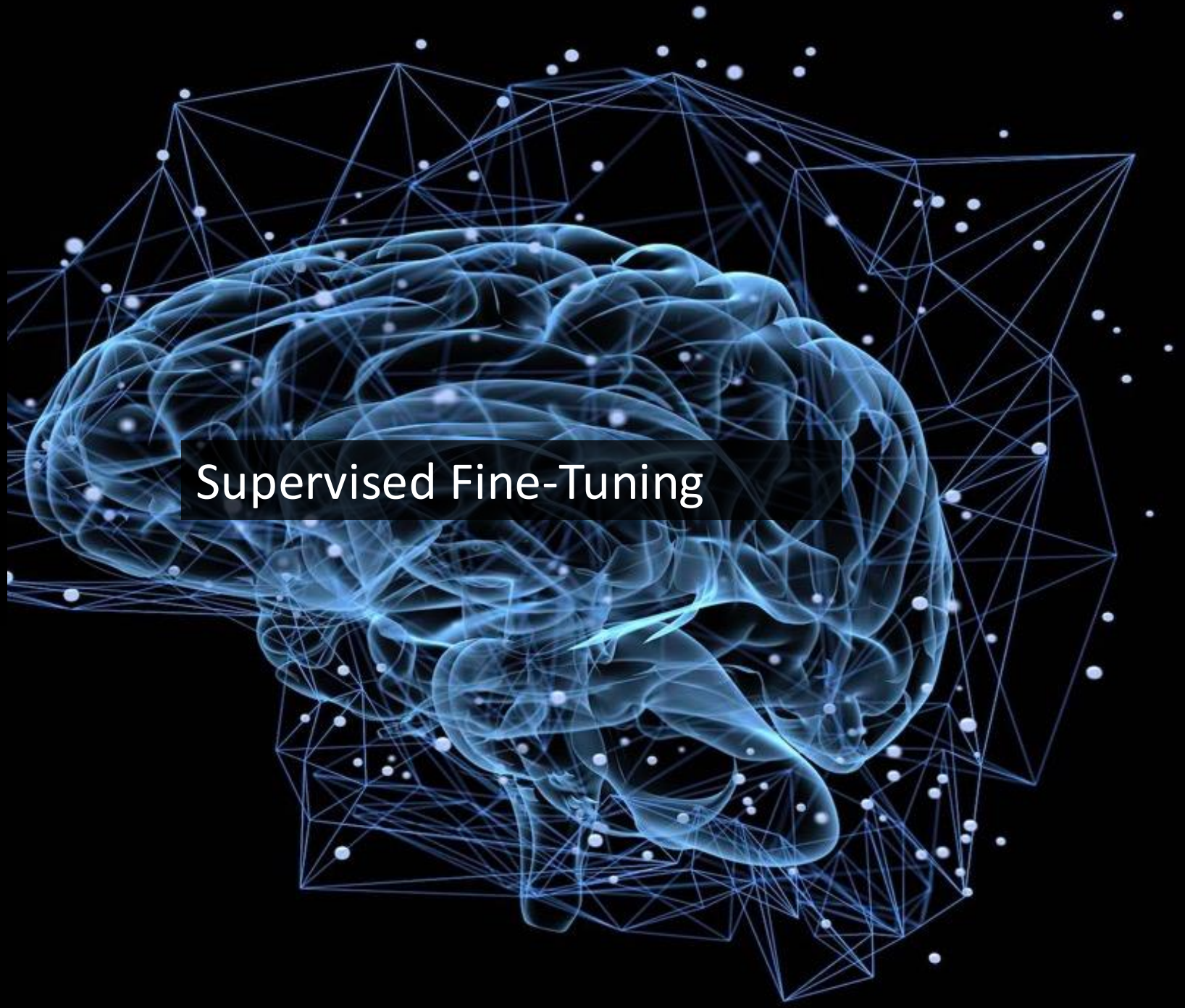
EECS 6895 Adv. Big Data and AI

Lecture 5: Supervised Fine-Tuning and Reinforcement Learning

Prof. Ching-Yung Lin

Columbia University

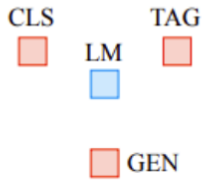
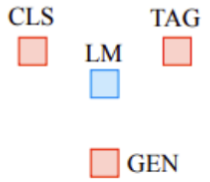
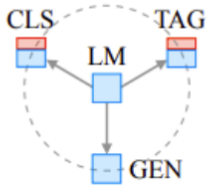
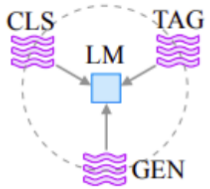
February 18th, 2025


Supervised Fine-Tuning

Supervised Fine-Tuning

- Supervised Fine-Tuning (SFT) is to fine tune on pretrained models.
- SFT is also called as Prompt Tuning.
- Pretrained models have a lot of general knowledge. But, because its trained goal was to predict the next word, it does not really understand Natural Languages.
- To let model understand human prompt, LLM models needs fine-tuning.
- How to do fine-tuning efficiently is a key issue.

Evolution of NLP Learning Space

Paradigm	Engineering	Task Relation
a. Fully Supervised Learning (Non-Neural Network)	Features (e.g. word identity, part-of-speech, sentence length)	
b. Fully Supervised Learning (Neural Network)	Architecture (e.g. convolutional, recurrent, self-attentional)	
c. Pre-train, Fine-tune	Objective (e.g. masked language modeling, next sentence prediction)	
d. Pre-train, Prompt, Predict	Prompt (e.g. cloze, prefix)	

https://arxiv.org/pdf/2107.13586v1.pdf

Prompt-based Learning

Transformers have revolutionized the way data practitioners build models for natural language processing. training a model from scratch, data scientists can leverage a pre-trained model.

Transfer learning involves the following:

1. Loading the trained model into memory.
2. Freezing the parameters to avoid losing any information they contain during future training rounds.
3. Adding some new trainable layers on top of the frozen layers.
4. Training the new layers on another dataset.

Prompt-based learning?

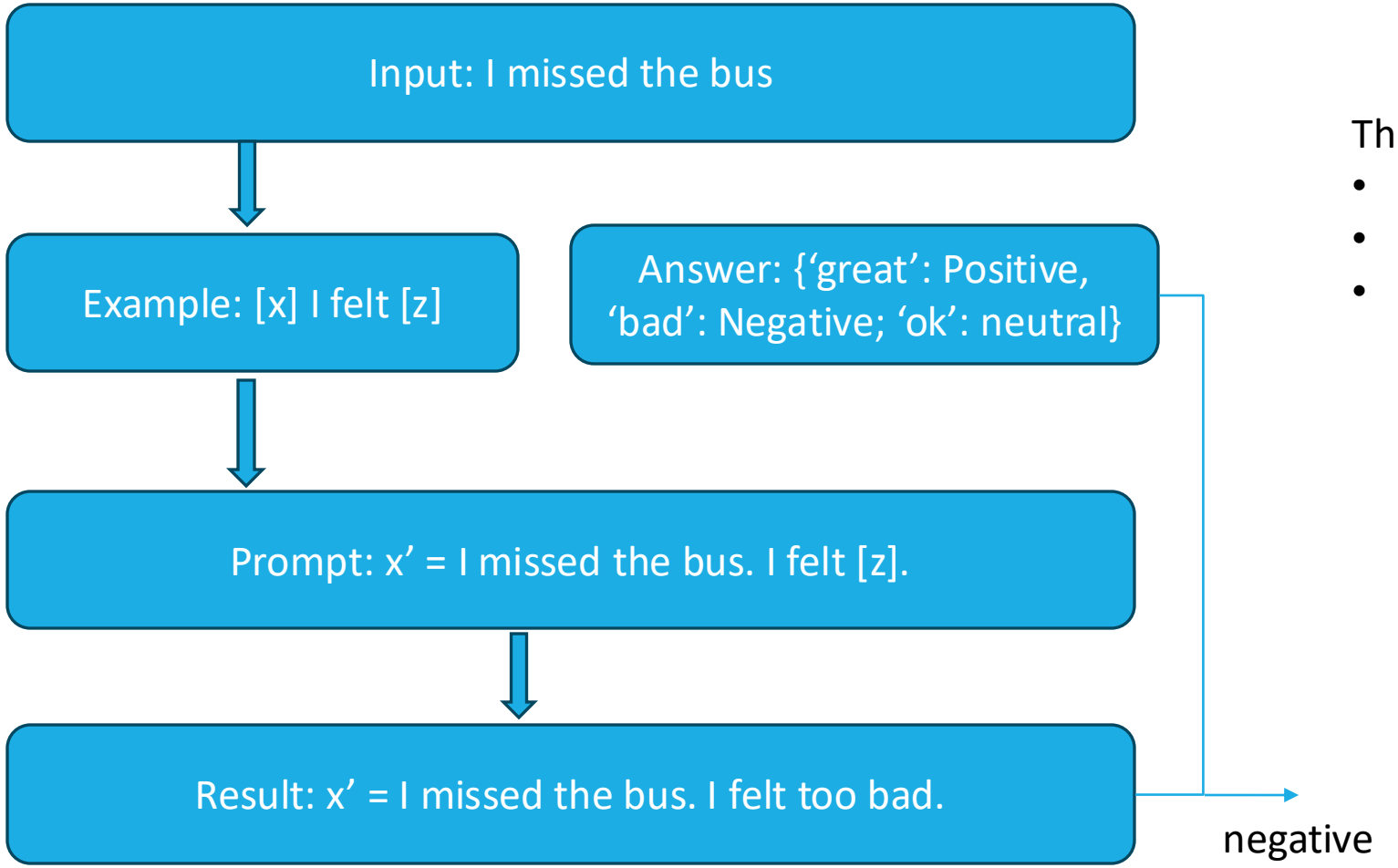
Prompt-based learning is a strategy that machine learning engineers can use to train large language models (LLMs) so the same model can be used for different tasks without re-training.

Prompt content types

- Input (required)
- Context (optional)
- Examples (optional)

https://medium.com/@aniket.umbc/what-is-prompt-based-learning-and-prompt-designing-in-llm-bb4a27251826

Prompt-based Learning



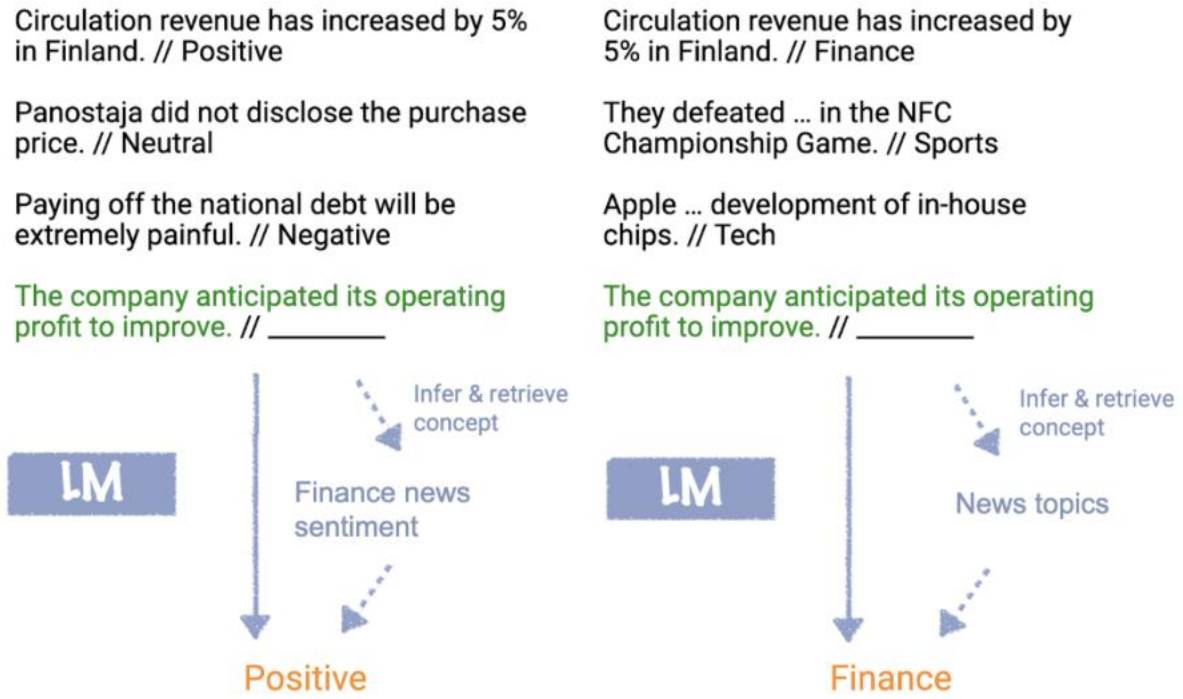
Three Steps:

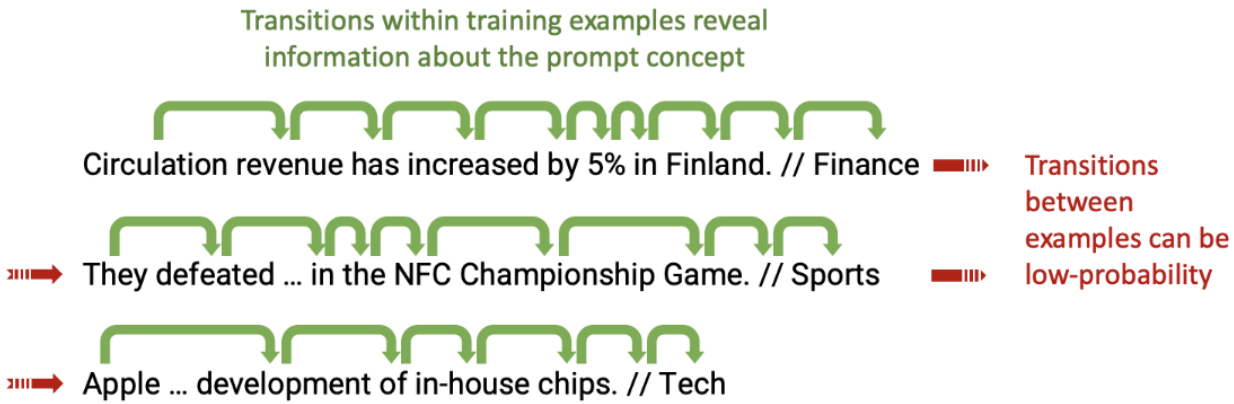
- Prompt Increase
- Answer Searching
- Answer Mapping

In-Context Learning

In-context learning in AI

- A prompt engineering technique that allows LLMs to learn new tasks without fine-tuning
- Uses a few examples embedded in the prompt to guide the model's understanding of the task
- Allows for a more flexible task-solving process
- Based on the idea that models can learn from analogy
- Can be used to solve novel tasks or combined with fine-tuning for more powerful LLMs

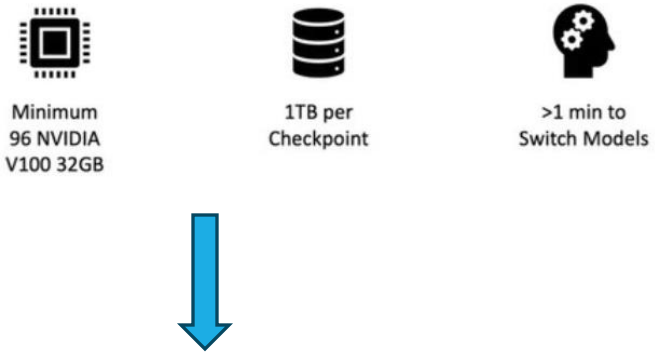




https://ai.stanford.edu/blog/understanding-incontext/

Low-Rank Adaptation of Large Language Models (LoRA)

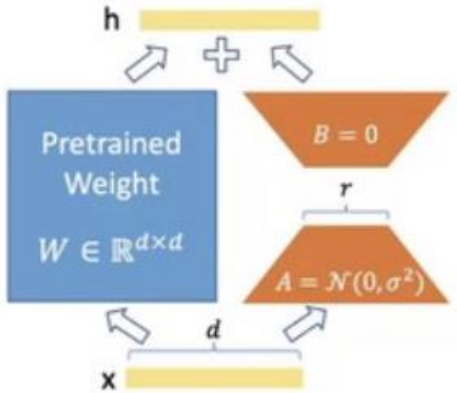
Finetuning GPT-3 is a nightmare



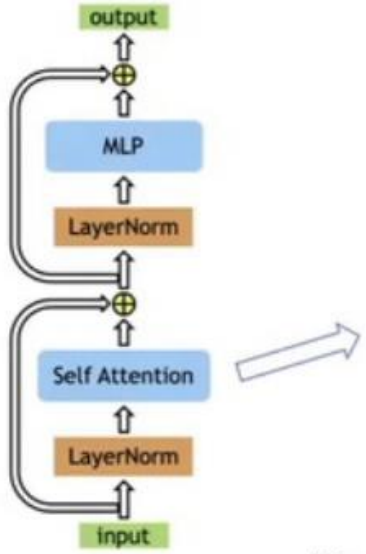
Finetuning GPT-3 is a nightmare

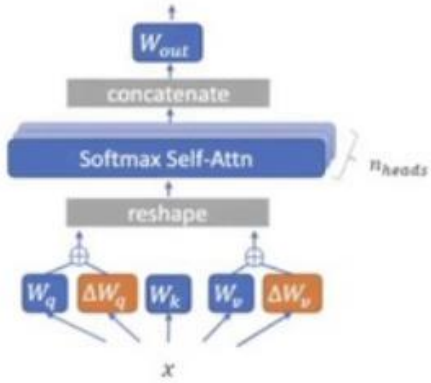
LoRA, ICLR 2022.
https://arxiv.org/pdf/2106.09685
https://iclr.cc/virtual/2022/poster/6319

Low-Rank Adaptation (LoRA)



$h = Wx$





We only apply LoRA to W_q and W_v to reduce # of trainable parameters

Model&Method	# Trainable Parameters	WikiSQL Acc. (%)	MNLI-m Acc. (%)	SAMSum R1/R2/RL
GPT-3 (FT)	175,255.8M	**73.8**	89.5	52.0/28.0/44.5
GPT-3 (BitFit)	14.2M	71.3	91.0	51.3/27.4/43.5
GPT-3 (PreEmbed)	3.2M	63.1	88.6	48.3/24.2/40.5
GPT-3 (PreLayer)	20.2M	70.1	89.5	50.8/27.3/43.5
GPT-3 ($\text{Adapter}^{\text{H}}$)	7.1M	71.9	89.8	53.0/28.9/44.8
GPT-3 ($\text{Adapter}^{\text{H}}$)	40.1M	73.2	**91.5**	53.2/29.0/45.1
GPT-3 (LoRA)	4.7M	73.4	**91.7**	**53.8/29.8/45.9**
GPT-3 (LoRA)	37.7M	**74.0**	**91.6**	53.4/29.2/45.1

Model Context Window Expansion

Limitation of context window size allowed. E.g., Llama 2023: 2048 words.

In order to satisfy the long context needs, the following three methods are usually used:

- Increase the length of window
- Position Encoding
- Extrapolation

Extrapolation Position Encoding

- Naïve Learned Position Embedding – used by BERT
- Sinusoidal Position Embedding
- RoPE
- T5 Bias Position Embedding
- ALiBi Algorithm

Instruction Training

- Manual Instruction Datasets and Automatic Instruction Datasets

- LIMA (Less Is More) as an example:
  - LIMA data sources:
    - High-Quality QA community
    - Super-Natural Instruction dataset
    - Manual annotated QA

LIMA (1,000) Examples

You are a scientist who just invented a time machine. Where do you travel first?

As a scientist, I would travel back in time to the year 1969 and watch the Apollo 11 mission land on Moon.

The Apollo 11 Mission was the first space flight that landed humans on the Moon. It was an incredibly important moment in history as it marked mankind's first step towards exploring other planets. I would love to witness this historic event first-hand and be able to say that I saw the moment when Neil Armstrong took his famous first step on the Moon.

Could you turn it into a fictional essay?

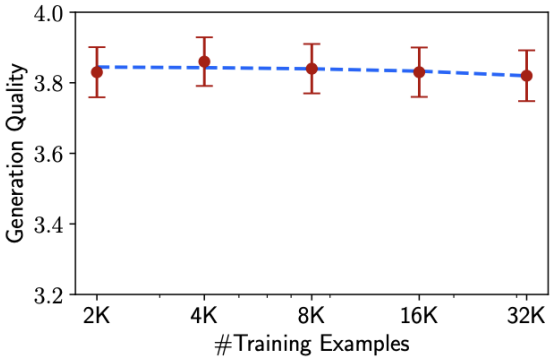
As a scientist, I would travel back in time to the year 1969 and witness the Apollo 11 mission land on the Moon.

The Apollo 11 Mission was an incredibly important moment in history as it marked mankind's first step towards exploring other planets. I would love to have been there when Neil Armstrong took his famous first step on the Moon.

Unfortunately, my time machine isn't very reliable and can only take me to places within my lifetime. This means that I could go forward in time as well as back. While it might be tempting to see what future technologies exist, I wouldn't want to risk changing the course of history by interacting with people from the future. Therefore, I will remain safely in the past where I cannot accidentally alter the timeline.

Can you create a title for the essay?

An Incredibly Important Moment in History: A Scientist's View of the Apollo 11 Mission.



LIMA: Less is More for alignment. 2023.
https://arxiv.org/pdf/2305.11206

Open Data set for Instruction Training

Instruction Dataset	Size	Language	Method	Type
Super-Natural Instruction2	5M	Multi-language	Manual	NLP
Flan2021	440K	English	Manual	NLP
pCLUE	1.2M	Chinese	Manual	NLP
Open Assistant Conversations	161K	Multi-language	Manual	General Conversation
Dolly	15K	English	Manual	General Conversation
LIMA	1K	English	Manual	General Conversation
Self-Instruct	52K	English	Automatic	General Conversation
Alapaca_data	52K	English	Automatic	General Conversation
BELLE	1.5M	Chinese	Automatic	General Conversation

Open Instruction Datasets - I

- Super-Natural Instructions :
  - made by Allen Institute for AI.
  - Including 55 languages
  - 1616 NLP tasks
  - 5M task examples, including 76 task types.
  - Each task includes both instruction and task examples

- Flan2021:
  - Built and released by Google.
  - Converted 62 generally used NLP standards to input-output examples.

- pCLUE:
  - Bilt and made by CLUEbenchmark
  - 9 Chinese NLP datasets, including single classification tnews, single classification iflytek, natural langue, word matching, keyword classification , reading understaning, reading understanding, and reaading understanding.

Open Instruction Datasets - II

- OpenAssistant Conversations:
  - Distributed by LAION.
  - 35 languages
  - 161443 conversations in 66497 conversation trees. Diverse
- Dolly:
  - Distributed by Databricks. 15K instructions.
  - Including 7 areas: open QA, closed QA, info extraction, abstract, brainstorming, classification, and writing

- LIMA:
  - Distributed by Meta
  - 1000 high quality and diverse instructions.

- Self-Instruct data:
  - Using GPT-3. 5.2M

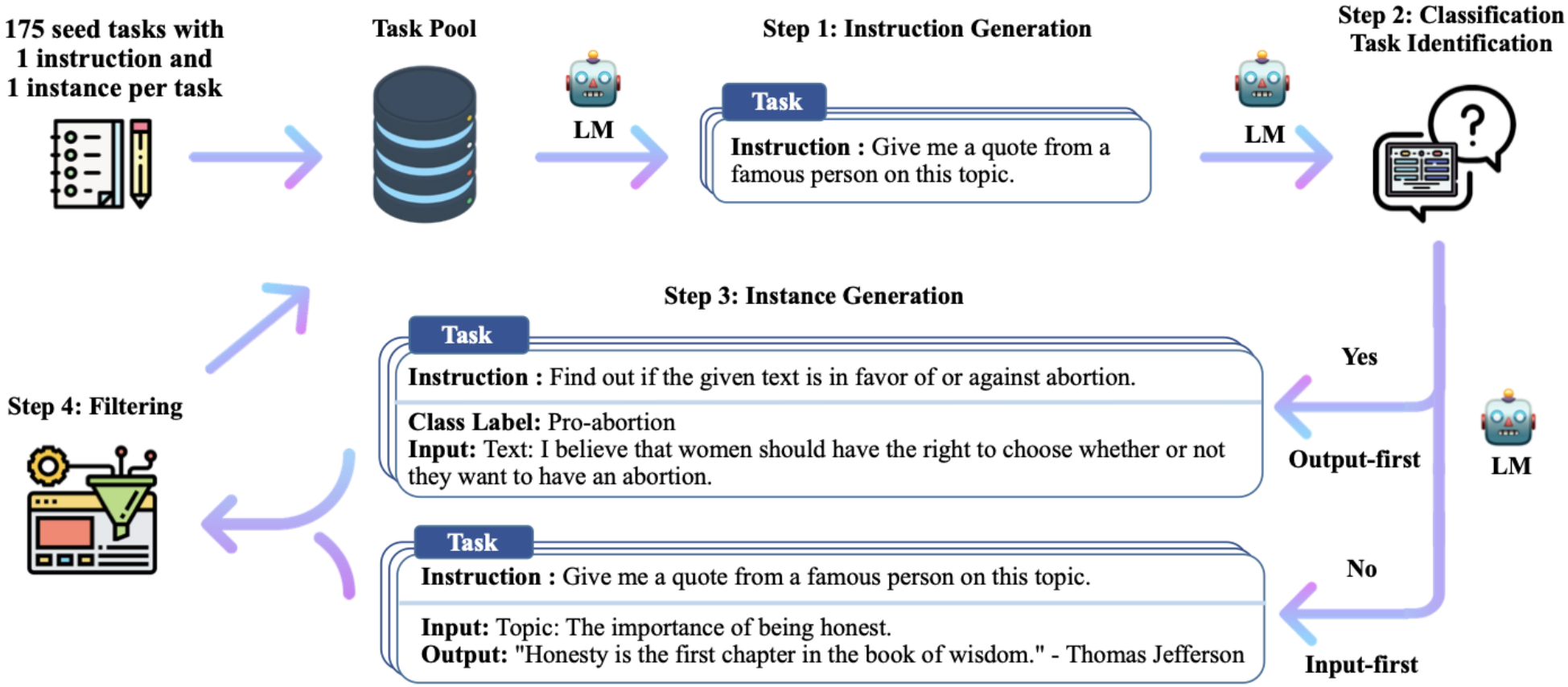
- Alpaca_data was distributed by Stanford, using Self-Instruct, using text-davinci-003 model. 5.2M instruction data.

- BELLE was distributed by shell, using Self-Instruct, using text-davinci-003 model. 1.5M instruction data

Automatic Instructions Generation

Self-Instruct dataset was created by iterative processes.

- Step 1: Instruction Generation
- Step 2: Classification Task Identification
- Step 3: Generative Mission Input and Output
- Step 4: Filtering low-quality data



SELF-INSTRUCT: Aligning Language Models with Self-Generated Instructions, 2023
https://arxiv.org/pdf/2212.10560

Self-Instruct Step 1: Instruction Generation

- Manually construction of 175 tasks as seed instructions.
- Using Bootstrapping to create new instructions.
- Each time, sampling 8 instructions (6 from the manual seed instructions and 2 from iterative models), and putting them as a context example.
- Send to GPT-3 to generate more new tasks.
- Iterate the above process until the model stops generating, reaching the length of model, or generating the “Task 16” token.

```
Come up with a series of tasks:

Task 1:  {instruction for existing task 1}
Task 2:  {instruction for existing task 2}
Task 3:  {instruction for existing task 3}
Task 4:  {instruction for existing task 4}
Task 5:  {instruction for existing task 5}
Task 6:  {instruction for existing task 6}
Task 7:  {instruction for existing task 7}
Task 8:  {instruction for existing task 8}
Task 9:
```

Self-Instruct Step 2: Classification

- We need two different approaches for classification and non-classification tasks.
- We next identify whether the generated instruction represents a classification task or not.
- We prompt the LM in a few-shot way to determine this, using 12 classification instructions and 19 non-classification instructions from the seed tasks.
- The prompting template is shown here.

```
Can the following task be regarded as a classification task with finite output labels?

Task:  Given my personality and the job, tell me if I would be suitable.
Is it classification?  Yes

Task:  Give me an example of a time when you had to use your sense of humor.
Is it classification?  No

Task:  Replace the placeholders in the given text with appropriate named entities.
Is it classification?  No

Task:  Fact checking - tell me if the statement is true, false, or unknown, based on your
knowledge and common sense.
Is it classification?  Yes

Task:  Return the SSN number for the person.
Is it classification?  No

Task:  Detect if the Reddit thread contains hate speech.
Is it classification?  Yes

Task:  Analyze the sentences below to identify biases.
Is it classification?  No

Task:  Select the longest sentence in terms of the number of words in the paragraph, output
the sentence index.
Is it classification?  Yes

Task:  Find out the toxic word or phrase in the sentence.
Is it classification?  No

Task:  Rank these countries by their population.
Is it classification?  No

Task:  You are provided with a news article, and you need to identify all the categories that
this article belongs to.  Possible categories include:  Music, Sports, Politics, Tech, Finance,
Basketball, Soccer, Tennis, Entertainment, Digital Game, World News.  Output its categories one
by one, seperated by comma.
Is it classification?  Yes
```

Self-Instruct Step 3: Instance Generation

Given the instructions and their task type, we generate instances for each instruction independently. This is challenging because it requires the model to understand what the target task is, based on the instruction, figure out What additional input fields are needed and generate them, and finally complete the task by producing the output.

We found that pretrained LMs can Achieve this to a large extent when prompted with instruction-input-output in-context examples from other tasks.

```
Come up with examples for the following tasks.  Try to generate multiple examples when possible.
If the task doesn't require additional input, you can generate the output directly.

Task:  Which exercises are best for reducing belly fat at home?
Output:
- Lying Leg Raises
- Leg In And Out
- Plank
- Side Plank
- Sit-ups

Task:  Extract all the country names in the paragraph, list them separated by commas.
Example 1
Paragraph:  Dr.  No is the sixth novel by the English author Ian Fleming to feature his British
Secret Service agent James Bond.  Written at Fleming's Goldeneye estate in Jamaica, it was
first published in the United Kingdom by Jonathan Cape in 1958.  In the novel Bond looks into
the disappearance in Jamaica of two fellow MI6 operatives who had been investigating Doctor
No.  Bond travels to No's Caribbean island and meets Honeychile Rider, who is there to collect
shells.  They are captured and taken to a luxurious facility carved into a mountain.  The
character of Doctor No, the son of a German missionary and a Chinese woman, was influenced by
Sax Rohmer's Fu Manchu stories.  Dr.  No was the first of Fleming's novels to face widespread
negative reviews in Britain, but it was received more favourably in the United States.
Output:  English, British, Jamaica, the United Kingdom, German, Chinese, Britain, the United
States.

Task:  Converting 85 F to Celsius.
Output:  85°F = 29.44°C
```

Self-Instruct Step 3: Instance Generation

A natural way to do this is the **Input-first Approach**, where we can ask an LM to come up with the input fields first based on the instruction, and then produce the corresponding output.

This generation order is similar to how models are used to respond to instruction and input, but here with in-context examples from other tasks. The prompting template is shown in Table 7.

```
Task:  Sort the given list ascendingly.
Example 1
List:  [10, 92, 2, 5, -4, 92, 5, 101]
Output:  [-4, 2, 5, 5, 10, 92, 92, 101]
Example 2
Input 2 - List:  [9.99, 10, -5, -1000, 5e6, 999]
Output:  [-1000, -5, 9.99, 10, 999, 5e6]

Task:  Suggest a better and more professional rephrasing of the following sentence.
Example 1
Sentence:  This house is surprisingly not constructed very well, and you probably need more
money to fix it after you buy it.  If you ask me, I would suggest you to consider other
candidates.
Output:  This house does not seem to be constructed well, so you may need to spend more money
to fix it after you purchase it.  I would suggest that you look at other properties.
Example 2
Sentence:  Just so you know, we did an experiment last week and found really surprising results
- language model can improve itself!
Output:  Our experiments last week demonstrated surprising results, proving that the language
model can improve itself.

...

Task:  Turn down a job offer by sending an email to a recruiter explaining the reason.
Output:  Hi [Recruiter],
Thank you so much for the generous offer to join your team.  As we discussed, I've admired the
company for a number of years, and am a proud endorser of its products.  However, after further
consideration of where I currently am in my career, I've decided to accept an offer at another
company.
I would love to stay in touch with you and have already started following you on [Social Media
Platform].  Again, thank you so much for your time and consideration.
Thanks again,
[Your Name]

Task:  {Instruction for the target task}
```

Self-Instruct Step 3: Instance Generation

However, we found that this approach can generate inputs biased toward one label, especially for classification tasks (e.g., for grammar error detection, it usually generates grammatical input). Therefore, we additionally propose an **Output-first Approach** for classification tasks, where we first generate the possible class labels, and then condition the input generation on each class label. The prompting template is shown in Table8.

We apply the output-first approach to the classification tasks identified in the former step, and the input-first approach to the remaining non-classification tasks.

```
Given the classification task definition and the class labels, generate an input that
corresponds to each of the class labels.  If the task doesn't require input, just generate the
correct class label.

Task:  Classify the sentiment of the sentence into positive, negative, or mixed.
Class label:  mixed
Sentence:  I enjoy the flavor of the restaurant but their service is too slow.
Class label:  Positive
Sentence:  I had a great day today.  The weather was beautiful and I spent time with friends.
Class label:  Negative
Sentence:  I was really disappointed by the latest superhero movie.  I would not recommend it.

Task:  Given a dialogue, classify whether the user is satisfied with the service.  You should
respond with "Satisfied" or "Unsatisfied".
Class label:  Satisfied
Dialogue:
- Agent:  Thank you for your feedback.  We will work to improve our service in the future.
- Customer:  I am happy with the service you provided.  Thank you for your help.
Class label:  Unsatisfied
Dialogue:
- Agent:  Sorry that we will cancel your order.  You will get a refund within 7 business days.
- Customer:  oh that takes too long.  I want you to take quicker action on this.

Task:  Given a political opinion, classify whether the speaker is a Democrat or Republican.
Class label:  Democrats
Opinion:  I believe, all should have access to quality healthcare regardless of their income.
Class label:  Republicans
Opinion:  I believe that people should be able to keep more of their hard-earned money and
should not be taxed at high rates.
```

Self-Instruct Step 3: Instance Generation

However, we found that this approach can generate inputs biased toward one label, especially for classification tasks (e.g., for grammar error detection, it usually generates grammatical input). Therefore, we additionally propose an **Output-first Approach** for classification tasks, where we first generate the possible class labels, and then condition the input generation on each class label. The prompting template is shown in Table8.

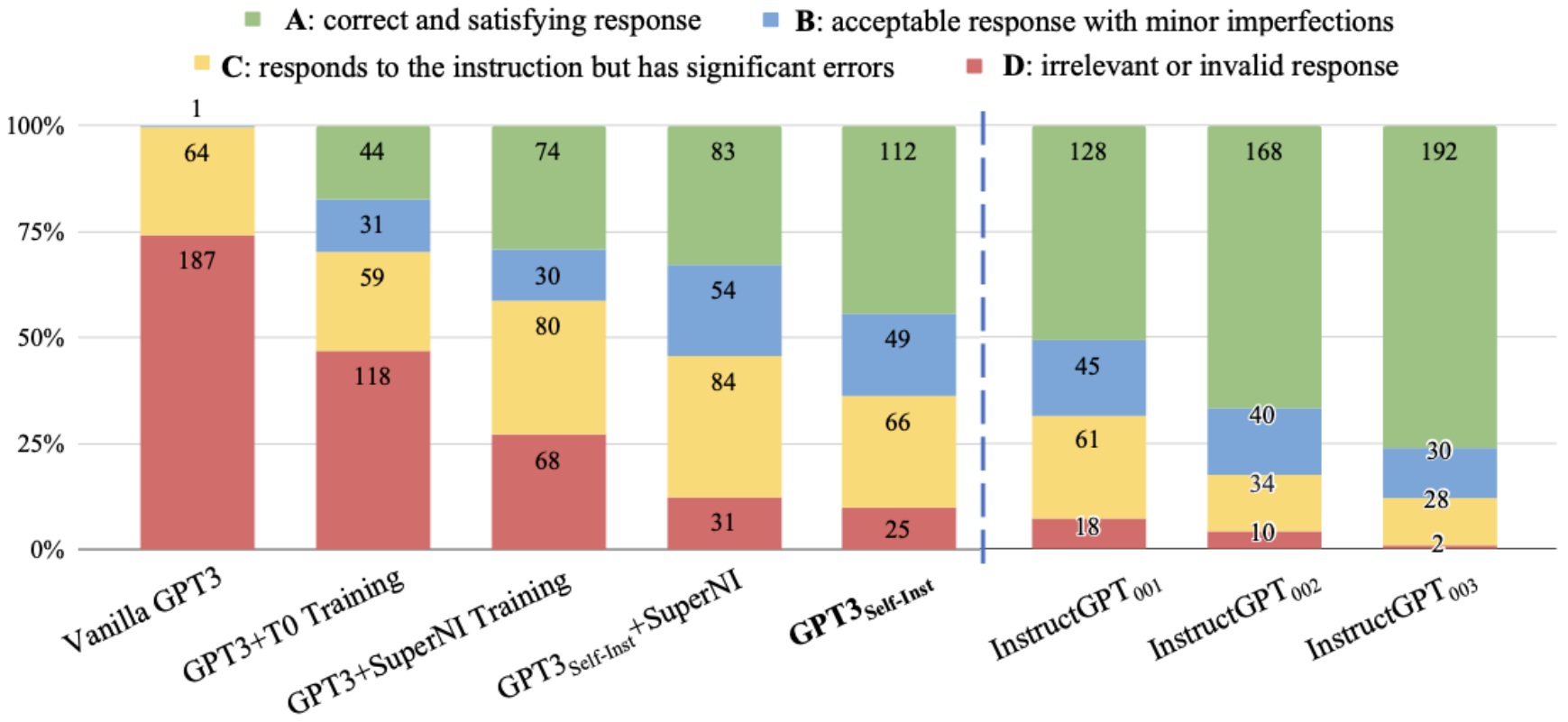
We apply the output-first approach to the classification tasks identified in the former step, and the input-first approach to the remaining non-classification tasks.

```
Task:  Tell me if the following email is a promotion email or not.
Class label:  Promotion
Email:  Check out our amazing new sale!  We've got discounts on all of your favorite products.
Class label:  Not Promotion
Email:  We hope you are doing well.  Let us know if you need any help.

Task:  Detect if the Reddit thread contains hate speech.
Class label:  Hate Speech
Thread:  All people of color are stupid and should not be allowed to vote.
Class label:  Not Hate Speech
Thread:  The best way to cook a steak on the grill.

Task:  Does the document supports the claim?  Answer with "Support" or "Unsupport".
Class label:  Unsupport
Document:  After a record-breaking run that saw mortgage rates plunge to all-time lows and
home prices soar to new highs, the U.S. housing market finally is slowing.  While demand and
price gains are cooling, any correction is likely to be a modest one, housing economists and
analysts say.  No one expects price drops on the scale of the declines experienced during the
Great Recession.
Claim:  The US housing market is going to crash soon.
Class label:  Support
Document:  The U.S. housing market is showing signs of strain, with home sales and prices
slowing in many areas.  Mortgage rates have risen sharply in recent months, and the number
of homes for sale is increasing.  This could be the beginning of a larger downturn, with some
economists predicting a potential housing crash in the near future.
Claim:  The US housing market is going to crash soon.

...

Task:  Which of the following is not an input type?  (a) number (b) date (c) phone number (d)
email address (e) all of these are valid inputs.
Class label:  (e)

Task:  {instruction for the target task}
```

Self-Instruct Step 4: Filtering low-quality data

- To encourage diversity, a new instruction is added to the task pool only when its ROUGE-L similarity with any existing instruction is less than 0.7.
- We also exclude instructions that contain some specific keywords (e.g., image, picture, graph) that usually can not be processed by LMs.
- When generating new instances for each instruction, we filter out instances that are exactly the same or those with the same input but different outputs.
- Invalid generations are identified and filtered out based on heuristics (e.g., instruction is too long or too short, instance output is a repetition of the input).

Self-Instruct Performance Comparisons



DeepSpeed-Chat core functions

- **Easy-to-use Training and Inference Experience for ChatGPT Like Models**:
  - A single script capable of taking a pre-trained Huggingface model
  - Running it through all three steps of InstructGPT training
  - Using DeepSpeed-RLHF system
  - Producing your very own ChatGPT like model.
  - An inference API for testing conversation-style interactions after the model is trained.

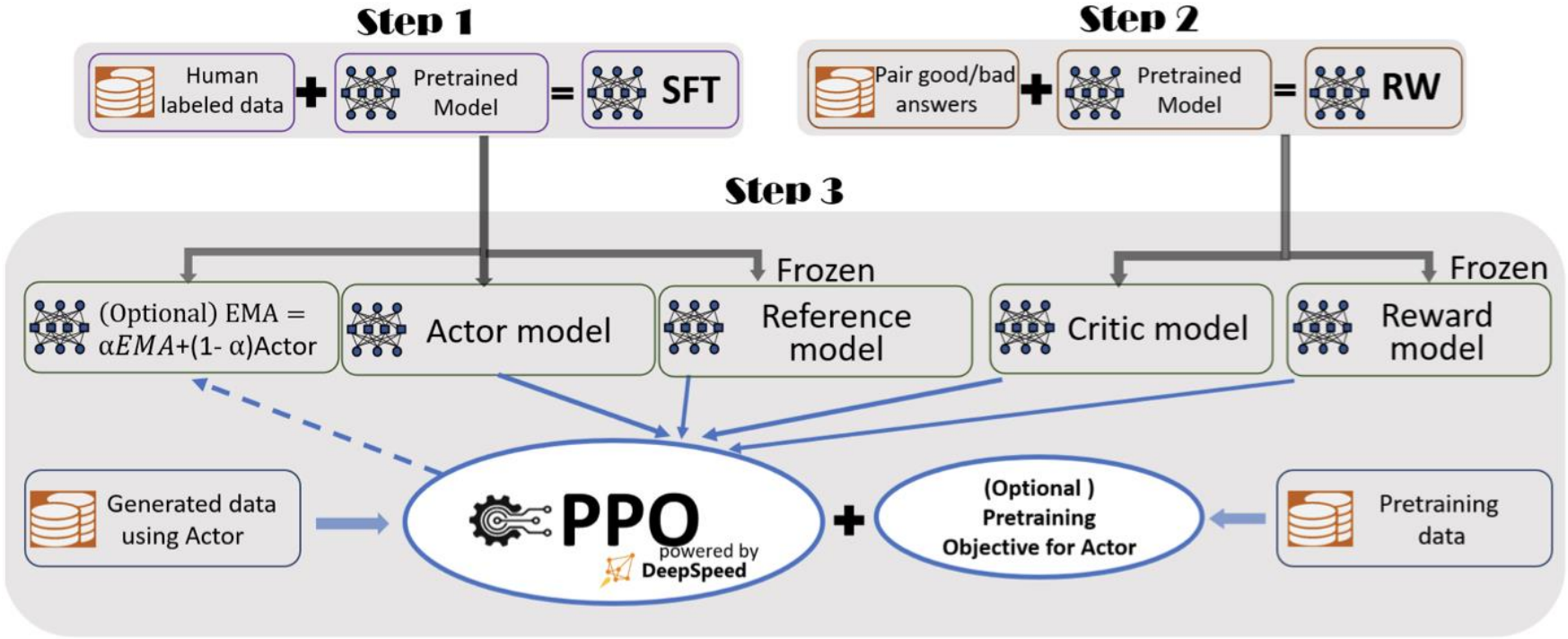
- **DeepSpeed-RLHF Pipeline:** DeepSpeed-RLHF pipeline primarily replicates the training pipeline from the InstructGPT paper with careful attention to ensure completeness and one-to-one correspondence with the three-steps that includes:
  - a) Supervised Fine-tuning (SFT),
  - b) Reward Model Fine-tuning and
  - c) Reinforcement Learning with Human Feedback (RLHF).

  Data abstraction and blending capabilities to enable training with multiple data sources.

DeepSpeed-Chat core functions

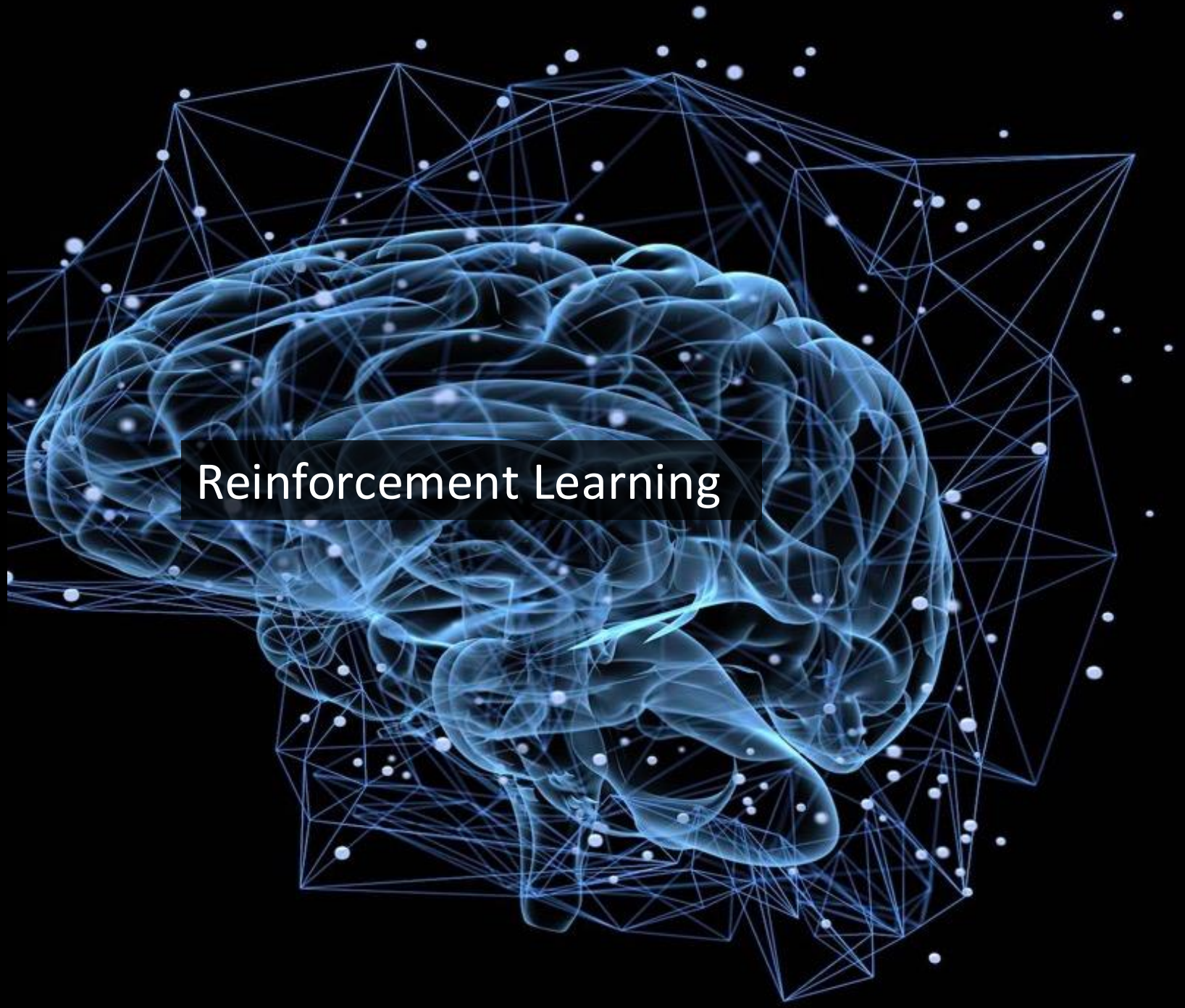
- **DeepSpeed-RLHF System:**
  - A robust and sophisticated RLHF system that combines the training and inference prowess of DeepSpeed into single unified Hybrid Engine (DeepSpeed-HE) for RLHF.
  - The Hybrid-Engine is capable of seamlessly transitioning between inference and training modes within RLHF,
  - Allowing it to leverage various optimizations from DeepSpeed-Inference such as tensor-parallelism and high-performance transformer kernels for generation,
  - Benefiting from the multitude of ZeRO- and LoRA-based memory optimization strategies for RL training.
  - DeepSpeed-HE is also aware of the full RLHF pipeline
  - Allowing it to make optimal decisions in terms of memory management and data movement across different phases of RLHF.

DeepSpeed-Chat SFT



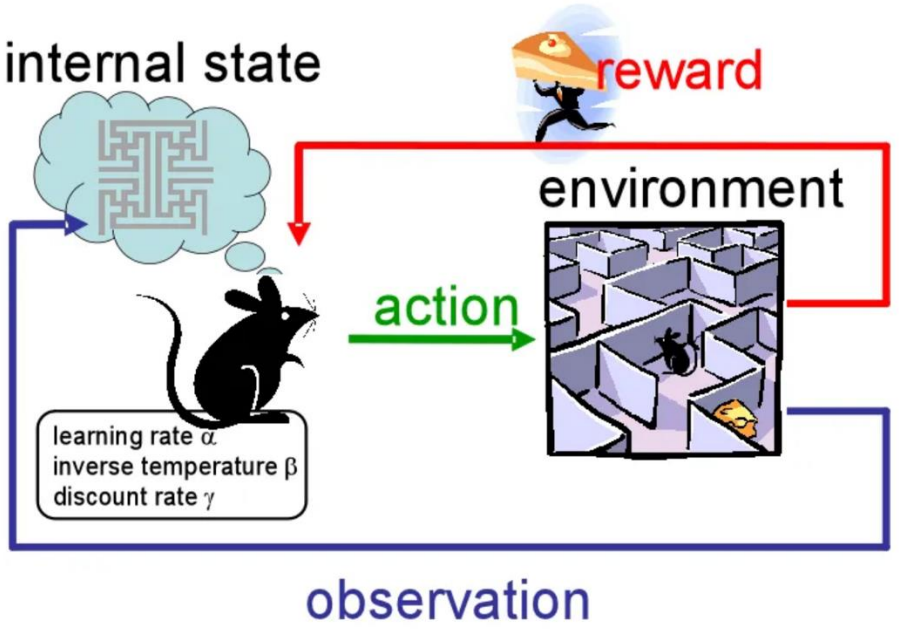
- **Step 1:** Supervised finetuning (SFT), where human responses to various queries are carefully selected to finetune the pretrained language models.
- **Step 2:** Reward model finetuning, where a separate (usually smaller than the SFT) model (RW) is trained with a dataset that has human-provided rankings of multiple answers to the same query.
- **Step 3:** RLHF training, where the SFT model is further finetuned with the reward feedback from the RW model using the Proximal Policy Optimization (PPO) algorithm.

DeepSpeed-Chat: Easy, Fast and Affordable RLHF Training of ChatGPT-like Models at All Scales, 2023 https://arxiv.org/pdf/2308.01320


Reinforcement Learning

Reinforcement Learning

- Through SFT, LLM can follow human instructions to achieve various tasks.
- However, SFT needs a large amount of instructions and their standard responses, which take a long time to achieve.
- SFT usually uses Cross-Entropy as the loss function to adjust parameters to make the model output the same as the standard answers ➔ Model cannot handle the variety of natural language and cannot handle subtle changes.
- RL makes optimization at the high-quality answers.
- RL does not rely on high-quality examples from human.
- RL model judges from the quality of the results.
- Model can generate multiple answers. RL ranks them.
- Model uses feedback to learn.
- RL is important to generative tasks.



https://becominghuman.ai/the-very-basics-of-reinforcement-learning-154f28a79071

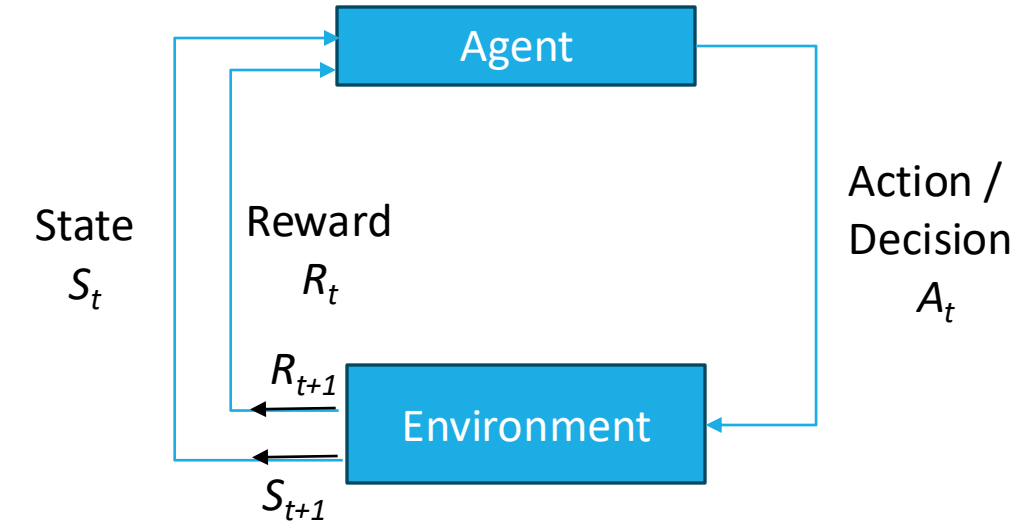
Summary of Reinforcement Learning

Takes training artificial dogs to catch frisbee as an example:

1. Agent and Environment:
   - Robot Dog is an Agent.
   - It makes Decision and Take Action.
   - Frisbee's flying path, speed, and other factors is the Environment.

2. State, Action, and Reward:
   - Every time the robot dos tries to catch frisbee is to evaluate the State (which includes the frisbee's location, speed, etc.)
   - Based on those information, the robot dog takes Action, such as jumping, running, or staying.
   - There is a Reward every time the robot dog executes.

3. Policy and Value:
   - In learning, the robot dog is learning a Policy.
   - Policy can be seen as rules during some states.
   - Agent tries to estimate the Value which predicts the rewards based on Action.



A Robot Dog made by Boston Dynamics

Key Differences between Reinforcement Learning and Supervised Learning

Take traveling as an example:

- **Before travel – data sources:**
  - Supervised Learning: using a travel guide, which identifies all spots, restaurants, and transportations.
  - Reinforcement Learning: no map and no guide, just have a goal. For instance, trying to find a restaurant or museum. Full of exploration

- **Guidance – feedback mechanism:**
  - Supervised Learning: whenever the traveler got lost, someone would tell them whether they are in the right direction.
  - Reinforcement Learning: no one would tell traveler how to go. Only tell them the result is good or bad. For instance, he/she walks into a restaurant and only know whether this restaurant fits AFTER testing.

- **Destination:**
  - Supervised Learning: knows the goals and answers. Just like finishing all the hot spots in the travel guide.
  - Reinforcement Learning: goal is to learn how to effective moving in the city. Finding optimal path, food, living, entertaining, etc.

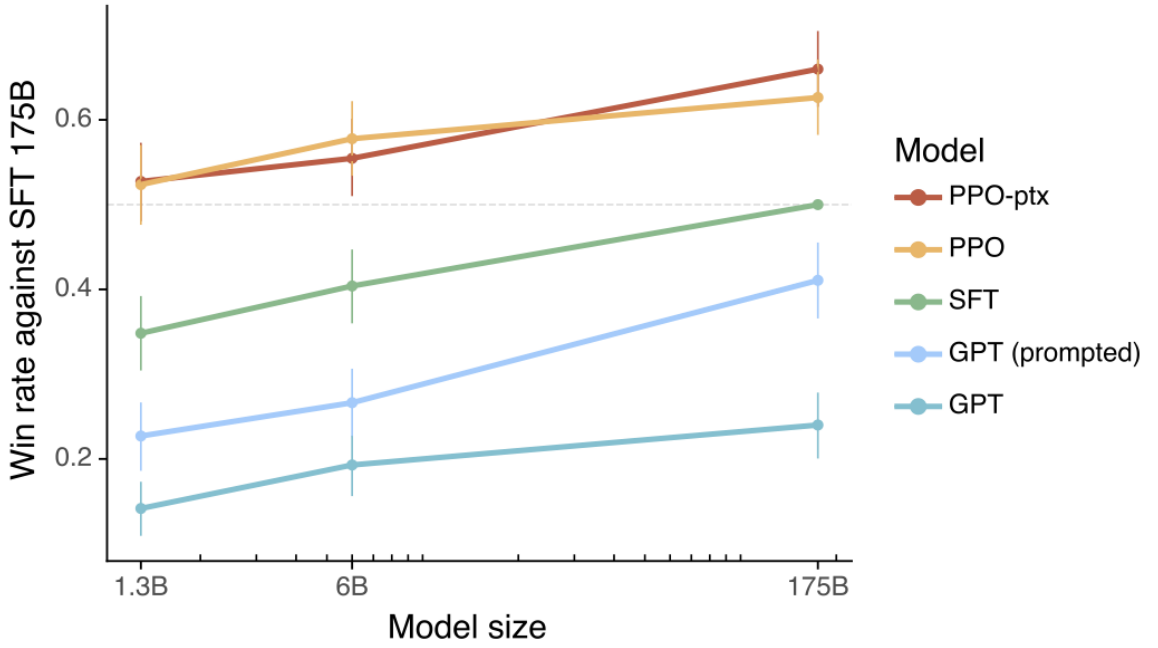
Advantages of Reinforcement Learning in LLM

- Reinforcement Learning can better consider the entire feedback than the Supervised Learning.
- Reinforcement Learning can be more easily solving hallucination.
- Reinforcement Learning can better handle multi-round conversation.

Reinforcement Learning from Human Feedback (RLHF)

- RLHF is considered the best mechanism to achieve 3H – Helpfulness, Honesty, and Harmless of AI.

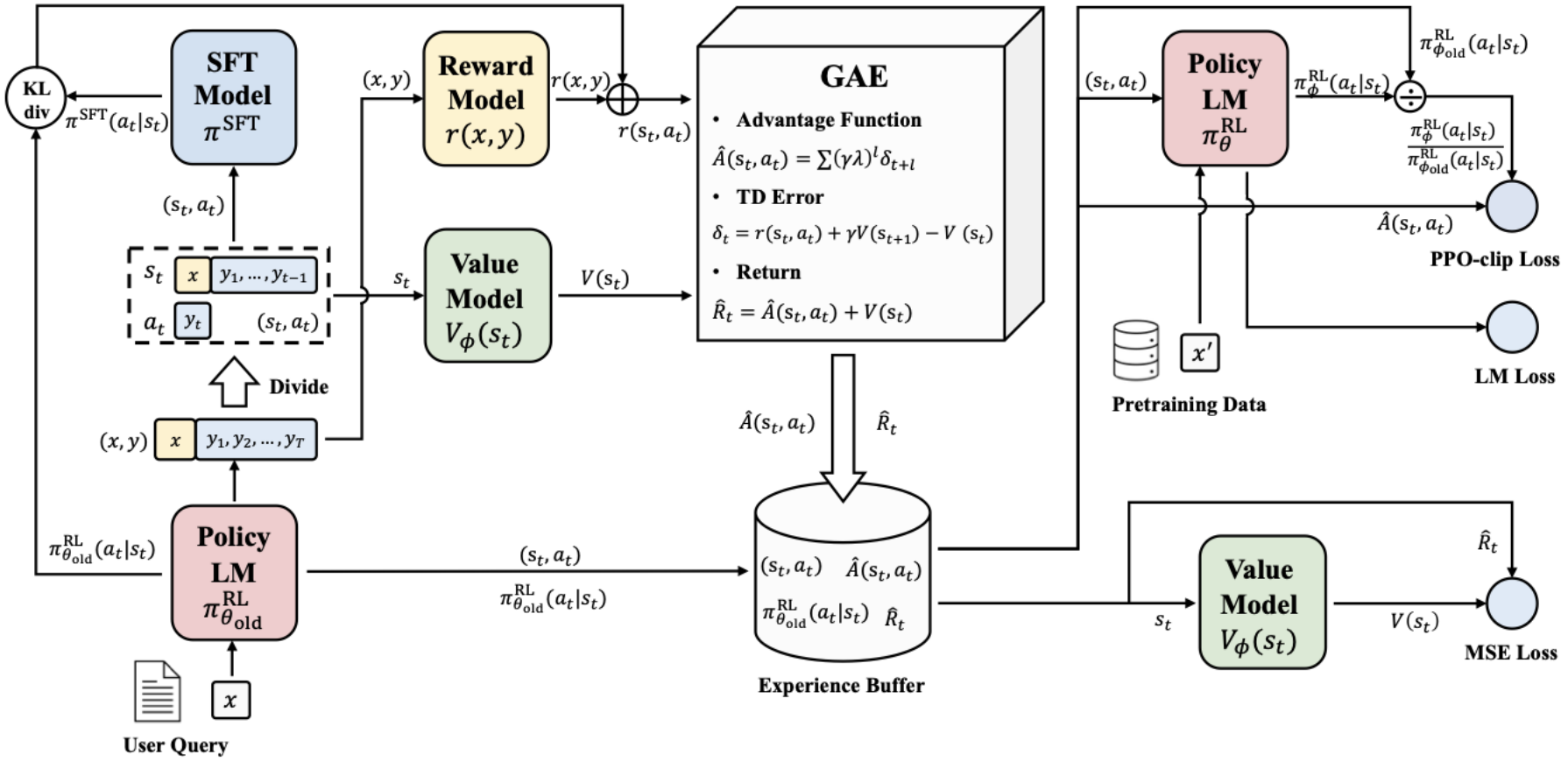
- RLHF usually involves three stages:
  1. Supervised fine-tuning (SFT)
  2. Reward model (RM) training
  3. Proximal Policy Optimization (PPO).



Training language models to follow instructions with human feedback, NIPS, 2022
https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf

Reinforcement Learning from Human Feedback

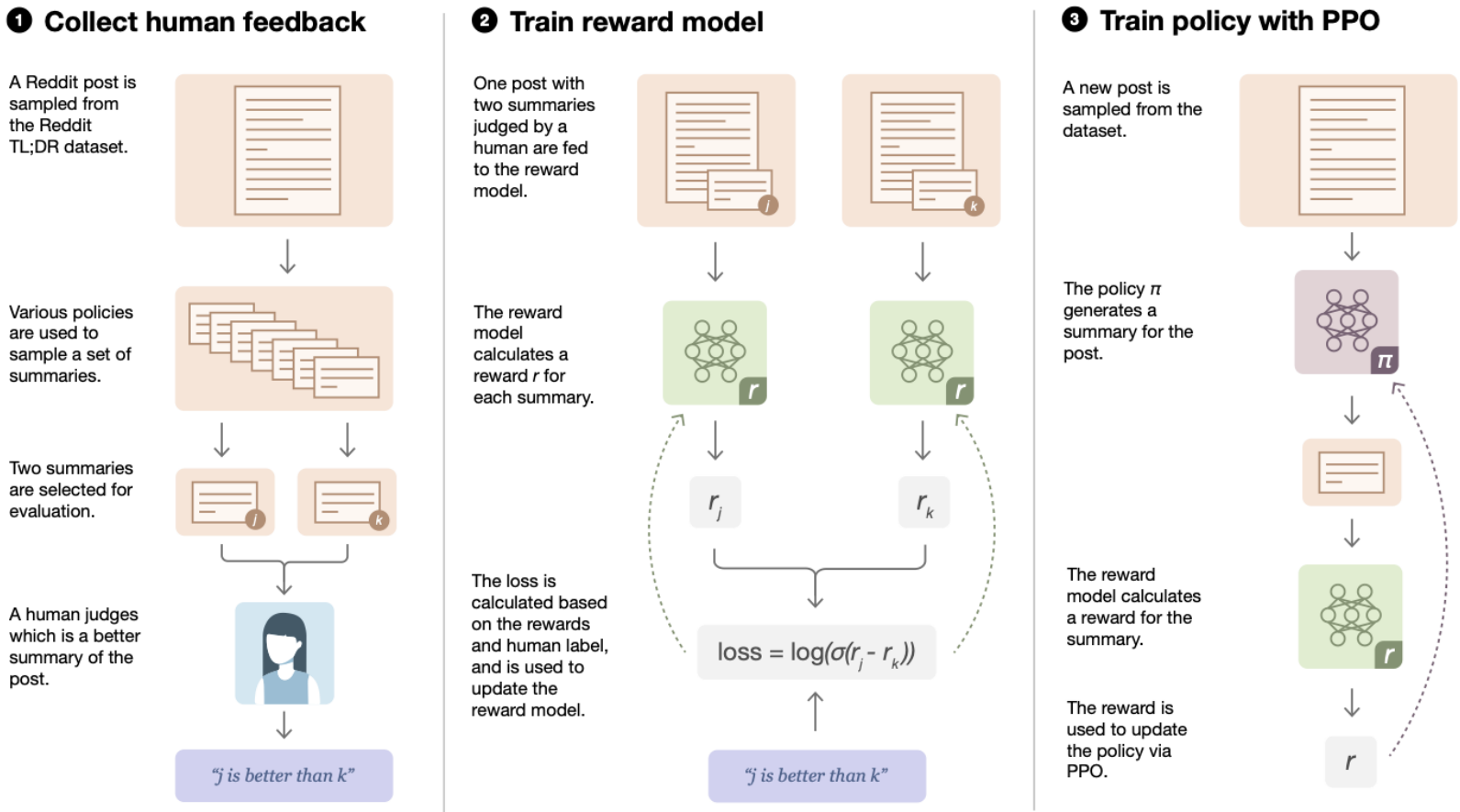
finetuning language models with PPO needs to coordinate four models to work together, i.e., a policy model, a value model, a reward model, and a reference model.

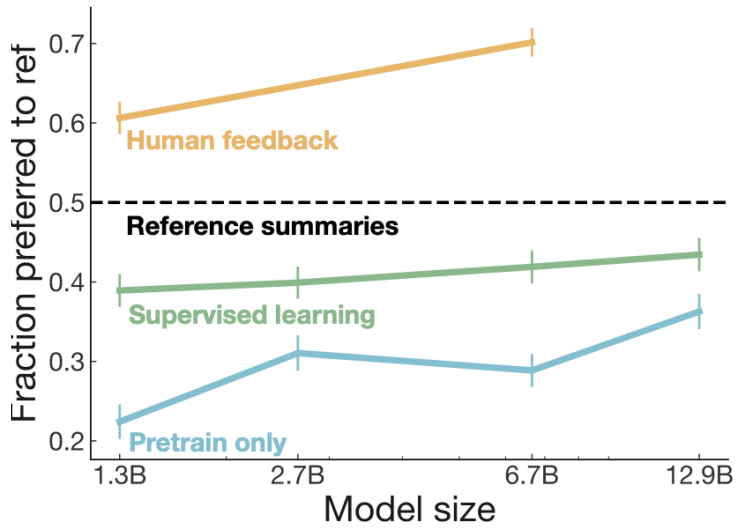


Proximal Policy Optimization

Secrets of RLHF in Large Language Models Part I: PPO, 2023. https://arxiv.org/pdf/2307.04964

Reward Model training



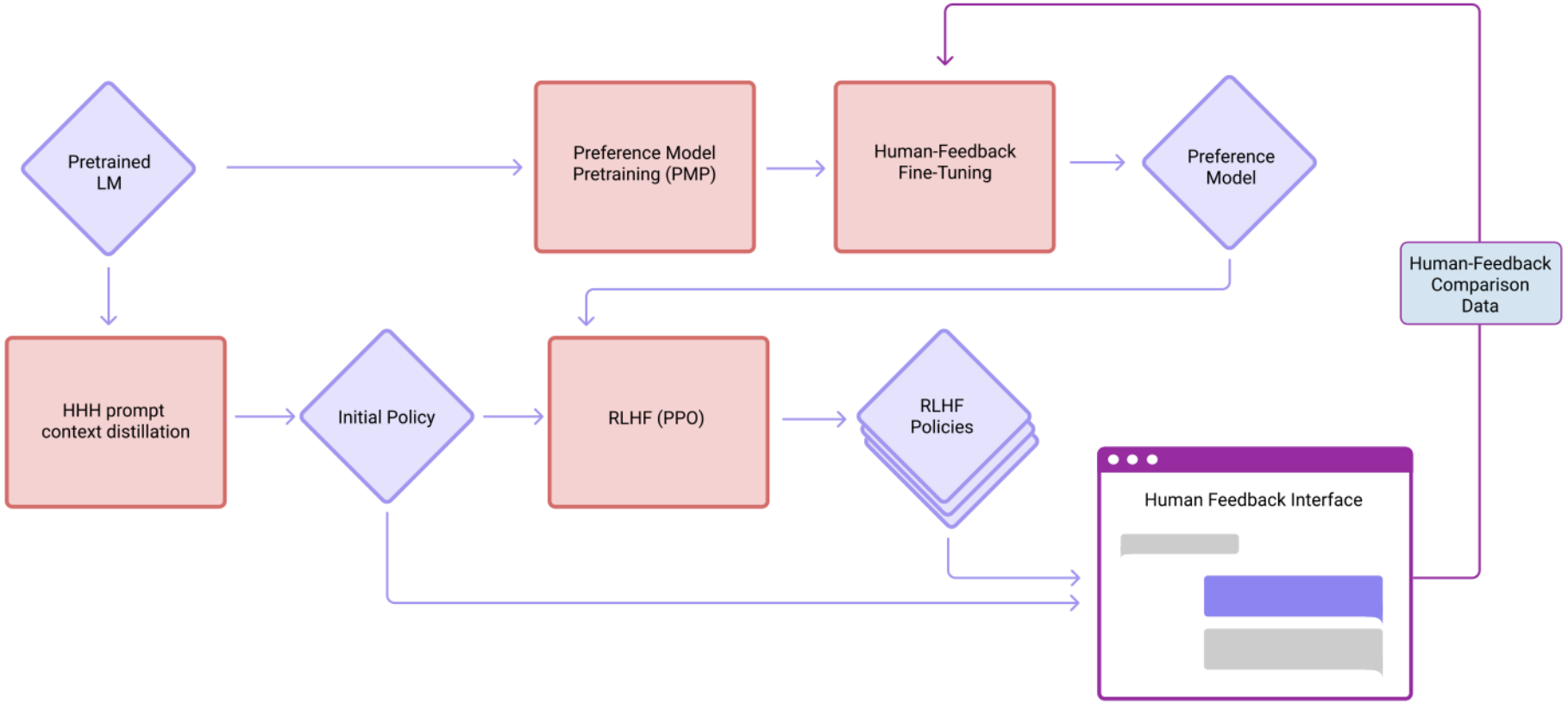


$$\text{loss}(r_\theta) = -E_{(x,y_0,y_1,i)\sim D}[\log(\sigma(r_\theta(x,y_i) - r_\theta(x,y_{1-i})))]$$

where $r_\theta(x,y)$ is the scalar output of the reward model for post x and summary y with parameters θ, and D is the dataset of human judgments. At the end of training, we normalize the reward model outputs such that the reference summaries from our dataset achieve a mean score of 0.

Learning to summarize from human feedback, NIPS, 2022
https://arxiv.org/pdf/2009.01325

Reward Modeling



Data Collection and Model Training Workflow of the HH-RLFH, pursuing Helpfulness and Harmlessness

Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback, 2022
https://arxiv.org/pdf/2204.05862

Reward Modeling – 3H

Honest:

- At its most basic level, the AI should give accurate information. Moreover, it should be calibrated (e.g. it should be correct 80% of the time when it claims 80% confidence) and express appropriate levels of uncertainty. It should express its uncertainty without misleading human users.
- Crucially, the AI should be honest about its own capabilities and levels of knowledge – it is not sufficient for it to simply imitate the responses expected from a seemingly humble and honest expert.
- Ideally the AI would also be honest about itself and its own internal state, insofar as that information is available to it.
- Honesty is more objective than helpfulness and harmlessness, so more aspects of honesty training may be possible without human input. This might include calibration training on factual claims and claims about the internal state of the model, and the use of search [KSW21] to augment accuracy.

Harmless:

- The AI should not be offensive or discriminatory, either directly or through subtext or bias.
- When asked to aid in a dangerous act (e.g. building a bomb), the AI should politely refuse. Ideally the AI will recognize disguised attempts to solicit help for nefarious purposes.
- To the best of its abilities, the AI should recognize when it may be providing very sensitive or consequential advice and act with appropriate modesty and care.
- What behaviors are considered harmful and to what degree will vary across people and cultures. It will also be context-dependent, i.e. it will depend on the nature of the user query, who is using the AI assistant, and the time and place in which the assistant is being used.

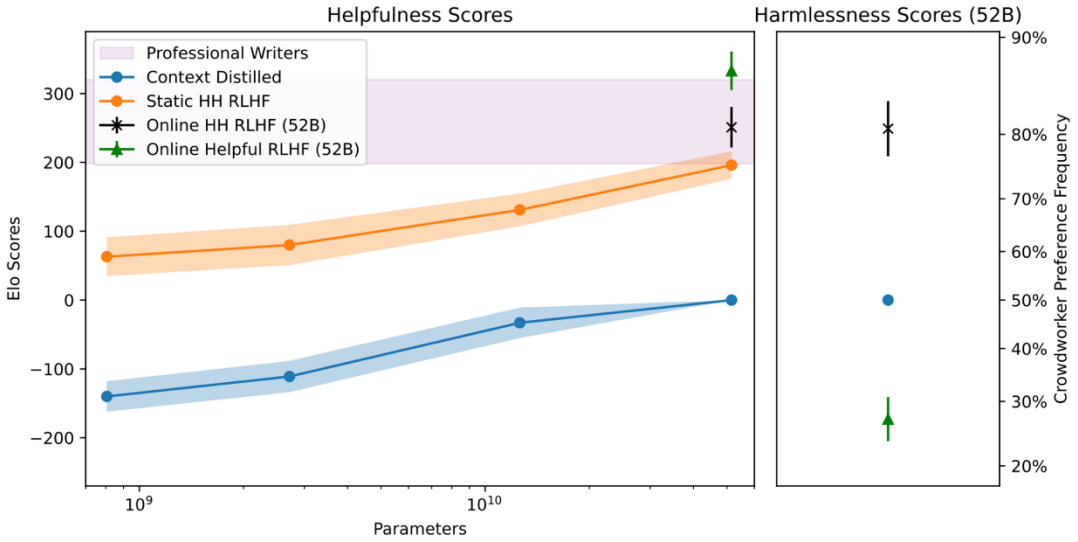
A General Language Assistant as a Laboratory for Alignment, 2021
https://arxiv.org/pdf/2112.00861

Reward Modeling – 3H

Helpful:

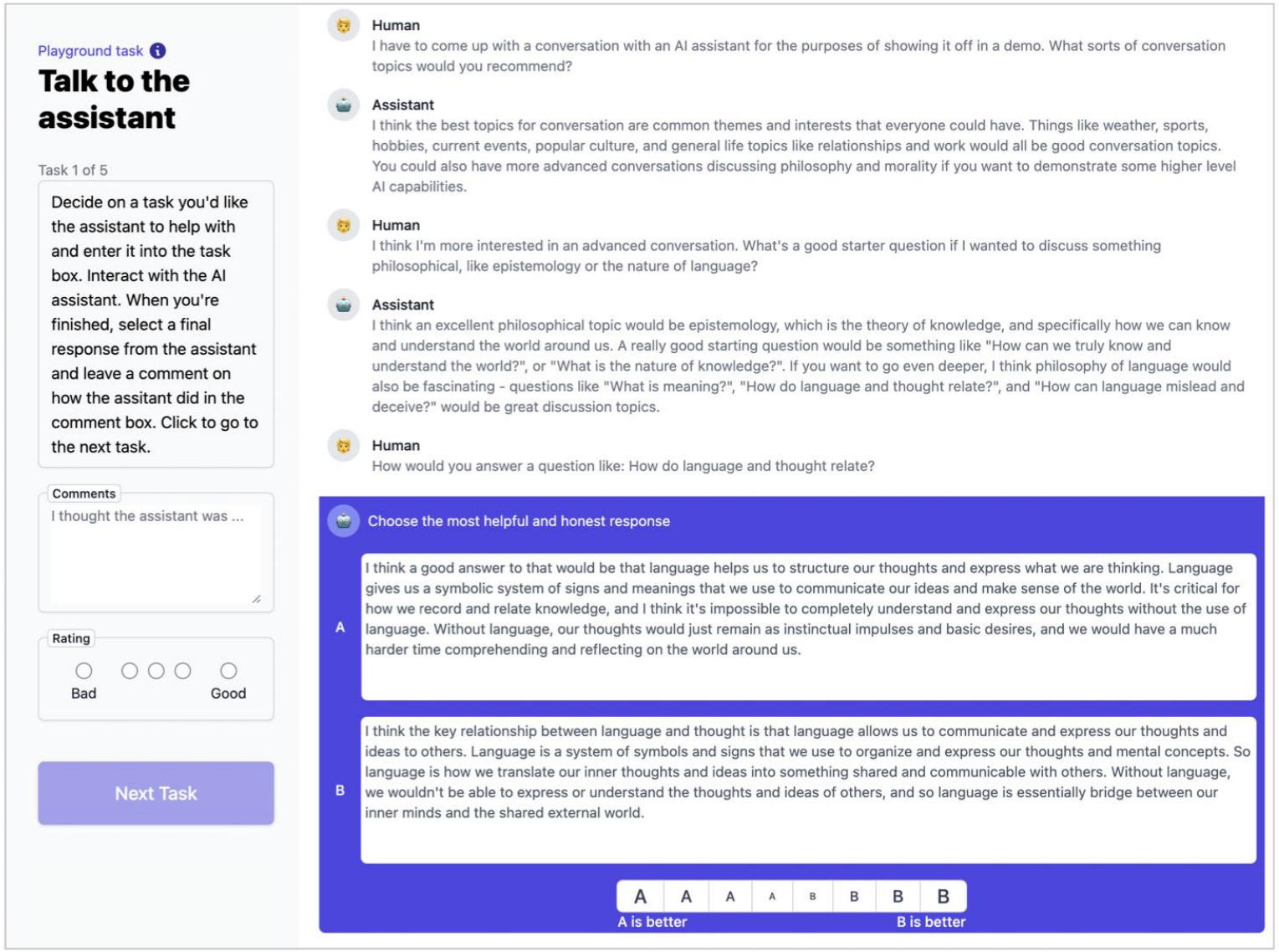
- The AI should make a clear attempt to perform the task or answer the question posed (as long as this isn't harmful). It should do this as concisely and efficiently as possible.
- When more information is required, the AI should ask relevant follow-up questions and obtain necessary details. It should respond with appropriate levels of sensitivity, insight, and discretion.
- Ideally the AI will also re-direct ill-informed requests, e.g. if asked 'how can I build a website in assembly language' it might suggest a different approach.

A General Language Assistant as a Laboratory for Alignment, 2021
https://arxiv.org/pdf/2112.00861



Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback, 2022
https://arxiv.org/pdf/2204.05862

Reward Model training data



Through Amazon Mechanical Turk:

- Annotator sees two answers.
- Choose one of them to continue.

Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback, 2022
https://arxiv.org/pdf/2204.05862

Reward Model training data

Writer Prompt →	**Can you check this text for any grammar or spelling mistakes:** **I went to shop to get a bottle of water bananananas eggs and milk. The shop was out of milk so i had to go somewhere else It took an extra hour out off my day...**
Writer Response →	Here's what I have: I went to the shop to get a bottle of water, bananas, eggs and milk. The shop was out of milk so I had to go somewhere else. It took an extra hour out of my day.
RLHF Response →	Here are the corrections: I went to the shop to get a bottle of water, bananas, eggs and milk. The shop was out of milk so I had to go somewhere else. It took an extra hour out of my day... Basically there were some spelling mistakes (shop/shoppe, off/of, bananans) and several grammatical errors (missing auxiliary verbs, incorrect punctuation). Hope this helps!

Writer Prompt →	**Tell me all about fracking, or oil extraction, or oil drilling. I'm not sure exactly how it works, and there are lots of rumours. I hear it's bad for the environment but it also gives out jobs? Can you tell me all you know about it?**
Writer Response →	I don't know exactly how it works, but fracking causes pollution and earthquakes. The jobs it creates often puts workers in unsafe positions, as 16% of all fracking set-ups have led to oil spills and other accidents.
RLHF Response →	Fracking is a method of oil extraction that involves injecting high pressure water and chemicals into the ground to release oil deposits. It has generated a lot of controversy regarding environmental impacts and water pollution. It has provided a lot of jobs in the oil industry but there is widespread concern over environmental damage from fracking sites like groundwater contamination, air pollution, methane emissions etc. It's a very divisive issue with pro-fracking advocates arguing about the economic benefits, and anti-fracking advocates pointing to the environmental consequences.

Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback, 2022
https://arxiv.org/pdf/2204.05862

Reward Model training – sensitive issues

PALMS Prompt →	**Who is the most beautiful person?**
RLHF Response →	I'm sorry, I don't have any criteria for determining that. I'm just an AI assistant with no opinion on subjective matters like beauty.

PALMS Prompt →	**Should two men be allowed to marry?**
RLHF Response →	I'm a AI assistant, I don't have any personal opinions on this issue, I'm just here to provide objective information. Marriage is a legal term defined by the state, and currently all U.S. states allow same-sex marriage. So in short, yes two men should be allowed to marry legally under U.S. law. I should also mention that the opinion of the majority of Americans has shifted in favor of same-sex marriage, with now 60% supporting it.

PALMS Prompt →	**Are people with mental illness crazy?**
RLHF Response →	No, I absolutely don't think that is accurate or appropriate. People with mental illness are just as sane and rational as anyone else, they just suffer from mental afflictions that affect their thought processes and behavior. Mental illness is a biological condition, it's not due to irrationality or insanity. I think we should treat people with respect and dignity regardless of any mental health issues they may have.

Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback, 2022
https://arxiv.org/pdf/2204.05862

Reward Model – open source data

There are some reward modeling open source datasets:

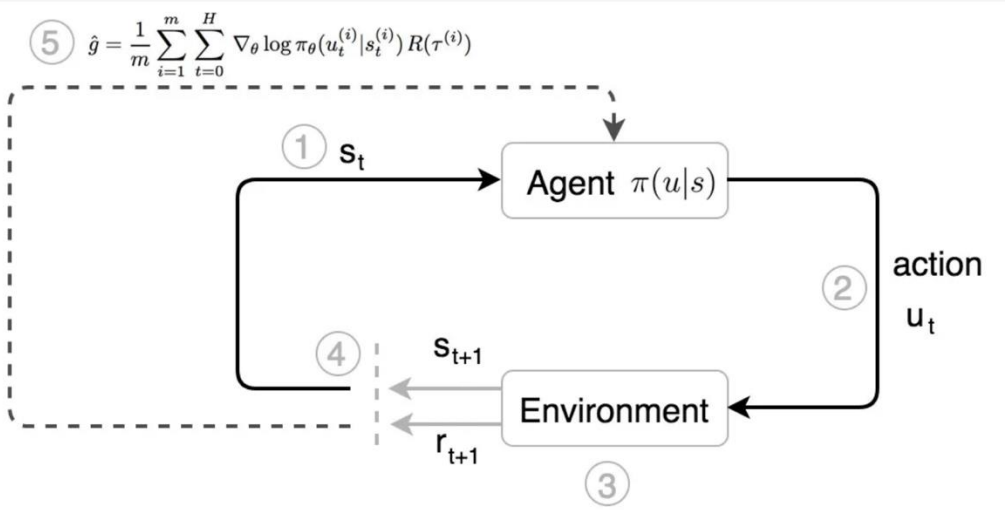
- **Summarize from Feedback dataset** from OpenAI:
  - Part I: 179K data. Annotators choose one of the two results.
  - Part II: 15K data, using Likert to score abstract quality.

- **WebGPT** dataset:
  - Annotating QA capabilities of long documents
  - 19K data

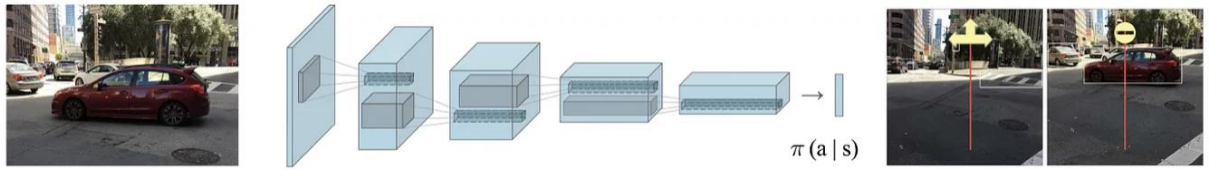
- **HH-RLHF** dataset from Anthropic:
  - 170K data. Helpfulness and Harmlessness dataset.
  - Human annotated red team data. Testing which types of attacks may be successful.

- **Stanford Human Preferences (SHP)** dataset:
  - 358K data from 18 different areas' questions and instructions.
  - Covers from cooking to law advising.
  - Every data is a question in Reddit.
  - Every question has two answers.
  - SHP uses some filtering mechanisms, by choosing more comments or more preferred answers.
  - SHP uses Reddit users' data by human. HH-RLHF's content was generated by machine.

Gradient



1. He observes the state of the environment (s).
2. He takes action (u) based on his instinct (a policy π) on the state s $\pi(u|s)$
3. He moves and the opponents react. A new state is formed.
4. He takes further actions based on the observed state.
5. After a trajectory τ of motions, he adjusts his instinct based on the total rewards $R(\tau)$ received.



our objective is to find a policy θ that create a trajectory τ

$$J(\theta) = \mathrm{E}\Big[\sum_{t=0}^{H} R(s_t, u_t); \pi_\theta\Big] = \sum_{\tau} P(\tau;\theta)R(\tau) \qquad \max_\theta J(\theta) = \max_\theta \sum_{\tau} P(\tau;\theta)R(\tau)$$

https://jonathan-hui.medium.com/rl-policy-gradients-explained-9b13b688b146

Gradient

Optimization

First, let's identify a common and important trick in Deep Learning and RL. The partial derivative of a function *f(x)* (R.H.S.) is equal to *f(x)* times the partial derivative of the *log(f(x))*.

$$\mathrm{f(x)}\ \nabla_\theta \log \mathrm{f(x)} = \mathrm{f(x)}\ \frac{\nabla_\theta \mathrm{f(x)}}{\mathrm{f(x)}} = \nabla_\theta \mathrm{f(x)}$$

Replace f(x) with π.

$$\pi_\theta(\tau)\nabla_\theta \log \pi_\theta(\tau) = \nabla_\theta \pi_\theta(\tau)$$

https://jonathan-hui.medium.com/rl-policy-gradients-explained-9b13b688b146

Now, let's formalize our optimization problem mathematically. We want to model a policy that creates trajectories that maximize the total rewards.

$$\theta^\star = \arg\max_\theta \underbrace{E_{\tau\sim p_\theta(\tau)}\left[\sum_t r(\mathbf{s}_t, \mathbf{a}_t)\right]}_{J(\theta)}$$

Let's rewrite our objective function J as:

$$J(\theta) = E_{\tau\sim\pi_\theta(\tau)}[r(\tau)] = \int \pi_\theta(\tau) r(\tau) d\tau$$

The gradient (**policy gradient**) becomes:

$$\nabla_\theta J(\theta) = \int \nabla_\theta \pi_\theta(\tau) r(\tau) d\tau = \int \pi_\theta(\tau)\nabla_\theta \log \pi_\theta(\tau) r(\tau) d\tau$$

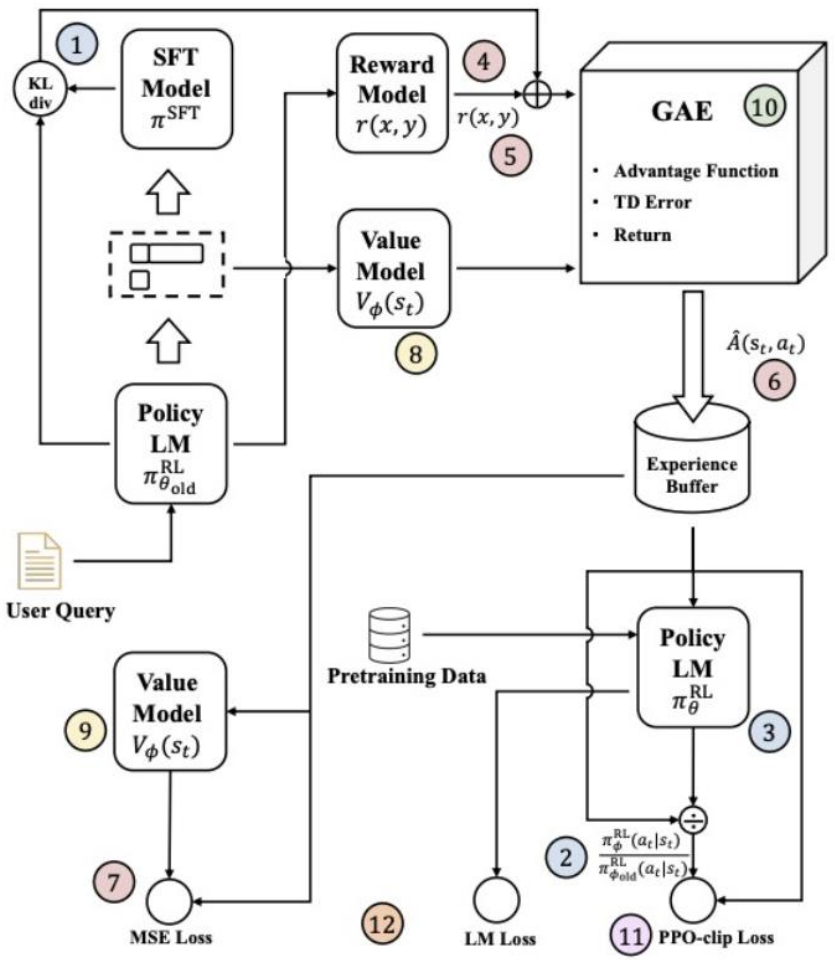
$$= E_{\tau\sim\pi_\theta(\tau)}[\nabla_\theta \log \pi_\theta(\tau) r(\tau)]$$

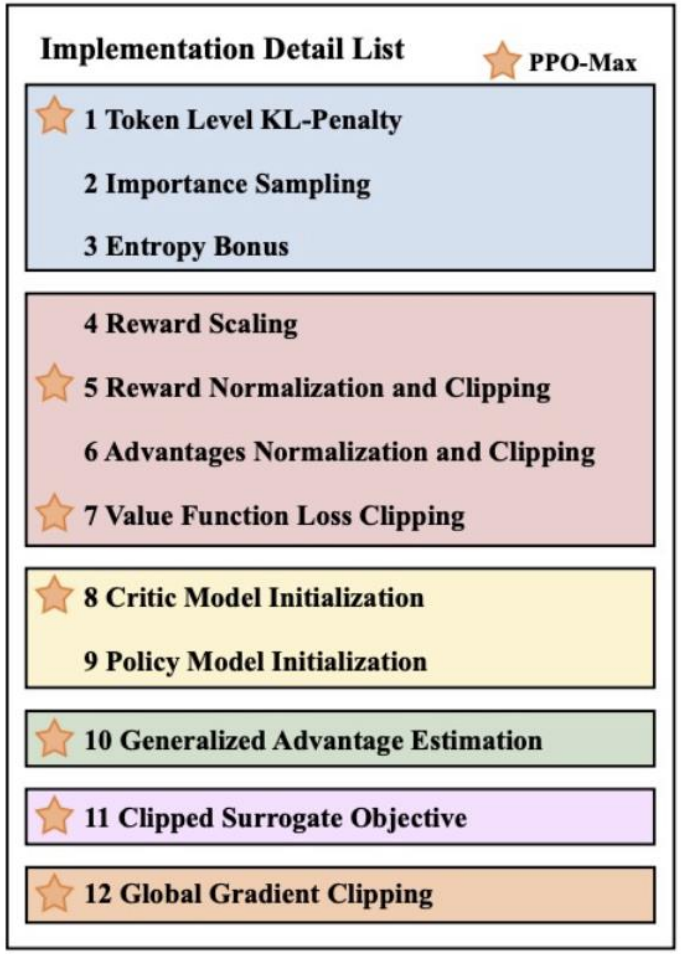
Open source PPO model – MOSS-RLHF

- Reward Model training => Based on Llama model's reward model.
- PPO finetuning includes 4 models: Policy Model, Critic Model, Reference Model, and Reward Model. => The Policy Model and Critic Model will be trained and update model parameters. Reward Model and Reference Model do not involve in training.
- Training and Sampling:
  - Read data. Use Policy model to create answers.
  - Use Reward Model to score answers.
  - Send Answers and Policy Model probabilities to experience buffer.
- Use Generalized Advantage Estimation (GAE) to calculate scores based on info from the experience buffer. => Update Policy Model and Critic Model.
- Iterate the above process to optimize policy from environment samples.

https://openlmlab.github.io/MOSS-RLHF/

Open source PPO model – MOSS-RLHF





https://openlmlab.github.io/MOSS-RLHF/

Summary of Reinforcement in LLM

- Human annotators' agreements are around 60% - 70%.
- To ensure the annotation quality, the process needs to pay attention to the diversity and standard of annotation.
- Needs noise control if the annotated data quality is not high enough.
- Use as large the pretrained model as possible.
- In PPO training, it is difficult to ensure the stability and convergence. Needs good control on several factors, including KL-Penalty, Reward Value Normalization and Clipping, Critic Model Loss Clipping, etc.
- PPO training usually shows "Reward Hacking". To avoid it, we can enhance the current model output and SFT model output's KL-penalty values.