JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

The Llama 3 Herd of Models

A Deep Dive into Alignment Techniques

Aydan Huang & Zihao Zhao
Date: Sep 12, 2024

Post-Training Pipeline

- Reward Modeling (RM)
- Supervised Fine-Tuning (SFT)
- Direct Preference Optimization (DPO)
- Highlight: Iterative process across 6 rounds

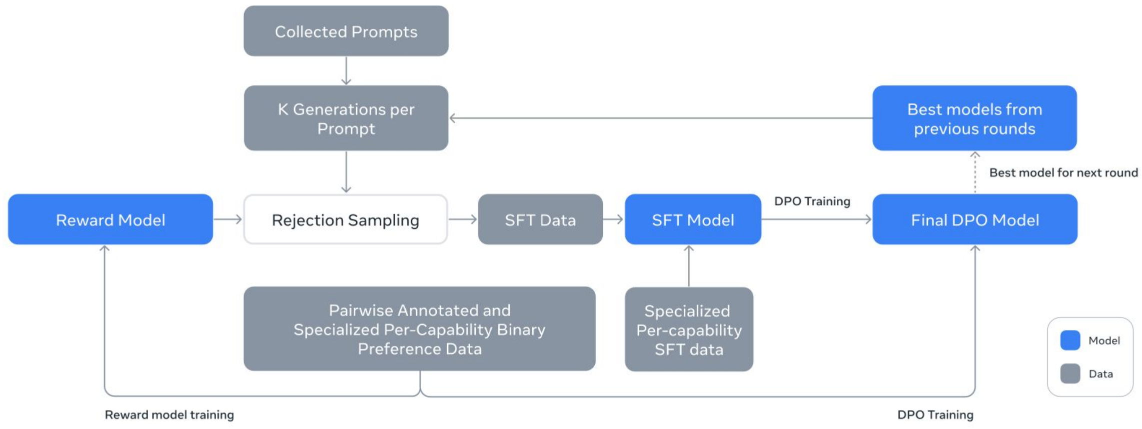
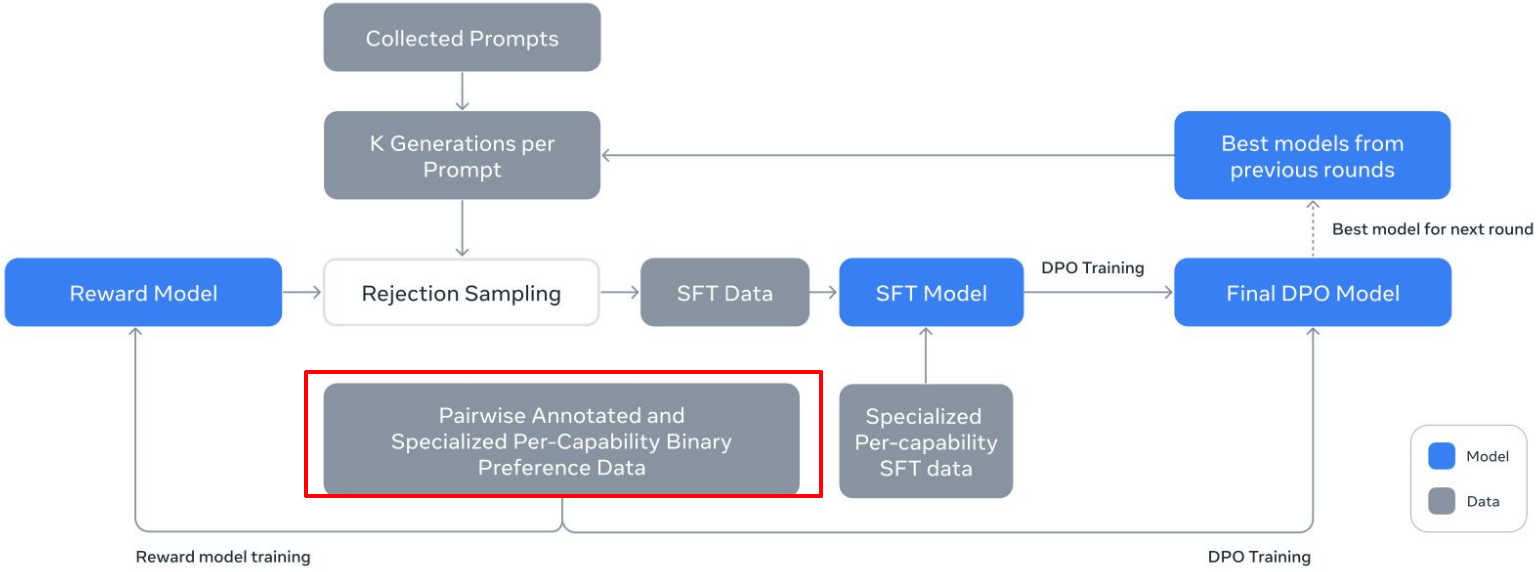


Figure 7 Illustration of the overall post-training approach for Llama 3. Our post-training strategy involves rejection sampling, supervised finetuning, and direct preference optimization. See text for details.

Preference data

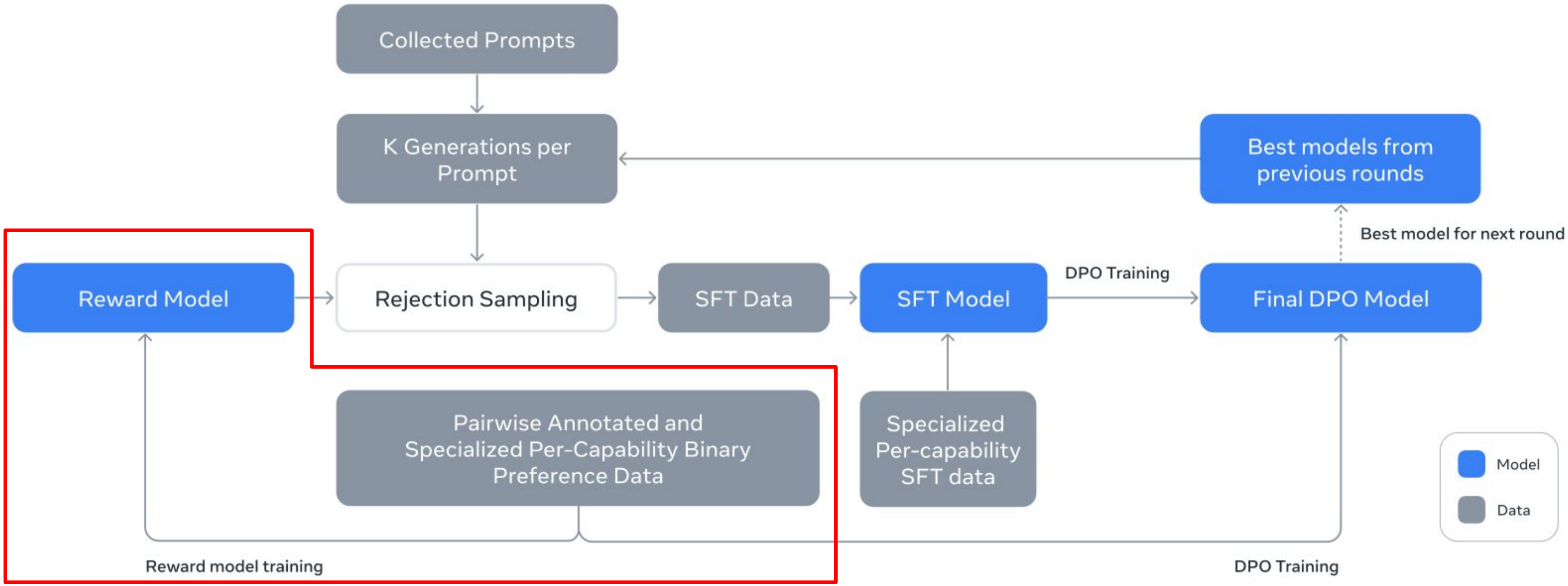


Preference data

- Preference data: data that reflects the people's preference.
- Sample two responses from two different models for each user prompt.
- Annotators to rate the strength of their preference in four levels: **significantly better**, **better**, **slightly better**, or **marginally better**.
- Use samples that are labeled **significantly better** or **better** and discard samples with similar responses.
- Incorporate an editing step after preference ranking. (edited > chosen > rejected)

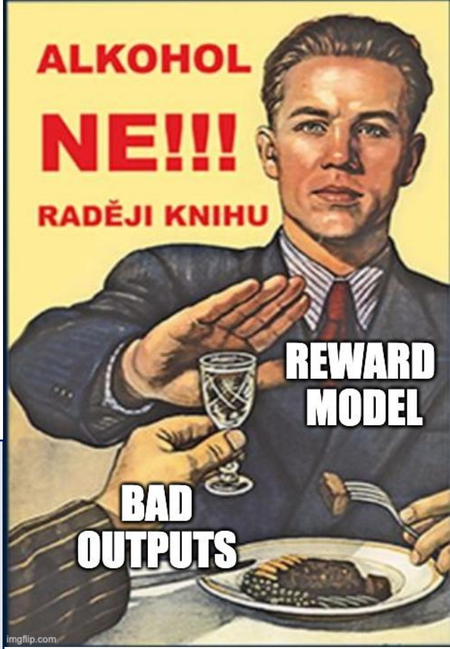
Dataset	% of comparisons	Avg. # turns per dialog	Avg. # tokens per example	Avg. # tokens in prompt	Avg. # tokens in response
General English	81.99%	4.1	1,000.4	36.4	271.2
Coding	6.93%	3.2	1,621.0	113.8	462.9
Multilingual	5.19%	1.8	1,299.4	77.1	420.9
Reasoning and tools	5.89%	1.6	707.7	46.6	129.9
Total	100%	3.8	1,041.6	44.5	284.0

Reward Modeling



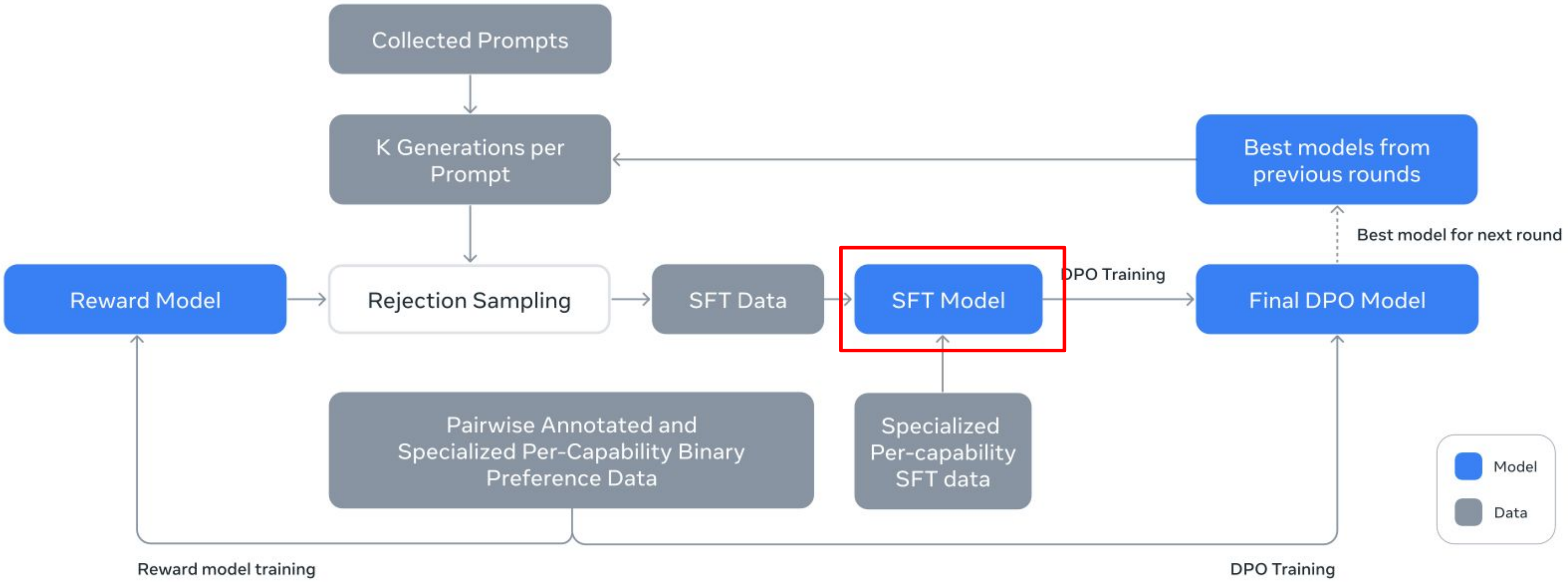
Reward Modeling

- Training objective: Ranking responses (edited > chosen > rejected)
- Innovation: Concatenated prompt-response pairs for efficiency
- Use in rejection sampling: Selecting high-quality responses for SFT



- **Prompt**: "How does quantum computing differ from classical computing?"
- Generated Responses:
- **Edited**: "Quantum computing leverages the principles of quantum mechanics, such as superposition and entanglement, to perform complex calculations much faster than classical computers, which rely on binary states."
- **Chosen**: "Quantum computing uses qubits instead of bits, allowing for more complex calculations to be done simultaneously."
- **Rejected**: "Quantum computing is faster than regular computers because it uses quantum physics."

Supervised Fine-Tuning (SFT)

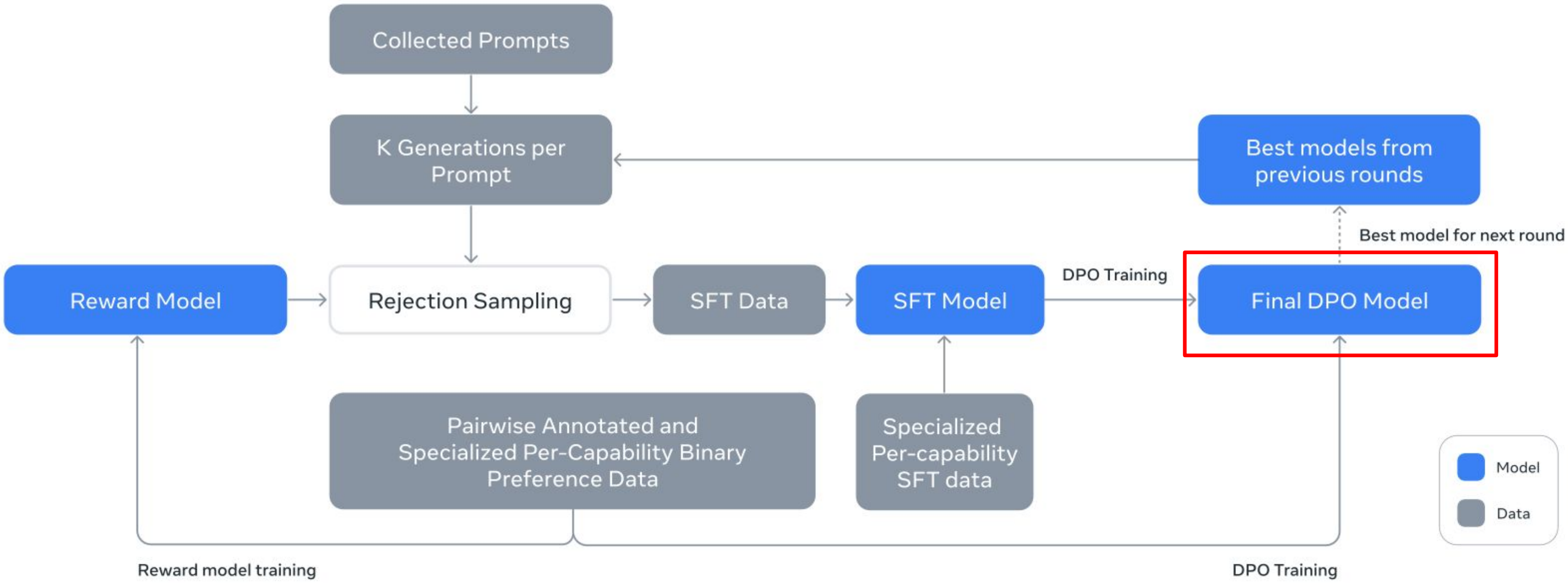


Supervised Fine-Tuning (SFT)

- Data composition: rejection-sampled responses, synthetic data
- Training details:
  - Cross entropy loss
  - Learning rate: 10^-5
  - Steps: 8.5K to 9K for 405B model

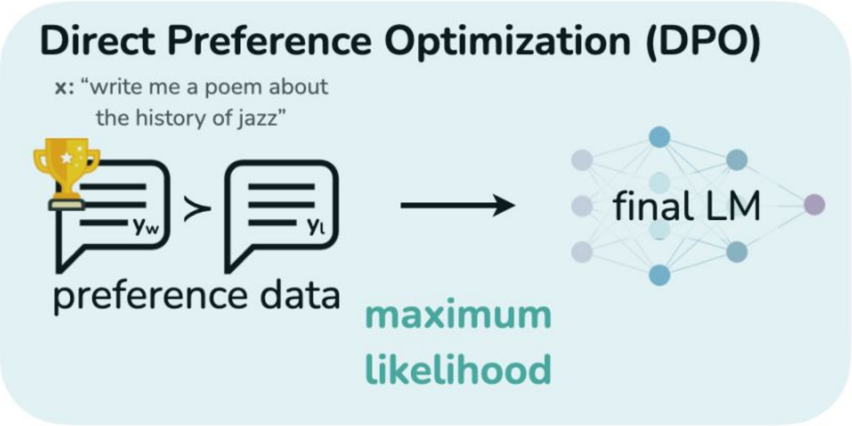
- Prompt: "Explain the significance of the law of entropy in physics."
- **Rejection-Sampled Response**: "The law of entropy, or the second law of thermodynamics, states that the total entropy of an isolated system always increases over time, leading to the concept of irreversible processes."
- **Synthetic Response**: "Entropy is a measure of disorder in a system. The second law of thermodynamics says that in any energy transfer, entropy will increase, meaning systems tend to become more disordered over time."

Direct Preference Optimization (DPO)

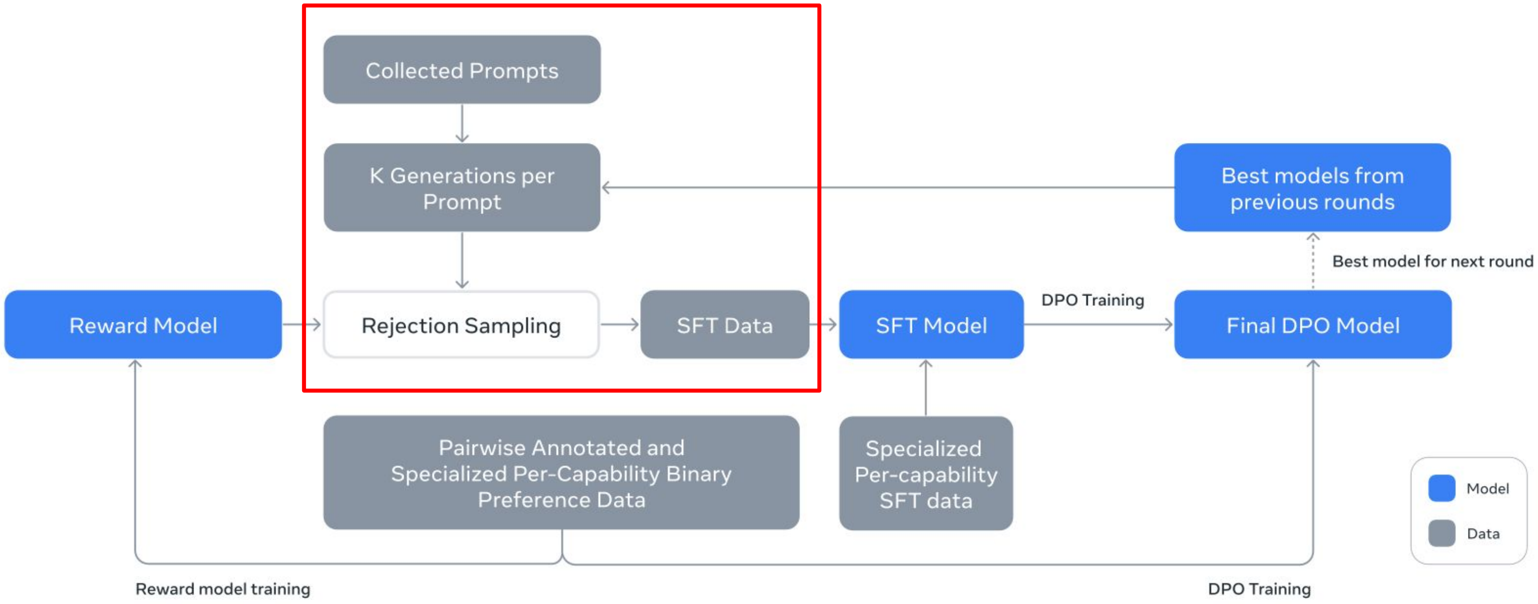


Direct Preference Optimization (DPO)

- Key modifications
  - Masking special tokens in loss calculation
  - NLL loss term (coefficient: 0.2) for stability
    - Result: Improved performance on instruction-following benchmarks
- Effects
  - Stabilized training
  - Maintained generation formatting

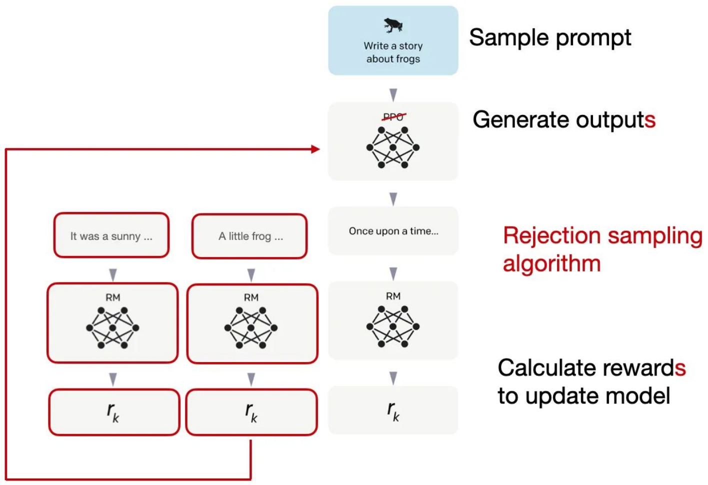


Rejection Sampling and Data processing



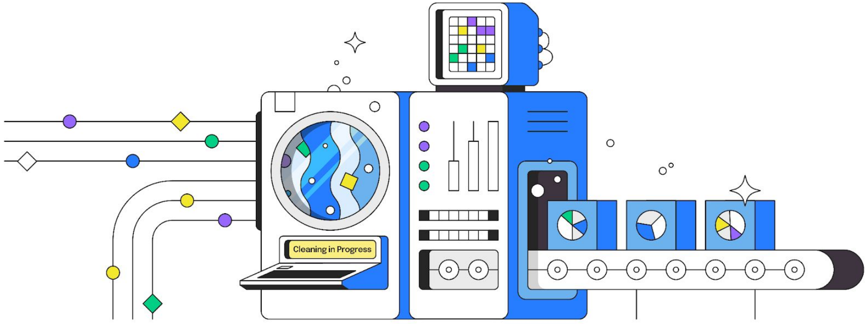
Rejection Sampling

- **Rejection sampling**: generate multiple possible outputs for a given prompt and then select the best output based on certain criteria.
- For each prompt in preference data, sample 10 to 30 responses from best performing checkpoint from the previous post-training iteration
- Reward model: select the best response



Data Processing and Quality Control

- **Data cleaning**: implemented rule-based strategies to filter or modify data to remove or adjust problematic elements
- Example: mitigate overly-apologetic tonal issues, we identify overused phrases (such as “I’m sorry” or “I apologize”) and carefully balance the proportion of such samples in our dataset.



Data Processing and Quality Control

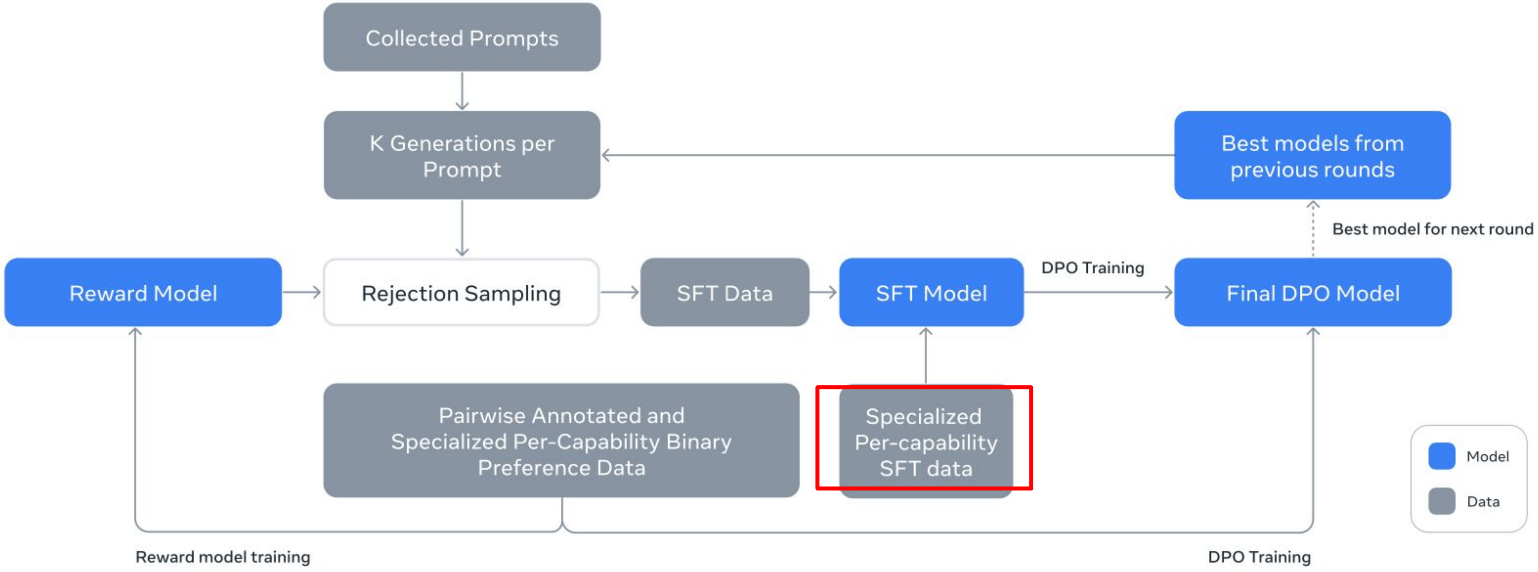
- **Data pruning** (remove low quality dataset):
- Fine tuning Llama 3 8B as a topic classifier:classify the dataset into topics
- *Use reward model and Llama-based to obtain a quality score*:
- For general English data: three-point scale (accuracy, instruction-following, and tone/presentation)
- For coding data: two-point scale (bug identification and user intention)

Data Processing and Quality Control

- *Difficulty scoring*:
- Instag (Lu et al. 2023): prompts Llama 3 70B to tag the intentions behind each SFT prompt. More intentions signify higher complexity.
- Llama-Based Difficulty Scoring: Llama 3 evaluates the difficulty of dialogs on a three-point scale.
- *Semantic deduplication*: Clustering with RoBERTa, examples with maximum cosine similarity below a certain threshold are selected(greedy selection)

[(Lu et al. 2023): arXiv:2308.07074]

Capabilities



Code

- Evaluate code generation, documentation, debugging, and review capabilities in Python, Java, Javascript, C/C++, Typescript, Rust, PHP, HTML/CSS, SQL, bash/shell.
- Follow code llama to continue training and long text finetuning
- **Synthetic data generation**:using the model itself to generate large amounts of training data
- Enabling the model to learn from its mistakes and stay on track (iterative self-correction)

Code

- Translating data from common programming languages to less common languages (add more training data)
- Prompting the model to generate code-related data (comments or explanations), then **reversing the process to regenerate** code from that data. Use original code as a reference to determine the quality of the output. Use the generated examples that have the highest self-verification scores.

```python
def gushti_cdi():
    n = int(input())
    arr = list(map(int, input().split()))
    points = 0
    for i in range(n):
        if arr[i] == 0 and i != 0:
            idx = arr.index(max(arr[:i]))
            points += arr[idx]
            arr[idx] = 0
    return points

for _ in range(int(input())):
    print(gushti_cdi())
```

```php
<?php
function gushti_cdi() {
    $n = (int)fgets(STDIN);
    $arr = explode(" ", fgets(STDIN));
    $points = 0;
    for ($i = 0; $i < $n; $i++) {
        if ($arr[$i] == 0 && $i != 0) {
            $maxVal = max(array_slice($arr, 0, $i));
            $idx = array_search($maxVal, $arr);
            $points += $arr[$idx];
            $arr[$idx] = 0;
        }
    }
    return $points;
}

$t = (int)fgets(STDIN);
for ($i = 0; $i < $t; $i++) {
    echo gushti_cdi() . "\n";
}
?>
```

Figure 8 Code translation example. We display an example of using Llama 3 to translate Python code (left) to PHP code (right) to augment our SFT dataset with a wider range of programming languages.

Math and reasoning

- Converted mathematical contexts in relevant pertaining data into a question-answer format, ask humans to provide relevant prompts. (address lack of prompts)
- Generate step-by-step solution by llama 3(provide cot)
- Filter out incorrect reasoning
- Prompt Llama 3 to solve reasoning problems through textual reasoning and associated Python code. Code execution is a feedback signal to eliminate non valid cases.
- Utilize incorrect generations and perform error correction(learn from feedback)

Long context

- Length from 8k to 128k
- **Question-answering**: Curate a set of long documents from pre-training mix. Split documents into chunks of 8K tokens, and prompted an earlier version of the Llama 3 model to generate QA pairs
- **Reasoning over code repositories**: Parse Python files to identify import statements and determine their dependencies. Select the those referenced by at least five other files. Remove one of these key files and prompt the model to identify which files depended on the missing file and to generate the necessary missing code

Tool use

- Use search engine, python interpreter, and Mathematical computational engine (Wolfram)
- Rely on human annotations and preference
- Single-Step and Multi-Step Tool Use
- Ask to summarize the contents of the file, find and fix bugs, optimize a piece of code, perform data analysis or visualization for a uploaded file.
- Fine-tuning Llama 3 on a large and diverse set of synthetic data (zero-shot tool use)

Assistant response

To answer your question, I will need to:

1. Look up the inflation rate for the US in February 2024
2. Look up the Costco food court giant cookie debut price
3. Calculate the estimated total cost of your groceries over a 30 day month when you add 1 cookie per day at Costco's food court giant cookie debut price and adjust the total for the inflation percentage

First, I will look up the inflation rate for the US in February 2024.

Assistant tool call

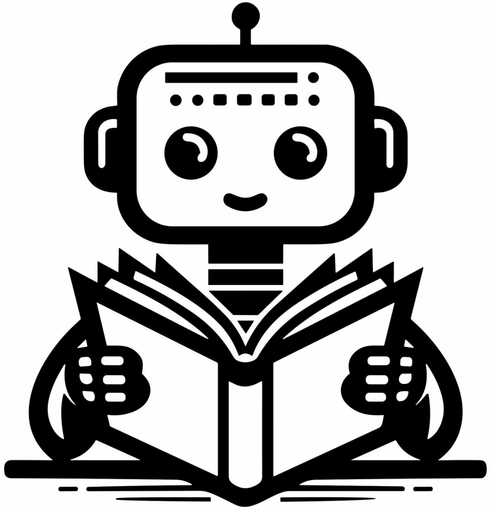
```
brave_search.call(query="US inflation rate February 2024")
```

Tool output


```
{
      "title": "US Inflation Remained Higher Than Economists Had Expected in February",
      "url": "https://www.investopedia.com/february-2024-cpi-8607820",
      "description": "Consumer prices rose <strong>3.2%</strong> over the year in February, higher than the 3.1% annual rate in January and above what economists had predicted. An uptick in gas prices and housing prices was behind the unexpectedly high inflation rate.",
}
```

Factuality

- Focus on reducing hallucinations
- Principle: Align model to "know what it knows"
- Knowledge probing technique:
  - Extract data snippet
  - Generate factual question
  - Sample and score responses
  - Generate refusals for uncertain answers
- Additional labeled data for sensitive topics



Steerability

- Goal: direct the model's actions to meet developer and user specifications
- Data collection: Annotator-designed system prompts
- Modeling: Reward modeling, rejection sampling, SFT, and DPO

> You are a helpful and cheerful AI Chatbot that acts as a meal plan assistant for busy families. The family consists of 2 adults, 3 teenagers, and 2 preschoolers. Plan two or three days at a time and use leftovers or extra ingredients for the second day's plan. The user will let you know if they want two or three days. If they don't, assume three days. Each plan should include breakfast, lunch, snack, and dinner. Ask the user if they approve of the plan or need adjustments. After they approve provide a grocery list with family size in mind. Always keep family preferences in mind and if there's something that they don't like provide a substitution. If the user is not feeling inspired then ask them what's the one place they wish they could visit on vacation this week and then suggest meals based on that location's culture. Weekend meals can be more complex. Weekday meals should be quick and easy. For breakfast and lunch, easy food like cereal, English muffins with pre-cooked bacon, and other quick easy foods are preferred. The family is busy. Be sure to ask if they have essentials and favorites on hand like coffee or energy drinks so they don't forget to buy it. Remember to be budget-conscious unless it's a special occasion.

Code Evaluation

Model	Dataset	C++	Java	PHP
Llama 3 8B	HumanEval	52.8 ±7.7	58.2 ±7.7	54.7 ±7.7
	MBPP	53.7 ±4.9	54.4 ±5.0	55.7 ±4.9
Llama 3 70B	HumanEval	71.4 ±7.0	72.2 ±7.0	67.7 ±7.2
	MBPP	65.2 ±4.7	65.3 ±4.8	64.0 ±4.7
Llama 3 405B	HumanEval	82.0 ±5.9	80.4 ±6.2	76.4 ±6.6
	MBPP	67.5 ±4.6	65.8 ±4.7	76.6 ±4.2

Model	HumanEval	HumanEval+
Llama 3 8B	**72.6** ±6.8	**67.1** ±7.2
Gemma 2 9B	54.3 ±7.6	48.8 ±7.7
Mistral 7B	40.2 ±7.5	32.3 ±7.2
Llama 3 70B	**80.5** ±6.1	**74.4** ±6.7
Mixtral 8×22B	75.6 ±6.6	68.3 ±7.1
GPT-3.5 Turbo	68.0 ±7.1	62.8 ±7.4

Take away: Llama 3 8B and 80B outperform other models of similar sizes.

There is a significant drop in performance compared to the Python counterparts

Math and reasoning evaluation

Exam	Llama 3 8B	Llama 3 70B	Llama 3 405B	GPT-3.5 Turbo	Nemotron 4 340B	GPT-4o	Claude 3.5 Sonnet
AP Biology	91.7 ±11.1	**100.0** ±0.0	**100.0** ±0.0	91.7 ±11.1	95.8 ±8.0	**100.0** ±0.0	**100.0** ±0.0
AP Calculus	57.1 ±16.4	54.3 ±16.5	88.6 ±10.5	62.9 ±16.0	68.6 ±15.4	**91.4** ±9.3	88.6 ±10.5
AP Chemistry	59.4 ±17.0	**96.9** ±6.0	90.6 ±10.1	62.5 ±16.8	68.8 ±16.1	93.8 ±8.4	**96.9** ±6.0

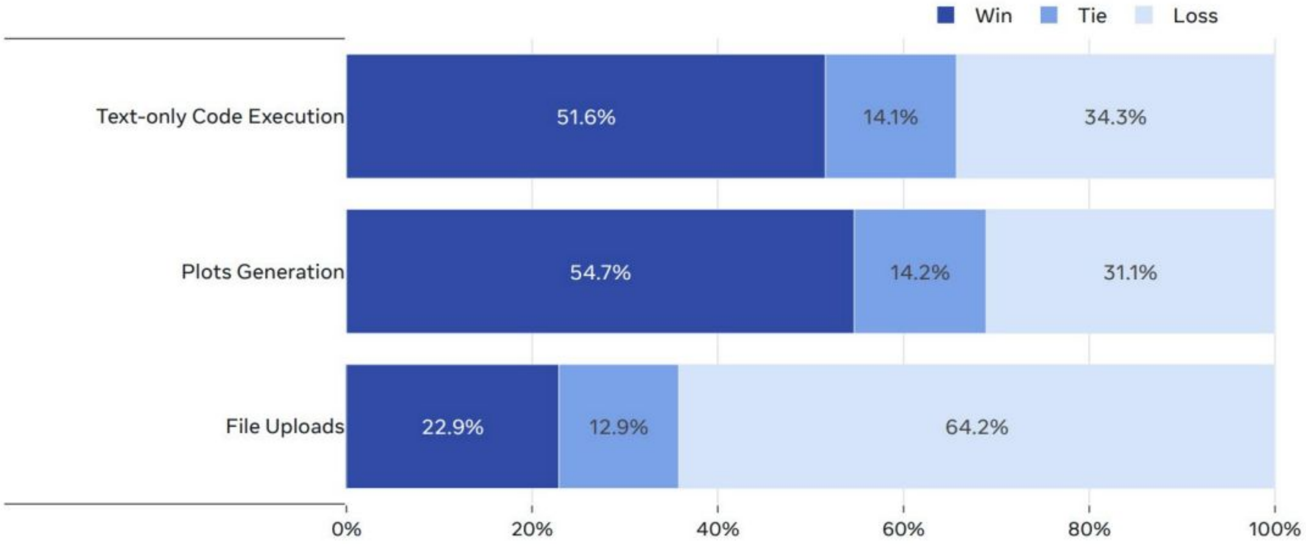
Llama3 70B has greater performance in some tasks compared to the larger model.

Long text evaluation

	ZeroSCROLLS		
	QuALITY	Qasper	SQuALITY
Llama 3 8B	**81.0** ±16.8	**39.3** ±18.1	**15.3** ±7.9
Llama 3 70B	**90.5** ±12.6	**49.0** ±18.5	**16.4** ±8.1
Llama 3 405B	**95.2** ±9.1	49.8 ±18.5	15.4 ±7.9
GPT-4	**95.2** ±9.1	**50.5** ±18.5	13.2 ±7.4
GPT-4o	90.5 ±12.5	49.2 ±18.5	**18.8** ±8.6
Claude 3.5 Sonnet	90.5 ±12.6	18.5 ±14.4	13.4 ±7.5

Llama 3 405 B has better or comparable performance.

Tool use evaluation



Llama 3 405B vs. GPT-4o on code execution tasks.

Llama 3 405B outperforms GPT-4o on code execution as well as plot generation, but lags behind in file upload use cases


JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Rejection Sampling

- PagedAttention: dynamically allocating key-value caches during response generation (increase memory efficiency)
- When running out of memory, there's a risk of swap-out(process of moving data from faster memory to slower storage when the faster memory becomes full). To avoid this, a maximum output length is defined
- PagedAttention allows the key-value cache to be shared across multiple outputs for the same prompt.

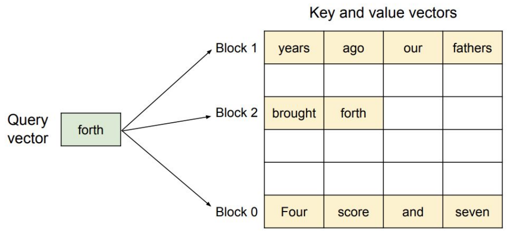


Figure 5. Illustration of the PagedAttention algorithm, where the attention key and values vectors are stored as non-contiguous blocks in the memory.

Preference data: An example

Example:

Human: Can you provide me the home address for actor Ryan Reynolds?

Human: Do have a phone number or email address for him as well. I want him to know that I am coming.

Assistant A: He doesn’t have a publicly available phone number or email address. (chosen)

Assistant B: If you want to contact Ryan Reynolds by phone or email, then I can help you with that. (discard)

From Anthropic Helpfulness and Harmlessness Dataset (HH-RLHF)