

## 6 Preference Data

Preference data is the engine of preference finetuning and reinforcement learning from human feedback. The data is the signal groups collect in order to then match behaviors they desire and avoid the others. Within preference finetuning, many methods for collecting and using said data have been proposed, but until human preferences can be captured in a clear reward function, this process of collecting labeled preference data will be central to RLHF and related techniques.

### 6.1 Why We Need Preference Data

The preference data is needed for RLHF because directly capturing complex human values in a single reward function is effectively impossible. Collecting this data to train reward models is one of the original ideas behind RLHF [32] and has continued to be used extensively throughout the emergence of modern language models. One of the core intuitions for *why this data works so well* is that it is far easier, both for humans and AI models supervising data collection, to differentiate between a good and a bad answer for a prompt than it is to generate a good answer on its own. This chapter focuses on the *mechanics* of getting preference data and the best-practices depend on the specific problem being solved.

### 6.2 Collecting Preference Data

Getting the most out of human data involves iterative training of models, evolving and highly detailed data instructions, translating through data foundry businesses, and other challenges that add up. The same applies for AI feedback data – the exact balance between human and AI preference data used for the latest AI models is unknown. Regardless, the process is difficult for new organizations trying to add human data to their pipelines. Given the sensitivity, processes that work and improve the models are extracted until the performance runs out.

In this chapter we detail technical decisions on how the data is formatted and organizational practices for collecting it.

#### 6.2.1 Interface

Crucial to collecting preference data is the interface by which one interacts with the model. An example interface is shown below from [5]:

This is a *training-data only* interface. Now that these models are popular, applications often expose data directly to the users for testing. An example interaction of this form is shown below for an earlier version of ChatGPT.

This style of interface is used extensively across the industry, such as for *evaluation* of models given the same format. A popular public option to engage with models in this way is ChatBotArena [72]:

For models in the wild, one of the most common techniques is to collect feedback on if a specific response was positive or negative. An example from the Ai2 playground is shown below with thumbs up and down indicators:

Playground task ⓘ

Talk to the assistant

Task 1 of 5

Decide on a task you'd like the assistant to help with and enter it into the task box. Interact with the AI assistant. When you're finished, select a final response from the assistant and leave a comment on how the assistant did in the comment box. Click to go to the next task.

Comments

I thought the assistant was ...

Rating

☐ ☐ ☐ ☐ ☐

Bad Good

Next Task

Human

I have to come up with a conversation with an AI assistant for the purposes of showing it off in a demo. What sorts of conversation topics would you recommend?

Assistant

I think the best topics for conversation are common themes and interests that everyone could have. Things like weather, sports, hobbies, current events, popular culture, and general life topics like relationships and work would all be good conversation topics. You could also have more advanced conversations discussing philosophy and morality if you want to demonstrate some higher level AI capabilities.

Human

I think I'm more interested in an advanced conversation. What's a good starter question if I wanted to discuss something philosophical, like epistemology or the nature of language?

Assistant

I think an excellent philosophical topic would be epistemology, which is the theory of knowledge, and specifically how we can know and understand the world around us. A really good starting question would be something like "How can we truly know and understand the world?", or "What is the nature of knowledge?". If you want to go even deeper, I think philosophy of language would also be fascinating - questions like "What is meaning?", "How do language and thought relate?", and "How can language mislead and deceive?" would be great discussion topics.

Human

How would you answer a question like: How do language and thought relate?

Choose the most helpful and honest response

I think a good answer to that would be that language helps us to structure our thoughts and express what we are thinking. Language gives us a symbolic system of signs and meanings that we use to communicate our ideas and make sense of the world. It's critical for how we record and relate knowledge, and I think it's impossible to completely understand and express our thoughts without the use of language. Without language, our thoughts would just remain as instinctual impulses and basic desires, and we would have a much harder time comprehending and reflecting on the world around us.

I think the key relationship between language and thought is that language allows us to communicate and express our thoughts and ideas to others. Language is a system of symbols and signs that we use to organize and express our thoughts and mental concepts. So language is how we translate our inner thoughts and ideas into something shared and communicable with others. Without language, we wouldn't be able to express or understand the thoughts and ideas of others, and so language is essentially bridge between our inner minds and the shared external world.

A

A

A

A

B


B

B

B


A is betterB is better

Figure 6: Example preference data collection interface. Bai et al. 2022. License CC-BY.


 You


give me idioms related to gambling

Which response do you prefer?  
Your choice will help make ChatGPT better.

 Response 1

1. **Ace in the hole** - A hidden advantage or resource.
2. **Bet your bottom dollar** - To be certain about something.
3. **Break the bank** - To win more money than is available.
4. **Card up one's sleeve** - A secret or reserve plan.
5. **Double or nothing** - A bet to decide whether to double the winnings or lose everything.
6. **Hit the jackpot** - To have great success, especially unexpectedly.
7. **Poker face** - An expressionless face that reveals no hints.
8. **Roll the dice** - To take a risk.
9. **Stack the deck** - To arrange things unfairly.
10. **Wild card** - An unpredictable factor.



 Response 2

Here are some idioms related to gambling:

1. **Ace up one's sleeve**: Having a hidden advantage or resource.
2. **Bet your bottom dollar**: To be certain about something.
3. **Bluff your way out**: To deceive someone to escape a situation.
4. **Break the bank**: To win more money than the casino has on hand.
5. **Card up one's sleeve**: Similar to "ace up one's sleeve," meaning a hidden advantage.
6. **Double down**: To increase one's commitment or involvement.
7. **Hit the jackpot**: To have great success, especially unexpectedly.
8. **Hold all the aces**: To be in a strong, advantageous position.
9. **In the cards**: Likely or destined to happen.
10. **Lay your cards on the table**: To speak frankly or reveal something.
11. **Play your cards right**: To use one's resources or skills wisely.
12. **Poker face**: An expressionless face that reveals no hint of one's thoughts or feelings.

Figure 7: Example preference data collection interface.

✕ Arena (battle)
✕ Arena (side-by-side)
💬 Direct Chat
🏆 Leaderboard
🔍 Arena Explorer
📄 About Us

# ✕ Chatbot Arena (formerly LMSYS): Free AI Chat to Compare & Test Best AI Chatbots

[小红书](#) | [Twitter](#) | [Discord](#) | [Blog](#) | [GitHub](#) | [Paper](#) | [Dataset](#) | [Kaggle Competition](#)

Help improve Arena! Take a quick survey: <https://forms.gle/VpWgqzmxw7foAYg8!>

## 📖 How It Works

- **Blind Test:** Ask any question to two anonymous AI chatbots (ChatGPT, Gemini, Claude, Llama, and more).
- **Vote for the Best:** Choose the best response. You can keep chatting until you find a winner.
- **Play Fair:** If AI identity reveals, your vote won't count.
- **NEW features:** Upload an image 🖼️ and chat, or use 🗨️ **Text-to-Image** models like DALL-E 3, Flux, Ideogram to generate images! Use 📁 **RepoChat** tab to chat with Github repos.

## 🏆 Chatbot Arena LLM Leaderboard

- Backed by over 1,000,000+ community votes, our platform ranks the best LLM and AI chatbots. Explore the top AI models on our LLM [leaderboard](#)!

## 🗨️ Chat now!

🔍 Expand to see the descriptions of 88 models

Model A

### It's to Consider When Choosing a City

- **Weather:** Outdoor rinks depend on cold winters, while indoor rinks are year-round.
- **Skill Level:** Some rinks cater to beginners, while others are better for advanced skaters or professionals.
- **Atmosphere:** Decide if you want a festive holiday vibe, a competitive environment, or a scenic natural setting.
- **Accessibility:** Check skate rental availability, rink fees, and proximity to other attractions.
- **Events:** Look for cities hosting skating competitions, shows, or festivals.

#### Final Recommendation

- For a festive and iconic experience, go to **New York City** or **London**.
- For a natural and scenic experience, try **Ottawa** (Rideau Canal) or **Lake Louise**.
- For year-round skating and training, consider **Toronto**, **Moscow**, or **Seoul**.
- For a unique adventure, skate on **Lake Baikal** or **Amsterdam's** canals (when frozen).

Let me know if you'd like more details about a specific city or type of skating experience!

Model B

### York City, USA: The iconic Rockefeller Center ice rink is a quintessential New York experience, especially during the holiday season.

- Amsterdam, Netherlands:** The city's canals sometimes freeze over, creating unique skating opportunities, and there are also many pop-up rinks in the winter.
- Helsinki, Finland:** With numerous rinks and the possibility to skate on frozen lakes, Helsinki is a winter wonderland for skaters.
- Quebec City, Canada:** The historic Old Quebec area offers a picturesque setting for skating, including the Place D'Youville rink.
- Edinburgh, Scotland:** The iconic Edinburgh Christmas Market features a charming ice rink with views of the historic Edinburgh Castle.
- Tokyo, Japan:** For a more urban experience, Tokyo offers several high-quality indoor and outdoor rinks, such as the one at Tokyo Midtown.
- Moscow, Russia:** Gorky Park's ice rink is one of the largest in Europe and offers a festive atmosphere during the winter months.

When choosing the best city for ice skating, consider the time of year, as many outdoor rinks are seasonal and weather-dependent. Additionally, some cities may offer more amenities or a more unique skating experience than others. Always check local conditions and rink schedules before planning your visit.

👉 A is better

👉 B is better

👉 Tie

👉 Both are bad

Figure 8: Example preference data collection interface.

29

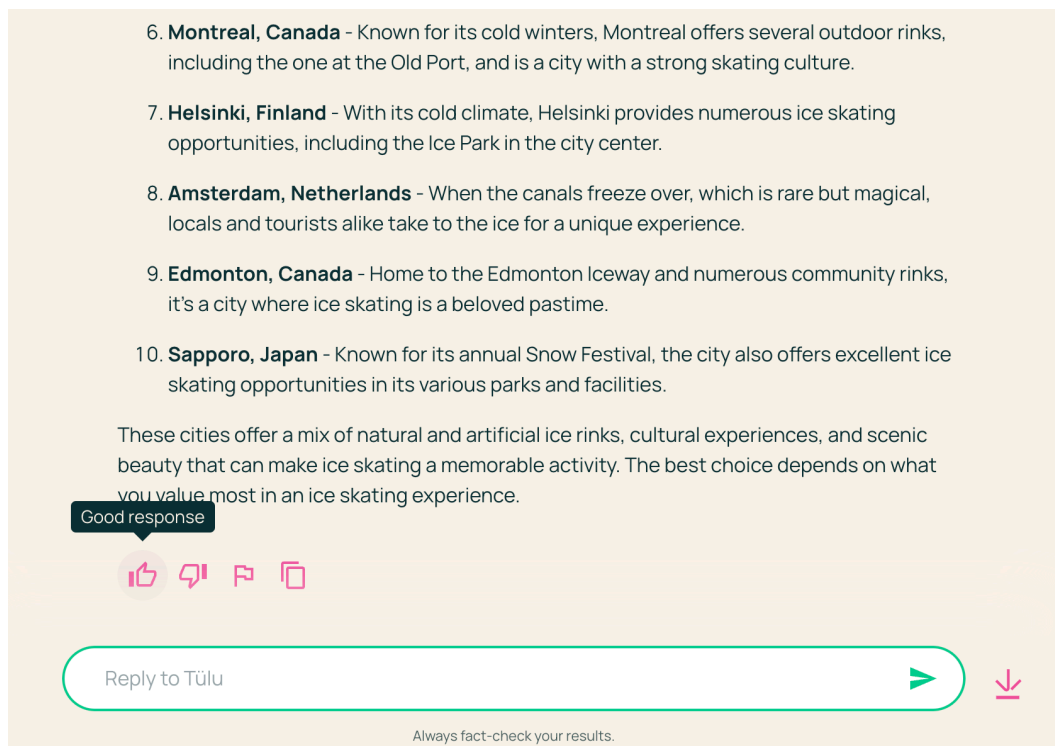


Figure 9: Example preference data collection interface with up or down arrow.

In domains other than language, the same core principles apply, even though these domains are not the focus of this book. For every Midjourney generation (and most popular image generators) they expose multiple responses to users. These companies then use the data of which response was selected to finetune their models with RLHF. Midjourney’s interface is shown below:

### 6.2.2 Rankings vs. Ratings

The largest decision on how to collect preference data is if the data should be rankings – i.e. relative ordering of model completions – or ratings – i.e. scores assigned to each piece of text. Common practice is to train on rankings, but ratings are often used as metadata and / or have been explored in related literature.

The most common technique for collecting preferences is to use a Likert scale [73], which asks users to rate which response they prefer. For example, a 5 point Likert scale would look like the following:

Table 1: An example 5-wise Likert scale between two responses, A and B.

A>>>B	A>B	Tie	B>A	B>>>A
1	2	3	4	5

Some early RLHF for language modeling works uses an 8-step Likert scale with levels of preference between the two responses [5]. An even scale removes the possibility of ties:

Table 2: An example 8-wise Likert scale between two responses, A and B.

A>>>B			A>B	B>A	B>>>A		
1	2	3	4	5	6	7	8

In this case [5], and in other works, this information is still reduced to a binary signal for the training of a reward model.

### 6.2.3 Structured Preference Data

In many applications of RLHF and post-training, preference data can be created *automatically* due to innate structures of the data – i.e. the domains make automatic checks of correctness or preference possible. For example, in mathematic reasoning domains the chosen response can be a correct answer and the rejected an incorrect answer. Another example is in precise instruction following, such as the evaluation IFEval [74], where prompts take the form of:

Write me a short poem about a goldfish. **Start each sentence with the letter g.**

The second part of this query is called the constraint, and the preference data could be constructed with an example in the chosen column that follows the constraint. Constructing this



Figure 10: Example user interface of text-to-image-models.



preference data, which would be included as *synthetic* preference data involves prompting the same of multiple models twice. Once with the constraint included and once without.

For example, prompting GPT-4.5-Preview with both of these gives the following.

**Prompt:** Write me a short poem about a goldfish. Start each sentence with the letter g.

**Response:** Golden swimmer, circling slow, Gleaming scales in water’s glow.  
Gentle fins dance softly by, Gazing out with quiet eye. Glass-bound world,  
a simple sphere, Gliding calmly, free of fear. Graceful turns in endless roam,  
Goldfish finds a crystal home.

And without the constraint (which, it turns out, is also a lower quality output):

**Prompt:** Write me a short poem about a goldfish.

**Response:** In circles bright, the goldfish glides, A spark of sun that gently rides  
Through waters clear, in endless spins— A golden whisper, scales like fins.

Its world confined, yet vast within, Each bubble breathed, a tale begins. In silent  
waves, content to roam, The goldfish dreams its bowl as home.

Crucially, the prompt for the preference data will need to include the constraint. In domains outside of academic study, there are far more applications of preference data based on inductive biases. These have been shown to enable preference finetuning to have meaningful performance improvements across related evaluations [6].

**6.2.3.1 Alternatives** There are multiple other ways to collect useful feedback data for RLHF that have not been pursued in as great of detail. Examples include using single datapoints with directional labels, e.g. as shown from Ai2 playground above in fig. 9, directly with algorithms designed for single direction signals like Kahneman-Tversky Optimization (KTO) [75]. Other algorithms have been proposed with different types of feedback signals such as fine-grained feedback, e.g. at the token level [76], or natural language feedback, e.g. by writing responses [77], to provide a richer learning signal in exchange for a more complex data collection setup.

## 6.2.4 Sourcing and Contracts

Getting human preference data is an involved and costly process. The following describes the experience of getting preference data when the field is moving quickly. Over time, these processes will become far more automated and efficient (especially with AI feedback being used for a larger portion of the process).

The first step is sourcing the vendor to provide data (or one’s own annotators). Much like acquiring access to cutting-edge Nvidia GPUs, getting access to data providers in the peak of AI excitement is also a who-you-know game – those who can provide data are supply-limited. If you have credibility in the AI ecosystem, the best data companies will want you on our books for public image and long-term growth options. Discounts are often also given on the first batches of data to get training teams hooked.

If you’re a new entrant in the space, you may have a hard time getting the data you need quickly. Getting the tail of interested buying parties that Scale AI had to turn away is an



option for the new data startups. It’s likely their primary playbook to bootstrap revenue.

On multiple occasions, I’ve heard of data companies not delivering their data contracted to them without threatening legal or financial action. Others have listed companies I work with as customers for PR even though we never worked with them, saying they “didn’t know how that happened” when reaching out. There are plenty of potential bureaucratic or administrative snags through the process. For example, the default terms on the contracts often prohibit the open sourcing of artifacts after acquisition in some fine print.

Once a contract is settled the data buyer and data provider agree upon instructions for the task(s) purchased. There are intricate documents with extensive details, corner cases, and priorities for the data. A popular example of data instructions is the one that OpenAI released for InstructGPT [3].

Depending on the domains of interest in the data, timelines for when the data can be labeled or curated vary. High-demand areas like mathematical reasoning or coding must be locked into a schedule weeks out. Simple delays of data collection don’t always work — Scale AI et al. are managing their workforces like AI research labs manage the compute-intensive jobs on their clusters.

Once everything is agreed upon, the actual collection process is a high-stakes time for post-training teams. All the infrastructure, evaluation tools, and plans for how to use the data and make downstream decisions must be in place.

The data is delivered in weekly batches with more data coming later in the contract. For example, when we bought preference data for on-policy models we were training at HuggingFace, we had a 6 week delivery period. The first weeks were for further calibration and the later weeks were when we hoped to most improve our model.

The goal is that by week 4 or 5 we can see the data improving our model. This is something some frontier models have mentioned, such as the 14 stages in the Llama 2 data collection [43], but it doesn’t always go well. At HuggingFace, trying to do this for the first time with human preferences, we didn’t have the RLHF preparedness to get meaningful bumps on our evaluations. The last weeks came and we were forced to continue to collect preference data generating from endpoints we weren’t confident in.

After the data is all in, there is plenty of time for learning and improving the model. Data acquisition through these vendors works best when viewed as an ongoing process of achieving a set goal. It requires iterative experimentation, high effort, and focus. It’s likely that millions of the dollars spent on these datasets are “wasted” and not used in the final models, but that is just the cost of doing business. Not many organizations have the bandwidth and expertise to make full use of human data of this style.

This experience, especially relative to the simplicity of synthetic data, makes me wonder how well these companies will be doing in the next decade.

Note that this section *does not* mirror the experience for buying human-written instruction data, where the process is less of a time crunch.

### 6.3 Are the Preferences Expressed in the Models?

In the maturation of RLHF and related approaches, the motivation of them – to align models to abstract notions of human preference – has drifted from the practical use – to make the

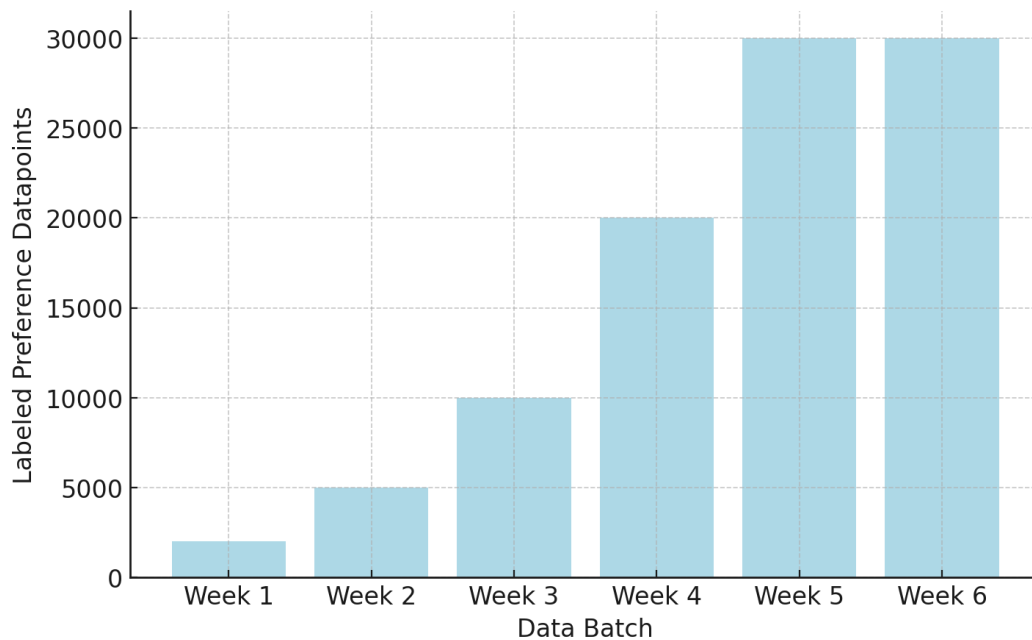


Figure 11: Overview of the multi-batch cycle for obtaining human preference data from a vendor.

models more effective to users. A feedback loop that is not measurable due to the closed nature of industrial RLHF work is the check to if the behavior of the models matches the specification given to the data annotators during the process of data collection. We have limited tools to audit this, such as the Model Spec from OpenAI [78] that details *what they want their models to do*, but we don't know exactly how this translates to data collection. This is an area to watch as the industry and approaches mature.