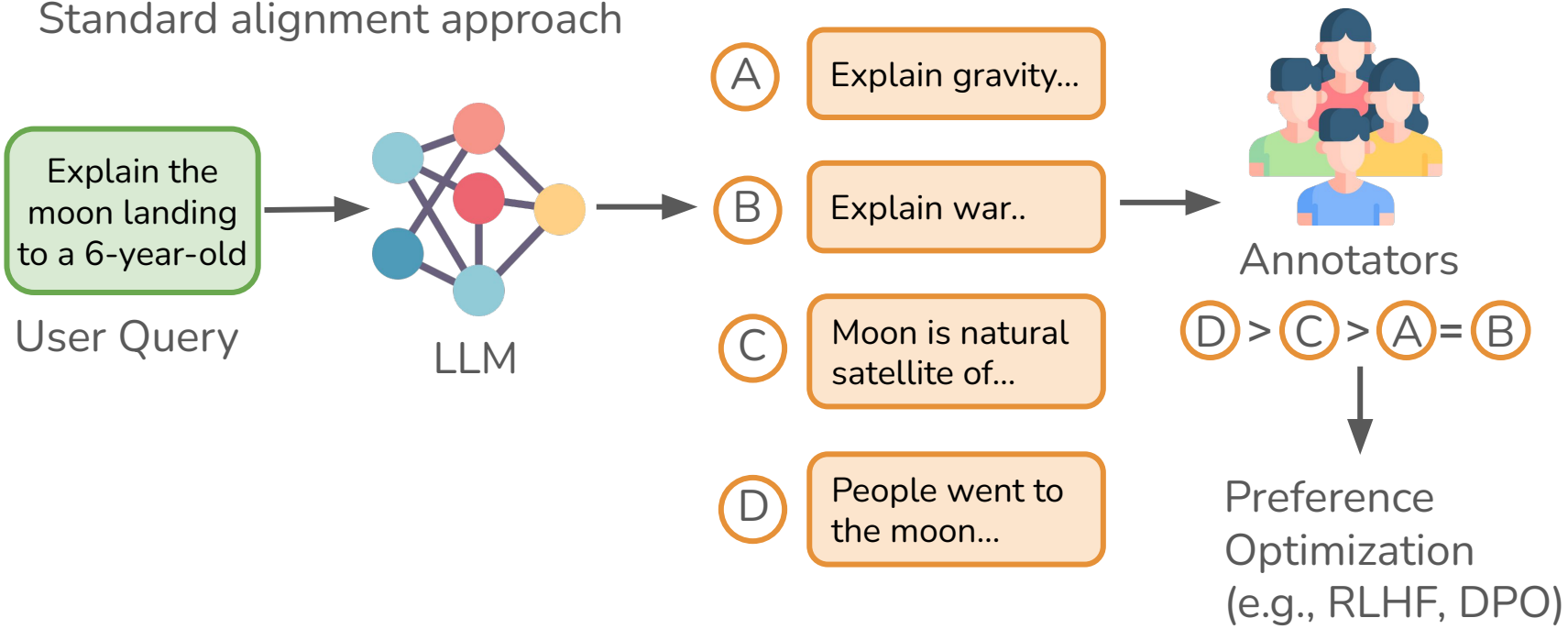
# Motivation

- Standard alignment approach



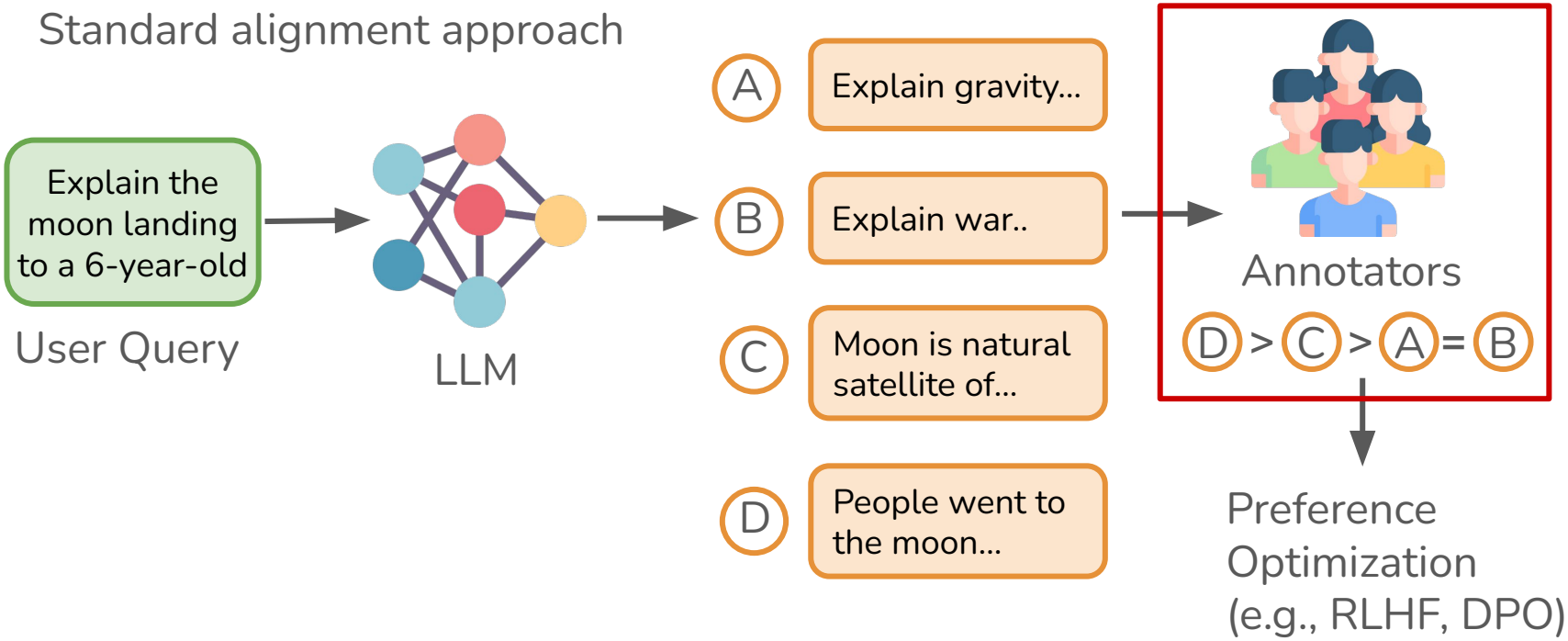
# Motivation

- Standard alignment approach



# Motivation

- Standard alignment approach



# Motivation

- However, as LLMs write **better and better** responses…

# Motivation

- However, as LLMs write **better and better** responses…
  - It becomes **harder and harder** for humans to process them, especially those that are lengthy and require domain expertise.

# Motivation

- However, as LLMs write **better and better** responses...
  - It becomes **harder and harder** for humans to process them, especially those that are lengthy and require domain expertise.

Images generated by GPT-4

# Research Question 🤔

- How can we continue improving superhuman models?

# Observations 🧐

# Observations 🧐

- Observation 1
  - LLMs can continue improving if provided good judgements on response quality
    - Exemplified by the success of iterative RLHF
      - Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback
      - Llama 2: Open Foundation and Fine-Tuned Chat Models

# Observations 🧐

- Observation 1
  - LLMs can continue improving if provided good judgements on response quality
    - Exemplified by the success of iterative RLHF
      - Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback
      - Llama 2: Open Foundation and Fine-Tuned Chat Models
- Observation 2
  - LLMs can provide good judgements on model generations
    - Exemplified by the line of works that use GPT-4 for evaluation
      - Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena
      - AlpacaEval: An Automatic Evaluator of Instruction-following Models

# Then, how about combining them together?

- Self-Rewarding LMs come to rescue!

# Our approach

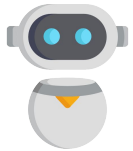
- Self-rewarding LMs
  - **Key idea**: train a self-rewarding language model that

# Our approach

- Self-rewarding LMs
  - **Key idea**: train a self-rewarding language model that
    - Has instruction following capability, i.e., given a user instruction, can respond to it appropriately

What is machine learning?

Machine learning is a subfield of artificial intelligence (AI) that ...



# Our approach

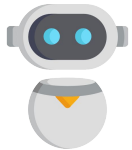
- Self-rewarding LMs
  - **Key idea**: train a self-rewarding language model that
    - Has instruction following capability, i.e., given a user instruction, can respond to it appropriately
    - Has evaluation capability, i.e., given a user instruction, a response, can judge the quality of that response

**Here is an instruction:** Can you explain contrastive learning in machine learning in simple terms for someone new to the field of ML?

**Here is the model response:** <MODEL_RESPONSE>

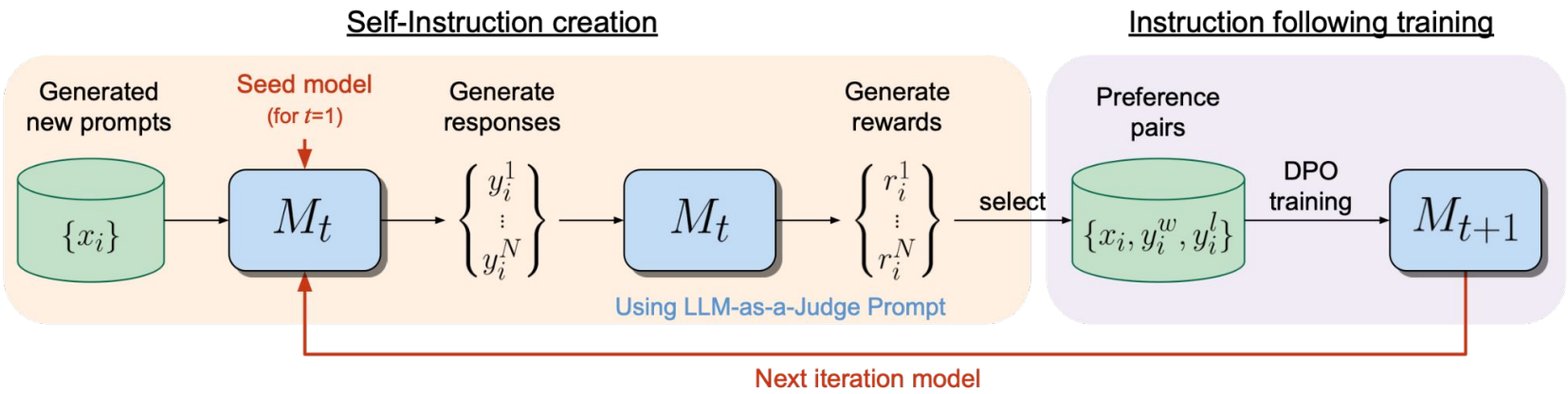
**Can you assign a score (0 to 5) to this response based on the following rubrics?** <RUBRICS>

<CoT reasoning process>
Therefore, I would assign 3 out of 5 to this response.



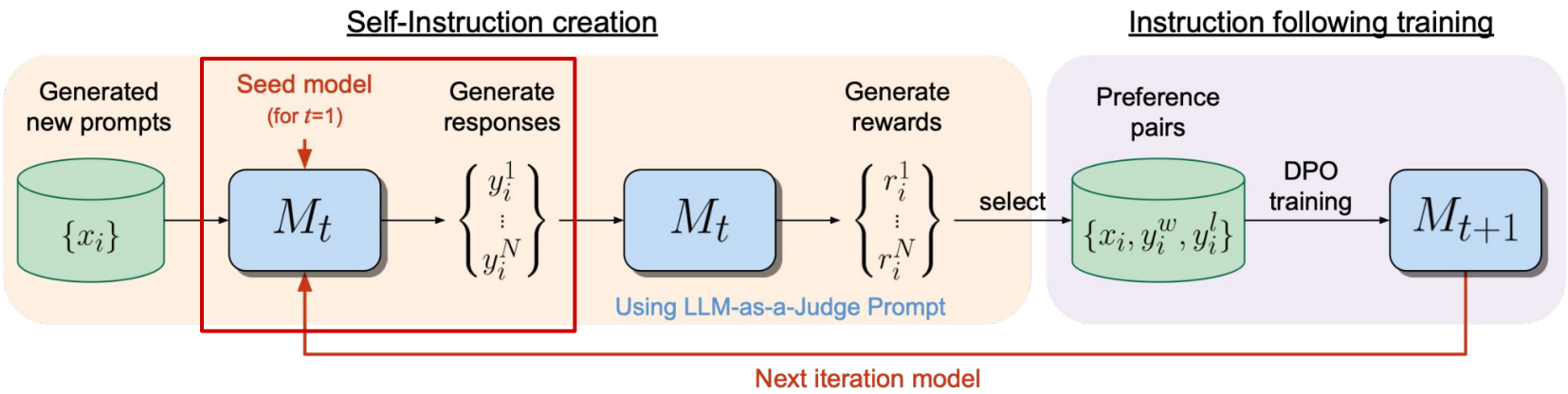
# Our approach

- Self-rewarding LMs



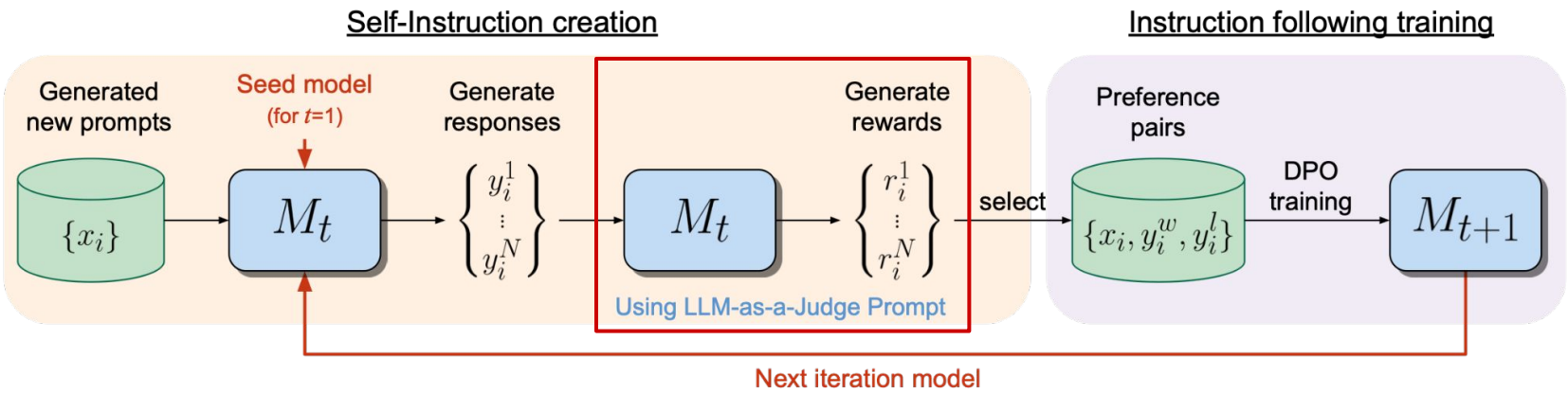
# Our approach

- Self-rewarding LMs



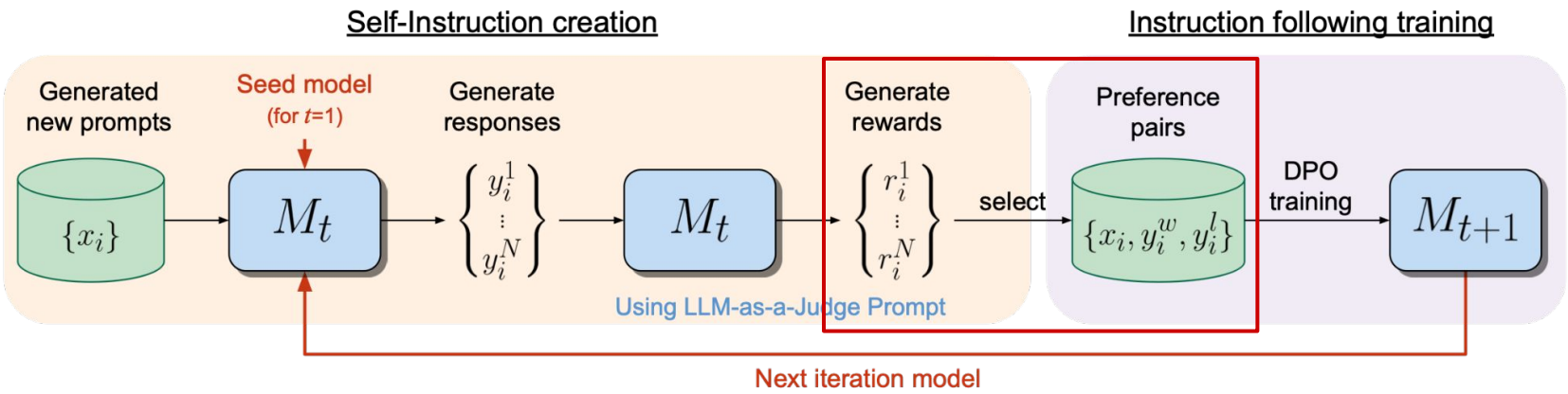
# Our approach

- Self-rewarding LMs



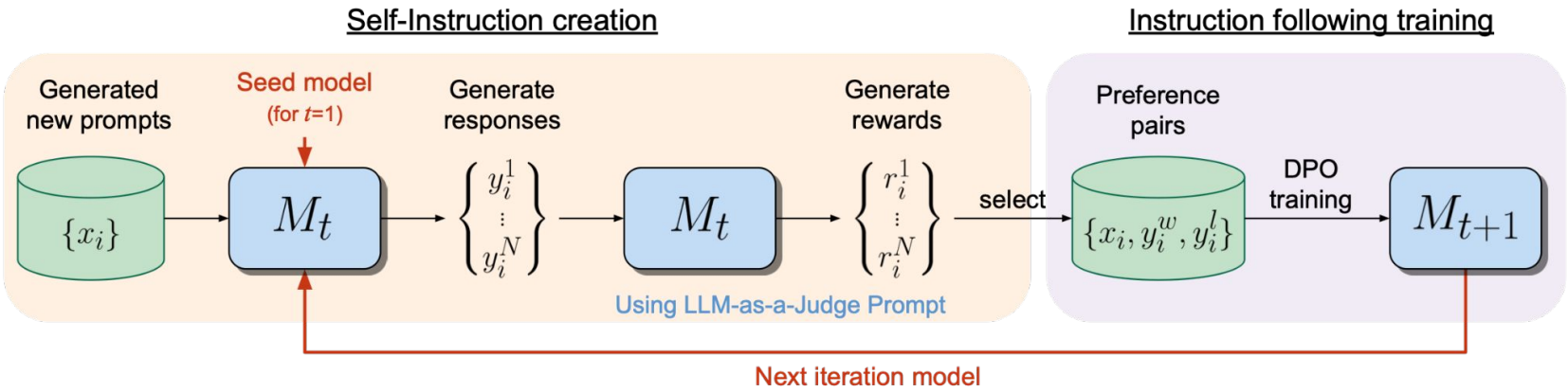
# Our approach

- Self-rewarding LMs



# Our approach

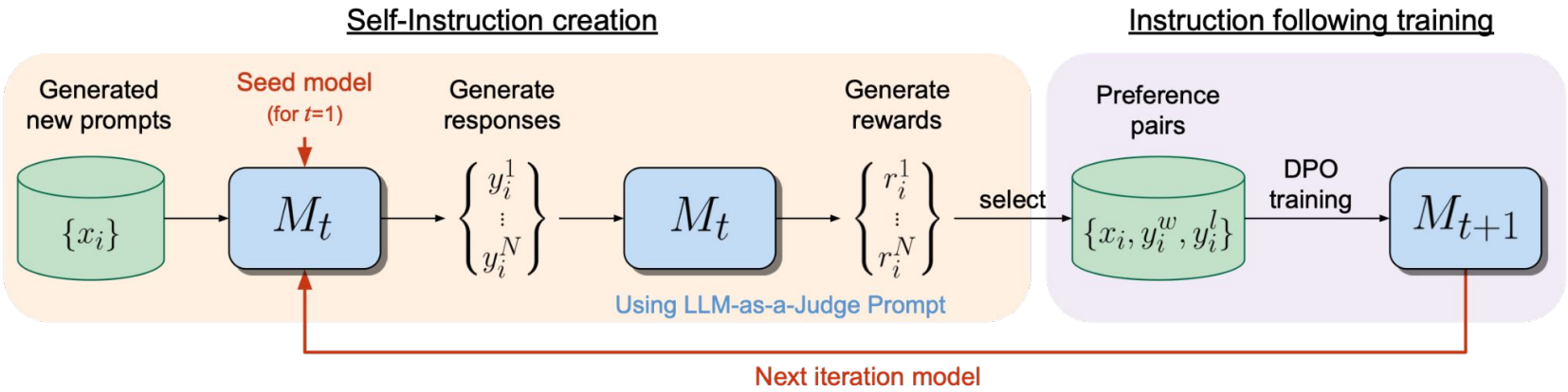
- Self-rewarding LMs




- Hopefully, the model can get better in terms of **both instruction following and evaluation capabilities** in each cycle

# Our approach

- Self-rewarding LMs




- Hopefully, the model can get better in terms of **both instruction following and evaluation capabilities** in each cycle

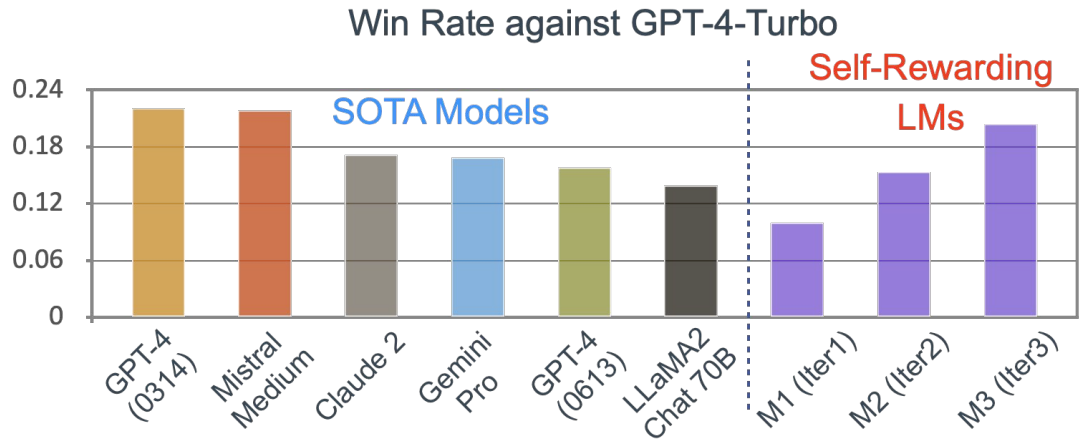
*Empirically, we have shown that this is possible !*

# Experiments

- We start from a Llama2-70b (base) model, aiming to improve it through iterations of training.
  - **Seed Data**: We construct seed data for instruction following tasks and evaluation tasks using OpenAssistant.
  - **Seed Model**: We fine-tune Llama2-70b (base) using the SFT seed data (to give M1).

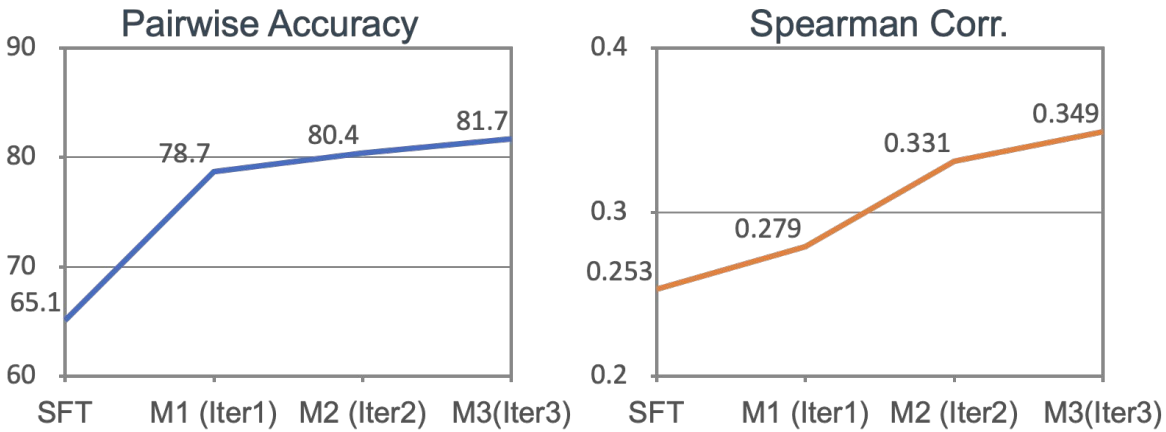
# Experiments

- Instruction following ability on AlpacaEval 2.0
  - **Our model is continuously improved on instruction following tasks through iterative training.**



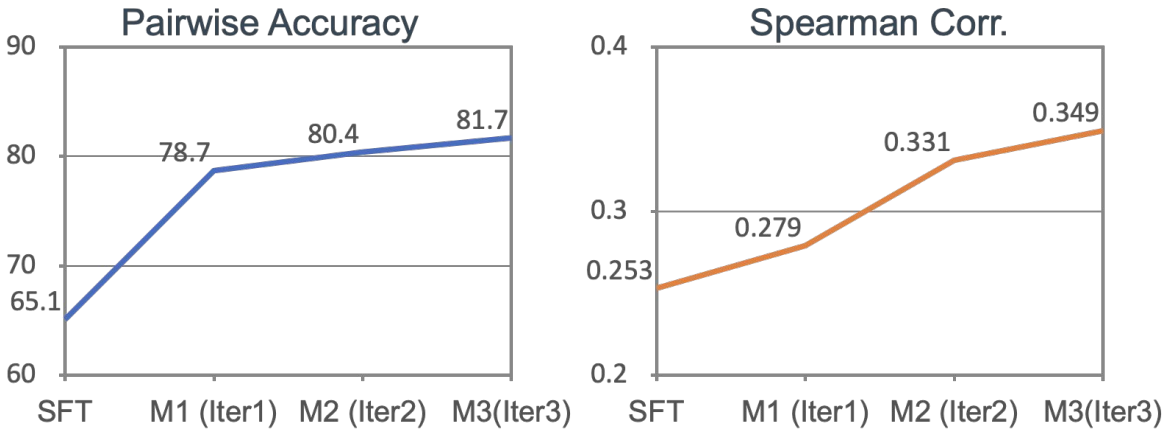
# Experiments

- Reward modeling ability on OpenAssistant validation set
  - **Our model is continuously improved on reward modeling tasks through iterative training.**



# Experiments

- Reward modeling ability on OpenAssistant validation set
  - **Our model is continuously improved on reward modeling tasks through iterative training.**




*Feel free to check out our paper or stop by our poster session (Tue Jul 23 Session 1) to see more results and analysis!*

# Thanks for Listening