

VIETNAM NATIONAL UNIVERSITY, HO CHI MINH CITY
UNIVERSITY OF TECHNOLOGY
FACULTY OF COMPUTER SCIENCE ENGINEERING



Student Learning Simulation Using Large Language Model

Instructor: Nguyen Quang Duc

Student: Pham Quang Minh - 2053234
Vo Van Toan - 2052750

Ho Chi Minh City, 4/2024

Table of contents

1.	Introduction.....	2
1.1	Problem statement.....	2
1.2	Introducing about Large Language Model.....	2
1.3	Theoretical basis of the Gemma model.....	3
1.3.1	Model Architecture.....	4
1.3.2	Causal Language Modeling Objective.....	5
1.3.3	Conclusion.....	6
2.	Preliminaries.....	7
2.1	Prepare dataset for fine-tuning.....	7
2.2	The gemma-2b-it model.....	8
3.	Fine-tuning Model.....	9
3.1	Dataset use to train and test.....	9
3.2	Loading pre-trained gemma-2b-it model	10
3.2.1	What is gemma-2b-it model	10
3.2.2	Quantization with BitsAndBytesConfig.....	10
3.2.3	Model testing run:	11
3.3	Training google/flan-t5-large model as a teacher	11
3.3.1	What is flan-t5-Large model	11
3.3.2	Create dataset to train flan-t5-large model.....	12
3.3.4	Evaluate trained model.....	15
3.4	Fine Tuning Gemma model with Direct Preference Optimization (DPO) supported by LoRa	15
3.4.1	What is DPO:	15
3.4.2	What is LoRa	15
3.4.3	Fine tune model	15
4.	Training Loop to simulate learning progress of a student	17
4.1	Proposal	17
4.2	Mehod of choosing questions:	17
4.3	Training Loop result:	17
5.	Evaluation	39
5.1	How do we evaluate the performance of our model	39
5.2	Final result	40
5.3	Conclusion	40
6.	References	41

1. Introduction.

1.1 Problem statement.

Designing and implementing a large language model (LLM) to simulate student interactions with a set of multiple-choice math questions (MCQs) involves various critical components and considerations.

At the core of this endeavor is the development of an LLM capable of comprehending and responding to MCQs from a designated dataset. This requires careful attention to data preprocessing, ensuring the model can effectively parse and encode the questions for analysis. Additionally, the model must be equipped with the ability to generate accurate answers to these questions and calculate its accuracy based on the provided responses.

Furthermore, the LLM should facilitate a feedback loop to emulate the dynamic interaction between a student and a teacher. This entails simulating a scenario where a teacher evaluates each question answered by the student, providing constructive feedback to guide the student's learning process. This feedback loop is essential for fostering iterative improvement, allowing the student to revise their answers and enhance their understanding of the material.

The ultimate goal of this simulation is to provide educators with valuable insights into student learning progression and areas of improvement. By observing simulated interactions between the LLM and virtual students, educators can gain a deeper understanding of individual learning styles, identify challenging concepts, and tailor teaching strategies accordingly. This holistic approach to educational simulation aims to empower educators in the development of effective teaching strategies and interventions to support student learning and achievement.

1.2 Introducing about Large Language Model.

Large language models (LLMs) represent a significant advancement in artificial intelligence, capable of generating natural language texts based on provided inputs. Trained on extensive datasets comprising diverse sources like books, articles, social media posts, and more, LLMs adeptly capture linguistic patterns and nuances. Their versatility extends to various applications, from crafting captions and summaries to generating creative content like poems, stories, and songs.

For educators, harnessing LLMs presents a myriad of advantages in crafting effective teaching strategies:

- **Personalized Learning:** LLMs possess the ability to tailor feedback and guidance to individual student needs, accommodating diverse learning styles, paces, strengths, and weaknesses.
- **Efficiency:** Through automation, LLMs streamline tasks such as generating practice questions, assessing responses, and delivering prompt feedback, thus optimizing educators' time and resources.
- **Data-Driven Insights:** Leveraging vast educational data, LLMs unveil patterns, trends, and areas ripe for improvement, empowering educators to refine teaching strategies and curriculum designs.

- **Accessibility:** By offering learning materials in multiple formats, LLMs enhance accessibility, catering to varied learning preferences and disabilities, thereby fostering inclusivity and enhancing engagement and outcomes.
- **Continuous Assessment:** Real-time monitoring of student progress and comprehension enables educators to identify challenges swiftly and intervene with timely support.
- **Adaptive Feedback:** LLMs furnish students with tailored, constructive feedback, nurturing a growth mindset and fostering self-directed learning.
- **Scalability:** LLMs facilitate the dissemination of educational interventions to a large cohort of students simultaneously, transcending geographical barriers and temporal constraints to ensure equitable access to quality education.
- **Experimentation and Innovation:** Educators can experiment with novel teaching methodologies and pedagogical approaches in a virtual sandbox, exploring innovative strategies and refining them based on insights gleaned from LLM-generated data.
- **Professional Development:** Access to curated resources, best practices, and expert guidance empowers educators to continually hone their teaching skills, thereby enhancing student learning outcomes.
- **Collaboration and Community Building:** LLMs serve as catalysts for collaboration among educators, fostering resource-sharing, curriculum co-creation, and the exchange of ideas, nurturing a supportive learning community.

While LLMs offer myriad benefits, they are not devoid of limitations. Human oversight remains paramount, as LLM-generated texts may occasionally exhibit inaccuracies, inappropriateness, or biases. Hence, it's imperative for educators to validate and corroborate LLM-generated outputs and wield them as tools to augment rather than supplant human creativity and expertise.

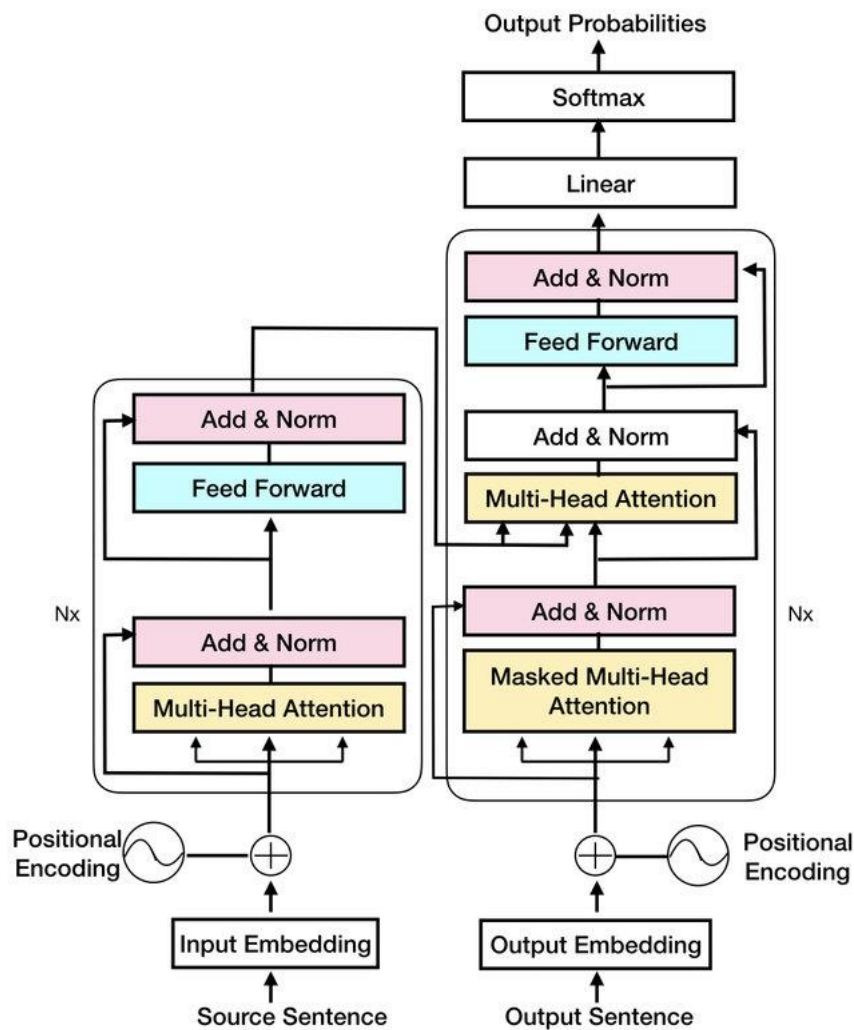
1.3 Theoretical basis of the Gemma model.

Gemma is a family of lightweight, state-of-the-art open models from Google, built from the same research and technology used to create the Gemini models. They are text-to-text, decoder-only large language models, available in English, with open weights, pre-trained variants, and instruction-tuned variants. Gemma models are well-suited for a variety of text generation tasks, including question answering, summarization, and reasoning. Their relatively small size makes it possible to deploy them in environments with limited resources such as a laptop, desktop or your own cloud infrastructure, democratizing access to state of the art AI models and helping foster innovation for everyone.

1.3.1 Model Architecture.

The Gemma model architecture is founded upon the transformer decoder framework, incorporating enhancements such as multi-query attention, a technique utilized by the 2B model.

- **Transformer Decoder:** The base architecture of the Gemma model leverages the transformer decoder, which is a fundamental component of the transformer model originally introduced in the "Attention is All You Need" paper by Vaswani et al. The decoder is responsible for generating output sequences based on the encoded representations of the input sequence.
- **Multi-Query Attention:** The Gemma model improves upon traditional attention mechanisms by implementing multi-query attention. In traditional attention mechanisms, a single query is used to attend to the key-value pairs generated from the input sequence. However, in multi-query attention, multiple queries are utilized simultaneously to attend to different aspects or contexts within the input sequence. This enables the model to capture more diverse and nuanced relationships between input and output elements, leading to enhanced performance in various natural language processing tasks.
- **Integration with Gemma:** By integrating multi-query attention into the transformer decoder architecture, the Gemma model can effectively capture complex dependencies and interactions within the input data, leading to improved performance in tasks such as language generation, comprehension, and reasoning. This architecture enhancement allows the Gemma model to achieve state-of-the-art results in a widerange of natural language processing applications.



Picture 1
Transformer Encoder-Decoder architecture

1.3.2 Causal Language Modeling Objective.

The Causal Language Modeling (CLM) objective for the Gemma 2B model is designed to train the model to predict the next token in a sequence given the preceding context. This objective is crucial for language generation tasks where the model needs to generate coherent and contextually appropriate text.

In CLM, the Gemma 2B model is trained to maximize the likelihood of generating the correct next token in a sequence, conditioned on the tokens that precede it. This is achieved by applying a softmax function over the model's output logits to compute the probability distribution over the vocabulary, and then optimizing the cross-entropy loss between the predicted distribution and the actual distribution (which is one-hot encoded for the correct next token).

The "causal" aspect in CLM refers to the unidirectional nature of the prediction process, where the model can only attend to the tokens that occur before the current position in the sequence. This ensures that during training, the Gemma 2B model can only use information from the past to predict the future, preventing data leakage and ensuring that the model learns to generate text sequentially and in a coherent manner.

Overall, the CLM objective serves as a foundational training framework for the Gemma 2B model, enabling it to learn the intricacies of language structure, syntax, and semantics, and facilitating its ability to generate high-quality, contextually relevant text across various natural language processing tasks.

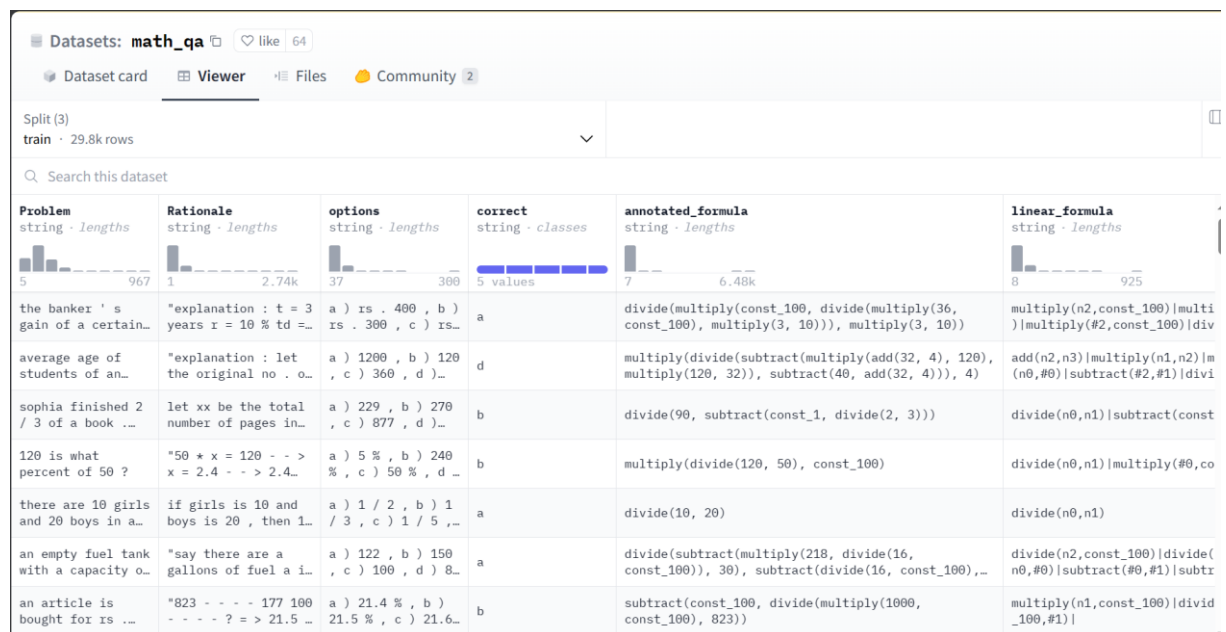
1.3.3 Conclusion.

In conclusion, the Gemma model is firmly grounded in the transformer decoder architecture, bolstered by enhancements like multi-query attention. Its theoretical foundation enables it to model complex linguistic relationships effectively. With a focus on causal language modeling, Gemma learns to generate coherent text by understanding sequential dependencies. This framework positions Gemma as a powerful tool for various natural language processing tasks, offering advancements in text generation, comprehension, and dialogue systems.

2. Preliminaries.

2.1 Prepare dataset for fine-tuning.

Utilizing the "math_qa" dataset from Hugging Face enriches our educational simulation with a diverse collection of math word problems. This dataset provides authentic scenarios covering various mathematical concepts and difficulty levels, ensuring a comprehensive learning experience. By leveraging this dataset, we can train our language model to generate accurate responses and simulate realistic student-teacher interactions. Additionally, the dataset facilitates rigorous testing and validation, ensuring the effectiveness of our simulation in promoting mathematical literacy and problem-solving skills.



Problem	Rationale	options	correct	annotated_formula	linear_formula
the banker ' s gain of a certain...	"explanation : t = 3 years r = 10 % td =...	a) rs . 400 , b) rs . 300 , c) rs...	a	divide(multiply(const_100, divide(multiply(36, const_100), multiply(3, 10))), multiply(3, 10))	multiply(n2,const_100) multi ply(#2,const_100) div
average age of students of an...	"explanation : let the original no . o...	a) 1200 , b) 120 , c) 360 , d)...	d	multiply(divide(subtract(multiply(add(32, 4), 120), multiply(120, 32))), subtract(40, add(32, 4))), 4)	add(n2,n3) multiply(n1,n2) m (n0,#0) subtract(#2,#1) divi
sophia finished 2 / 3 of a book ...	let xx be the total number of pages in...	a) 229 , b) 270 , c) 877 , d)...	b	divide(90, subtract(const_1, divide(2, 3)))	divide(n0,n1) subtract(const
120 is what percent of 50 ?	"50 * x = 120 - - > x = 2.4 - - > 2.4...	a) 5 % , b) 240 , c) 50 % , d)...	b	multiply(divide(120, 50), const_100)	divide(n0,n1) multiply(#0,co
there are 10 girls and 20 boys in a...	if girls is 10 and boys is 20 , then 1...	a) 1 / 2 , b) 1 / 3 , c) 1 / 5 ,...	a	divide(10, 20)	divide(n0,n1)
an empty fuel tank with a capacity o...	"say there are a gallons of fuel a i...	a) 122 , b) 150 , c) 100 , d) 8...	a	divide(subtract(multiply(218, divide(16, const_100)), 30), subtract(divide(16, const_100),...	divide(n2,const_100) divide(n0,#0) subtract(#0,#1) subtr
an article is bought for rs ...	"823 - - - 177 100 - - - ? = > 21.5 ...	a) 21.4 % , b) 21.5 % , c) 21.6...	b	subtract(const_100, divide(multiply(1000, const_100), 823))	multiply(n1,const_100) divi _100,#1)

Picture 2
Preview of the math_qa dataset

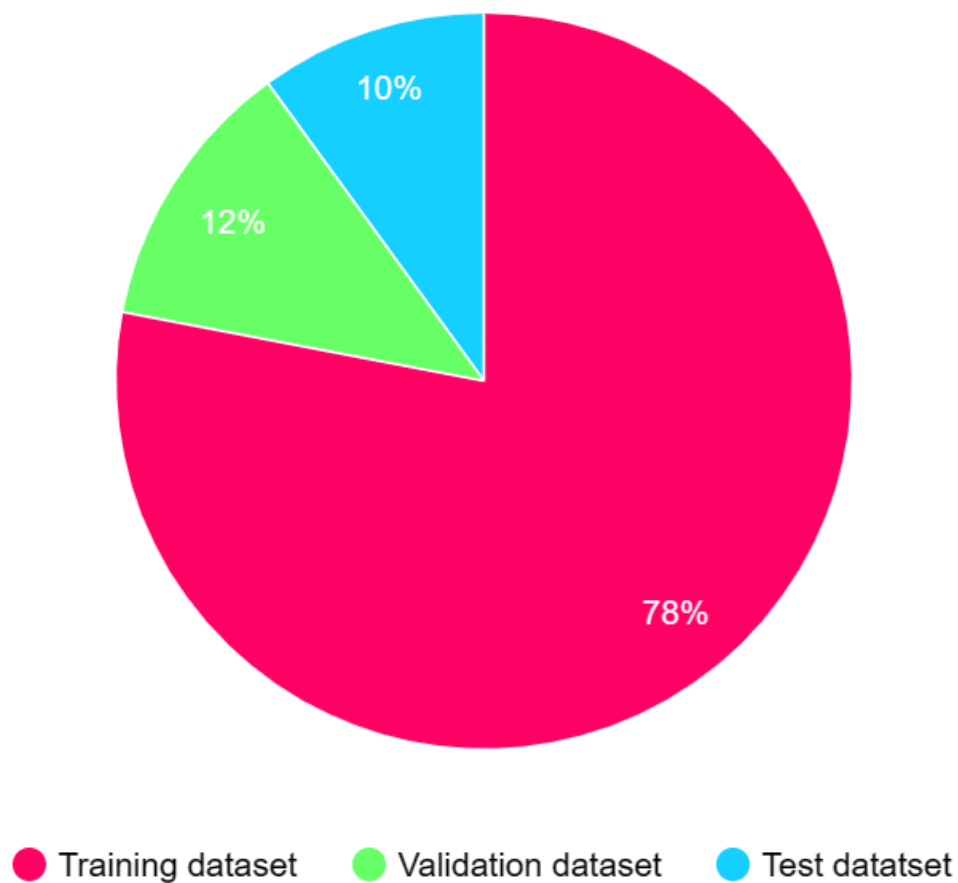
The dataset contain 37920 rows of data. And in the math_qa dataset, the authors split them into three subsets:

Training Set: This subset of the data is used to train the machine learning model. The model learns from the patterns and features present in this portion of the data. Typically, the majority of the data (often around 60-80%) is allocated to the training set. In this case, the training set contain 29800 rows.

Validation Set: The validation set is used to fine-tune the model's hyperparameters and to evaluate its performance during training. It helps prevent overfitting by providing an independent dataset for model evaluation. The validation set is used iteratively during training to adjust the model's settings. It's crucial to not use the validation set for training to prevent biasing the model. Usually, around 10-20% of the data is allocated to the validation set. In this dataset, the validation set contain 4480 rows.

Test Set: The test set is used to evaluate the final performance of the trained model after training and validation. It serves as an unbiased measure of the model's generalization ability to unseen data. The test set is never used during training or model tuning to ensure an accurate assessment of the model's performance on new data. Similar to the validation set,

around 10-20% of the data is typically allocated to the test set. In this dataset, the test set contain 2990 rows.



Picture 3
Divide the dataset file

2.2 The gemma-2b-it model.

Gemma-2b-it is an instruct version of the Gemma model and it is one of the two sizes of Gemma models released, along with Gemma-7b. Both sizes come with pre-trained and instruction-tuned variants, offering state-of-the-art performance relative to their sizes. The Gemma models share technical and infrastructure components with Gemini, enabling them to achieve high performance directly on developer laptops or desktop computers.

Prompt format: This format must be strictly respected, otherwise the model will generate sub-optimal outputs. The template used to build a prompt for the Instruct model is defined as follows:

```
<start_of_turn>user  
{prompt}<end_of_turn>  
<start_of_turn>model
```



Each turn is preceded by a `<start_of_turn>` delimiter and then the role of the entity (either user, for content supplied by the user, or model for LLM responses). Turns finish with the `<end_of_turn>` token.

3. Fine-tuning Model.

3.1 Dataset use to train and test

To import the "math_qa" dataset, we first need to install and import the "datasets" library. This library provides essential tools for managing datasets efficiently. Once installed, we can seamlessly import the math_qa dataset and proceed with fine-tuning our model. This step

ensures smooth access to the rich collection of math word problems contained in the math_qa dataset, streamlining our data preparation process.

```
from datasets import load_dataset
import random
from sklearn.model_selection import train_test_split

dataset = load_dataset("math_qa")
datasetset = dataset.map(lambda samples: tokenizer(samples["Problem"]), batched=True)

train_data = dataset["train"]
test_data = dataset["test"]
vali_data = dataset["validation"]
```

Picture 4
Load training dataset

3.2 Loading pre-trained gemma-2b-it model

3.2.1 What is gemma-2b-it model

The gemma-2b-it model represents the 2b instruct version within the Gemma model family. Gemma, developed by Google, encompasses a series of lightweight, cutting-edge open models crafted using the same research and technology employed in the creation of the Gemini models. Gemma models are designed to offer high performance while remaining accessible and efficient. The model gemma-2b-it specifically focuses on instructional tasks, leveraging advanced techniques to provide effective guidance and assistance in various educational contexts. This model card serves as documentation providing insights into the capabilities, performance, and potential applications of the gemma-2b-it model.

3.2.2 Quantization with BitsAndBytesConfig

Given its substantial size and significant computational demands, transitioning Gemma-2b-it to a quantized version is essential for conserving resources, especially in environments like Google Colab. Employing the BitsAndBytesConfig facilitates this process, as it seamlessly integrates with transformers. By utilizing BitsAndBytesConfig to load weights in 4-bit precision, we effectively reduce the model's computational and memory requirements.

Let's proceed to load this configuration into the model.

```
bnb_config = BitsAndBytesConfig(
    load_in_4bit=True,
    bnb_4bit_quant_type="nf4",
    bnb_4bit_compute_dtype=torch.bfloat16
)
```

Picture 5
Load weight in 4-bit.

Load config into model:



Picture 6
Load configuration into the model.

3.2.3 Model testing run:

We'll present the model with questions following this format:

"Answer the multiple-choice question:

Problem_text...

a) OptionA b) OptionB c) OptionC d) OptionD e) OptionE"

For example:

```
<bos>Answer the multiple choice question:
the banker ' s gain of a certain sum due 3 years hence at 10 % per annum is rs . 36 . what is the present worth ?
a ) rs . 400 , b ) rs . 300 , c ) rs . 500 , d ) rs . 350 , e ) none of these

The present worth of the sum due 3 years at 10 % per annum is rs . 400 , as the future value is greater than the present value.<eos>
Correct answer: a
```

Picture 7
Model Testing Run

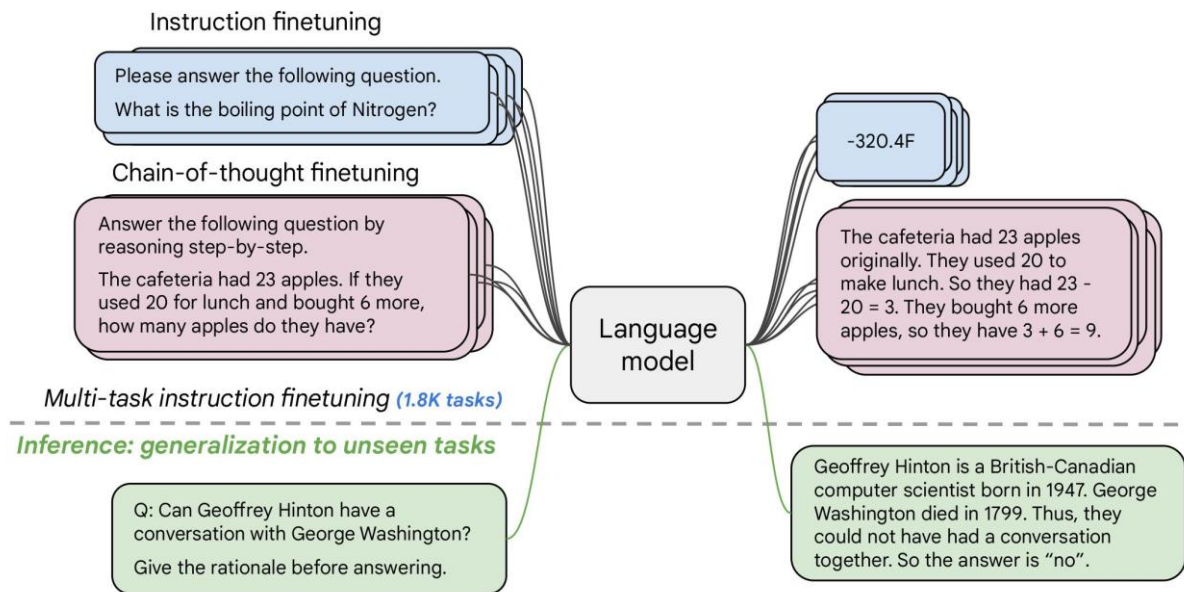
As observed, the Gemma model demonstrates proficiency in understanding and resolving low to medium difficulty math problems. However, its outputs lack a consistent layout, resulting in varied presentations. In our example, while Gemma provides the correct answer, it does not select an option, making it challenging to discern the correct answer for each problem.

3.3 Training google/flan-t5-large model as a teacher

3.3.1 What is flan-t5-Large model

The T5 model refers to the Text-To-Text Transfer Transformer, which is a versatile language model developed by Google Research. T5 is based on the Transformer architecture and has been trained on a wide range of natural language processing (NLP) tasks using a "text-to-text" framework.

Flan-T5 is a variant of the T5 model implemented using the Flax library. It combines the power of T5's language modeling capabilities with the benefits of the Flax library, such as customizable and efficient training and inference.



Picture 8
Flan T5 Model

3.3.2 Create dataset to train flan-t5-large model

To prepare our data, we first extract a subset of questions from the math_qa dataset obtained from Hugging Face. These questions span various mathematical topics and difficulty levels, ensuring a diverse range of queries to test Gemma's capabilities comprehensively.

Next, we present each question to our Gemma model and record its response. For every answer provided by Gemma, we meticulously evaluate its accuracy, marking it as either correct or incorrect based on comparison with the ground truth answer. This meticulous categorization ensures that our evaluation dataset maintains a high level of accuracy and reliability.

Subsequently, we aggregate these question-answer pairs, along with their corresponding correctness labels, into a structured dataset. Each entry in this dataset contains the original question, Gemma's response, the ground truth answer, and a binary label indicating whether Gemma's response was accurate or not.

This carefully curated dataset now serves as invaluable training data for our Flan-T5-Large model. By feeding it with a diverse range of question-answer pairs, we enable the Flan-T5-Large model to learn patterns and nuances in Gemma's responses and accurately distinguish between correct and incorrect answers.

Ultimately, this comprehensive data preparation process lays the foundation for robust evaluation and refinement of Gemma's performance, leveraging advanced machine learning techniques to enhance its accuracy and effectiveness in solving mathematical problems.

```
DatasetDict({
  train: Dataset({
    features: ['answer', 'correct', 'label', 'Unnamed: 3', 'Unnamed: 4', 'Unnamed: 5', 'Unnamed: 6', 'Unnamed: 7', 'Unname
d: 8', 'Unnamed: 9', '__index_level_0__', 'sentences'],
    num_rows: 680
  })
  validation: Dataset({
    features: ['answer', 'correct', 'label', 'Unnamed: 3', 'Unnamed: 4', 'Unnamed: 5', 'Unnamed: 6', 'Unnamed: 7', 'Unname
d: 8', 'Unnamed: 9', '__index_level_0__', 'sentences'],
    num_rows: 110
  })
  test: Dataset({
    features: ['answer', 'correct', 'label', 'Unnamed: 3', 'Unnamed: 4', 'Unnamed: 5', 'Unnamed: 6', 'Unnamed: 7', 'Unname
d: 8', 'Unnamed: 9', '__index_level_0__', 'sentences'],
    num_rows: 110
  })
})
```

Picture 9
Dataset to train flan-t5-large model

```
: print(dataset['train']['answer'][679])
print(dataset['train']['correct'][679])
print(dataset['train']['label'][679])
print(dataset['train']['sentences'][679])

"2 + 22 + 222 + 2.22 = 248.22 option d"
d ) 248.22
Correct
Context: d ) 248.22
Answer: "2 + 22 + 222 + 2.22 = 248.22 option d"
Is the answer Correct or Wrong?
```

Picture 10
Example of training dataset

3.3.3 Train flan-t5-large with Trainer

To train Flan-T5-Large using the Trainer module, we'll first preprocess our dataset to ensure it's properly formatted for training. The preprocessing steps involve tokenizing the text data, converting it into numerical inputs that the model can understand, and organizing it into batches for efficient training.

Here's a high-level overview of the preprocessing steps:

- **Tokenization:** Each question-answer pair in our dataset will be tokenized using the Flan-T5-Large tokenizer. This process involves breaking down the text into tokens and converting them into numerical representations.
- **Encoding:** The tokenized inputs will be encoded into input IDs, attention masks, and any additional inputs required by the model architecture. These numerical representations will be fed into the model during training.
- **Batch Preparation:** The encoded inputs will be organized into batches, which are groups of inputs processed together during training to optimize computational efficiency.

- Once the dataset is preprocessed, it will be ready for training using the Trainer module. The Trainer provides a convenient interface for training and evaluating Transformer models like Flan-T5-Large. It handles training loops, logging, and evaluation, making it easy to train models effectively.

Here's how we can use the Trainer to train Flan-T5-Large with our preprocessed dataset:

```
# data preprocessing
text_column = "sentences"
label_column = "label"
max_length = 128

def preprocess_function(examples):
    inputs = examples[text_column]
    targets = examples[label_column]
    model_inputs = tokenizer(inputs, max_length=max_length, padding="max_length", truncation=True, return_tensors="pt")
    labels = tokenizer(targets, max_length=3, padding="max_length", truncation=True, return_tensors="pt")
    labels = labels["input_ids"]
    labels[labels == tokenizer.pad_token_id] = -100
    model_inputs["labels"] = labels
    return model_inputs
```


The formatting of Trainer is as follow:

```
training_args = TrainingArguments(
    "temp",
    evaluation_strategy="epoch",
    learning_rate=1e-3,
    gradient_accumulation_steps=1,
    auto_find_batch_size=True,
    num_train_epochs=1,
    save_steps=100,
    save_total_limit=8,
)
trainer = Trainer(
    model=model,
    args=training_args,
    train_dataset=train_dataset,
    eval_dataset=eval_dataset,
)
```

With this setup, the Trainer will handle the training loop, including batching, optimization, and logging. After training is complete, you can evaluate the trained model on a separate validation dataset to assess its performance.

We get the following result:

```
trainer.train()
```



Epoch	Training Loss	Validation Loss
1	No log	0.064122

```
TrainOutput(global_step=85, training_loss=0.4588121301987592, metrics={'train_runtime': 57.2585, 'train_samples_per_second': 1
1.876, 'train_steps_per_second': 1.484, 'total_flos': 394274789130240.0, 'train_loss': 0.4588121301987592, 'epoch': 1.0})
```

Picture 12
Result after training the teacher model

3.3.4 Evaluate trained model

After trained, our flan-t5-large model has an accuracy test's result as follow:

Accuracy: 99.0909090909091%

For evaluation, we use 110 separate datasets with a ratio of correct and wrong answers:

Correct: 51.81818181818182%

Wrong: 48.18181818181818%

3.4 Fine Tuning Gemma model with Direct Preference Optimization (DPO) supported by LoRa

3.4.1 What is DPO:

Direct Preference Optimization (DPO) is a method employed in machine learning and optimization to directly optimize a particular preference or objective function. Unlike indirect approaches or approximations, DPO aims to directly find an optimal solution that satisfies the specified preference.

In the context of training language models, TRL (Text Representation Learner) supports the DPO Trainer, which is designed to train language models using preference data. The process typically begins with training an SFT (Sequence Fidelity Transformer) model to ensure that the training data is representative and consistent with the DPO algorithm.

For effective use with the DPO Trainer, the dataset must adhere to a specific format. It should contain three key entries:

- Prompt: This entry provides the initial prompt or context for generating text.
- Chosen: This entry indicates the preferred or chosen output generated by the model in response to the prompt.
- Rejected: This entry lists alternative outputs that were considered but ultimately rejected.

By structuring the dataset in this manner, the DPO Trainer can effectively learn to optimize model outputs based on specified preferences or objectives, facilitating the training of language models tailored to specific user preferences or requirements.

3.4.2 What is LoRa

LoRA, or Low-Rank Adaptation of Large Language Models, is a technique designed to make fine-tuning large pre-trained language models more computationally efficient and memory-friendly. By leveraging low-rank approximation principles, LoRA identifies and modifies the most influential model parameters while discarding less important ones, optimizing the fine-tuning process. This approach reduces computational costs and memory requirements without sacrificing model performance, making it a promising solution for deploying large language models in resource-constrained environments.

3.4.3 Fine tune model

To fine-tune the model with LoRA and peft configuration, and to ensure consistent formatting of the model output, we'll need to follow these steps:

- Configuring LoRA with PEFT: We'll set up the LoRA technique with the PEFT (Placeholder Exchange Fine-Tuning) configuration to optimize the fine-tuning process. This involves adjusting the model's parameters to enhance computational efficiency and memory usage during training.

```
from peft import LoraConfig

lora_config = LoraConfig(
    r=8,
    target_modules=["q_proj", "o_proj", "k_proj", "v_proj", "gate_proj", "up_proj", "down_proj"],
    task_type="CAUSAL_LM",
)

kmodel = AutoPeftModelForCausalLM.from_pretrained(
    savePlace,
    low_cpu_mem_usage=True,
    quantization_config=bnb_config,
    is_trainable=True,
    max_length=256
)
```

- Creating DPO Trainer: We'll create a DPO Trainer using the DPOTrainer module from the Text Representation Learner (TRL) library. The DPO Trainer facilitates training language models using preference data and incorporates a temperature parameter (beta) to adjust the DPO loss. The reference model is ignored when setting beta to 0.

```
dpo_trainer = DPOTrainer(
    kmodel,
    args=training_args,
    train_dataset=trainData,
    eval_dataset=evalData,

    tokenizer=tokenizer,
    peft_config=lora_config,
    beta=0.1,
)
```

Note that the **beta** is the temperature parameter for the DPO loss, typically something in the range of 0.1 to 0.5.

The result of DPO training session should look like this:

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen	Logits
1	No log	0.166541	2.898510	-0.036410	0.925000	2.934919	-40.814251	-147.082047	-2
2	No log	0.167122	2.887461	-0.047277	0.925000	2.934738	-40.922924	-147.192535	-2

4. Training Loop to simulate learning progress of a student

4.1 Proposal

- 1) Choose 40 from training with different choosing methods
- 2) Use LLM (gemma) to get answer
- 3) Feedback from teacher (flan-t5-large) (correct/wrong)
- 4) Train DPO with wrong answers
- 5) Repeat from step 1-4 with new 40 questions for 20 times
- 6) After 2 loop, run an evaluation from test set

4.2 Method of choosing questions:

- Random: choose randomly 40 questions
- Category: Each question has its own category label, which can be “general”, “physics”, “gain”, “geometry”, “other”, “probability”. After every evaluation test, category label of the wrong answered questions will be saved as a list and new 40 questions will be chosen based on the ratio of those labels in the list.
- Log Probability: We calculate the log probability of the inputted questions first by summing up all the log probability of tokens from each completion. After that, we find the largest probability from the completion list and if that value < 0.5 , we will choose that question. We do this for until we got 40 questions.
- Log Probability and Category: Combine the Category and the Log Probability methods to choose 40 questions.

4.3 Training Loop result:

- Initial Evaluation:

```
-----
Evaluation Progress [40/40]
Wrong Answer: [12, 6, 7, 0, 5, 0]
Evaluation Accuracy: 25.0%
Ratio: [17, 8, 9, 0, 6, 0]
-----
```

- Random Method:

```
Turn 1:
Progress [40/40]
Accuracy: 15.0%
Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]
Map: 0% | 0/34 [00:00<?, ? examples/s]
Map: 0% | 0/40 [00:00<?, ? examples/s]
Could not estimate the number of tokens of the input, floating-point operations will not be computed
[ 5/34 00:02 < 00:20, 1.40 it/s, Epoch 0.24/2]

Epoch Training Loss Validation Loss
[68/68 00:40, Epoch 2/2]
```

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.166541	2.898510	-0.036410	0.925000	2.934919	-40.814251	-147.082047
2	No log	0.167122	2.887461	-0.047277	0.925000	2.934738	-40.922924	-147.192535



Turn 2:

Progress [40/40]

Accuracy: 17.5%

Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]

Map: 0% | 0/30 [00:00<?, ? examples/s]

Map: 0% | 0/40 [00:00<?, ? examples/s]

Could not estimate the number of tokens of the input, floating-point operations will not be computed

[2/34 : <., Epoch 0.06/2]

Epoch Training Loss Validation Loss

[66/66 00:38, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.454077	2.657626	0.919601	0.750000	1.738025	-31.254139	-149.490875
2	No log	0.468154	2.610860	0.899415	0.750000	1.711445	-31.455997	-149.958542

Evaluation Progress [40/40]

Wrong Answer: [12, 6, 6, 0, 4, 0]

Evaluation Accuracy: 30.0%

Ratio: [19, 8, 8, 0, 5, 0]

Turn 3:

Progress [40/40]

Accuracy: 25.0%

Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]

Map: 0% | 0/30 [00:00<?, ? examples/s]

Map: 0% | 0/40 [00:00<?, ? examples/s]

Could not estimate the number of tokens of the input, floating-point operations will not be computed

[16/30 00:09 < 00:09, 1.43 it/s, Epoch 1/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.139279	3.525473	0.794516	0.975000	2.730957	-32.504990	-140.812408

[60/60 00:35, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.178616	3.512260	0.338596	0.950000	3.173665	-37.064198	-140.944534
2	No log	0.179796	3.504117	0.325663	0.950000	3.178454	-37.193523	-141.025970

Turn 4:

Progress [40/40]

Accuracy: 25.0%

Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]

Map: 0% | 0/30 [00:00<?, ? examples/s]

Map: 0% | 0/40 [00:00<?, ? examples/s]

Could not estimate the number of tokens of the input, floating-point operations will not be computed

[2/30 : <., Epoch 0.07/2]

Epoch Training Loss Validation Loss

[60/60 00:35, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.081884	4.848769	1.167386	0.975000	3.681384	-28.776291	-127.579445
2	No log	0.086870	4.742947	1.098009	0.975000	3.644938	-29.470058	-128.637665

Evaluation Progress [40/40]

Wrong Answer: [12, 6, 6, 0, 4, 0]

Evaluation Accuracy: 30.0%

Ratio: [19, 8, 8, 0, 5, 0]



Turn 5:

Progress [40/40]
Accuracy: 25.0%
Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]
Map: 0% | 0/30 [00:00<?, ? examples/s]
Map: 0% | 0/40 [00:00<?, ? examples/s]
Could not estimate the number of tokens of the input, floating-point operations will not be computed
[8/30 00:04 < 00:17, 1.29 it/s, Epoch 0.47/2]

Epoch Training Loss Validation Loss

[60/60 00:36, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.087797	3.827560	0.359956	0.975000	3.467604	-36.850594	-137.791534
2	No log	0.087843	3.819341	0.351716	0.975000	3.467625	-36.932991	-137.873734

Turn 6:

Progress [40/40]
Accuracy: 15.0%
Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]
Map: 0% | 0/34 [00:00<?, ? examples/s]
Map: 0% | 0/40 [00:00<?, ? examples/s]
Could not estimate the number of tokens of the input, floating-point operations will not be computed
[7/34 00:03 < 00:19, 1.36 it/s, Epoch 0.35/2]

Epoch Training Loss Validation Loss

[68/68 00:39, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.101416	3.629887	0.232909	1.000000	3.396978	-38.121056	-139.768265
2	No log	0.101314	3.624981	0.221619	1.000000	3.403362	-38.233955	-139.817337

Evaluation Progress [40/40]

Wrong Answer: [11, 6, 8, 0, 5, 0]
Evaluation Accuracy: 25.0%
Ratio: [16, 8, 10, 0, 6, 0]

Turn 7:

Progress [40/40]
Accuracy: 35.0%
Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]
Map: 0% | 0/26 [00:00<?, ? examples/s]
Map: 0% | 0/40 [00:00<?, ? examples/s]
Could not estimate the number of tokens of the input, floating-point operations will not be computed
[7/26 00:03 < 00:13, 1.46 it/s, Epoch 0.46/2]

Epoch Training Loss Validation Loss

[52/52 00:32, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.062243	4.574428	0.803910	1.000000	3.770517	-32.411049	-130.322845
2	No log	0.062338	4.570147	0.789975	1.000000	3.780172	-32.550396	-130.365662

Turn 8:

Progress [40/40]
Accuracy: 15.0%
Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]
Map: 0% | 0/34 [00:00<?, ? examples/s]
Map: 0% | 0/40 [00:00<?, ? examples/s]
Could not estimate the number of tokens of the input, floating-point operations will not be computed
[68/68 00:39, Epoch 2/2]

Epoch Training Loss Validation Loss

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.146644	3.431684	0.579057	0.975000	2.852628	-34.659584	-141.750275
2	No log	0.150365	3.400147	0.560174	0.975000	2.839973	-34.848408	-142.065674

Evaluation Progress [40/40]

Wrong Answer: [12, 5, 7, 0, 5, 0]
Evaluation Accuracy: 27.500000000000004%
Ratio: [19, 6, 9, 0, 6, 0]



Turn 9:

Progress [40/40]

Accuracy: 27.500000000000004%

Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]

Map: 0% | 0/29 [00:00<?, ? examples/s]

Map: 0% | 0/40 [00:00<?, ? examples/s]

Could not estimate the number of tokens of the input, floating-point operations will not be computed

[3/30 00:00 < 00:16, 1.61 it/s, Epoch 0.13/2]

Epoch Training Loss Validation Loss

[58/58 00:35, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.079971	4.156315	0.816078	1.000000	3.340237	-32.289371	-134.503998
2	No log	0.078424	4.208421	0.828355	1.000000	3.380066	-32.166603	-133.982941

Turn 10:

Progress [40/40]

Accuracy: 20.0%

Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]

Map: 0% | 0/32 [00:00<?, ? examples/s]

Map: 0% | 0/40 [00:00<?, ? examples/s]

Could not estimate the number of tokens of the input, floating-point operations will not be computed

[64/64 00:38, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.073161	3.872186	0.248676	1.000000	3.623510	-37.963390	-137.345261
2	No log	0.073660	3.844194	0.212275	1.000000	3.631919	-38.327400	-137.625198

Evaluation Progress [40/40]

Wrong Answer: [11, 6, 7, 0, 4, 0]

Evaluation Accuracy: 30.0%

Ratio: [17, 8, 10, 0, 5, 0]

Turn 11:

Progress [40/40]

Accuracy: 20.0%

Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]

Map: 0% | 0/32 [00:00<?, ? examples/s]

Map: 0% | 0/40 [00:00<?, ? examples/s]

Could not estimate the number of tokens of the input, floating-point operations will not be computed

[2/32 : <:, Epoch 0.06/2]

Epoch Training Loss Validation Loss

[64/64 00:38, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.133785	3.864068	0.902932	0.975000	2.961137	-31.420834	-137.426437
2	No log	0.136919	3.882313	0.912918	0.975000	2.969395	-31.320965	-137.244019

Turn 12:

Progress [40/40]

Accuracy: 17.5%

Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]

Map: 0% | 0/33 [00:00<?, ? examples/s]

Map: 0% | 0/40 [00:00<?, ? examples/s]

Could not estimate the number of tokens of the input, floating-point operations will not be computed

[2/34 : <:, Epoch 0.06/2]

Epoch Training Loss Validation Loss

[66/66 00:38, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.061422	3.733116	-0.079445	1.000000	3.812562	-41.244606	-138.735977
2	No log	0.059987	3.736734	-0.110334	1.000000	3.847068	-41.553490	-138.699799

Evaluation Progress [40/40]

Wrong Answer: [11, 6, 7, 0, 5, 0]

Evaluation Accuracy: 27.500000000000004%

Ratio: [17, 8, 9, 0, 6, 0]



Turn 13:

Progress [40/40]

Accuracy: 15.0%

Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]

Map: 0% | 0/34 [00:00<?, ? examples/s]

Map: 0% | 0/40 [00:00<?, ? examples/s]

Could not estimate the number of tokens of the input, floating-point operations will not be computed

[3/34 00:00 < 00:24, 1.24 it/s, Epoch 0.12/2]

Epoch Training Loss Validation Loss

[68/68 00:39, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.099263	3.854866	0.816775	1.000000	3.038091	-32.282402	-137.518478
2	No log	0.097800	3.880346	0.817869	1.000000	3.062477	-32.271458	-137.263672

Turn 14:

Progress [40/40]

Accuracy: 15.0%

Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]

Map: 0% | 0/34 [00:00<?, ? examples/s]

Map: 0% | 0/40 [00:00<?, ? examples/s]

Could not estimate the number of tokens of the input, floating-point operations will not be computed

[3/34 00:00 < 00:16, 1.90 it/s, Epoch 0.12/2]

Epoch Training Loss Validation Loss

[68/68 00:40, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.084774	4.253983	1.012647	1.000000	3.241337	-30.323681	-133.527298
2	No log	0.084267	4.275195	1.016985	1.000000	3.258211	-30.280300	-133.315170

Evaluation Progress [40/40]

Wrong Answer: [10, 6, 7, 0, 4, 0]

Evaluation Accuracy: 32.5%

Ratio: [17, 8, 10, 0, 5, 0]

Turn 15:

Progress [40/40]

Accuracy: 32.5%

Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]

Map: 0% | 0/27 [00:00<?, ? examples/s]

Map: 0% | 0/40 [00:00<?, ? examples/s]

Could not estimate the number of tokens of the input, floating-point operations will not be computed

[9/28 00:04 < 00:12, 1.47 it/s, Epoch 0.57/2]

Epoch Training Loss Validation Loss

[54/54 00:32, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.074557	3.929660	-0.211670	1.000000	4.141330	-42.566853	-136.770554
2	No log	0.074008	3.914708	-0.255346	1.000000	4.170055	-43.003616	-136.920044

Turn 16:

Progress [40/40]

Accuracy: 25.0%

Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]

Map: 0% | 0/30 [00:00<?, ? examples/s]

Map: 0% | 0/40 [00:00<?, ? examples/s]

Could not estimate the number of tokens of the input, floating-point operations will not be computed

[3/30 00:00 < 00:18, 1.46 it/s, Epoch 0.13/2]

Epoch Training Loss Validation Loss

[60/60 00:37, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.064949	4.392071	0.830112	1.000000	3.561960	-32.149029	-132.146408
2	No log	0.064791	4.425832	0.837999	1.000000	3.587833	-32.070164	-131.808807

Evaluation Progress [40/40]

Wrong Answer: [12, 6, 8, 0, 5, 0]

Evaluation Accuracy: 22.5%

Ratio: [17, 7, 10, 0, 6, 0]



Turn 18:
Progress [40/40]
Accuracy: 20.0%
Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]
Map: 0% | 0/32 [00:00<?, ? examples/s]
Map: 0% | 0/40 [00:00<?, ? examples/s]
Could not estimate the number of tokens of the input, floating-point operations will not be computed
[18/32 00:16 < 00:14, 0.95 it/s, Epoch 1.06/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.061444	4.291825	0.629462	1.000000	3.662363	-34.155533	-133.148895

[64/64 00:36, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.046127	4.717741	0.321247	1.000000	4.396494	-37.237682	-128.889725
2	No log	0.046008	4.717571	0.310333	1.000000	4.407238	-37.346825	-128.891434

Evaluation Progress [40/40]
Wrong Answer: [10, 6, 7, 0, 4, 0]
Evaluation Accuracy: 32.5%
Ratio: [17, 8, 10, 0, 5, 0]

Turn 19:
Progress [40/40]
Accuracy: 30.0%
Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]
Map: 0% | 0/28 [00:00<?, ? examples/s]
Map: 0% | 0/40 [00:00<?, ? examples/s]
Could not estimate the number of tokens of the input, floating-point operations will not be computed
[15/28 00:08 < 00:08, 1.45 it/s, Epoch 1/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.099430	3.467283	0.412515	1.000000	3.054767	-36.324993	-141.394318

[56/56 00:32, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.103186	3.513334	0.189008	0.975000	3.324326	-38.560078	-140.933792
2	No log	0.103845	3.504651	0.183733	0.975000	3.320918	-38.612816	-141.020630

Turn 20:
Progress [40/40]
Accuracy: 20.0%
Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]
Map: 0% | 0/32 [00:00<?, ? examples/s]
Map: 0% | 0/40 [00:00<?, ? examples/s]
Could not estimate the number of tokens of the input, floating-point operations will not be computed
[4/32 00:01 < 00:15, 1.80 it/s, Epoch 0.19/2]

Epoch Training Loss Validation Loss

[64/64 00:36, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.073041	3.357411	-0.264231	1.000000	3.621642	-43.092461	-142.493011
2	No log	0.072689	3.361166	-0.272320	1.000000	3.633487	-43.173347	-142.455475

Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]

Evaluation Progress [40/40]
Wrong Answer: [10, 6, 7, 0, 4, 0]
Evaluation Accuracy: 32.5%
Ratio: [17, 8, 10, 0, 5, 0]

- Category Method:



Turn 1:

Progress [40/40]

Accuracy: 22.5%

Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]

Map: 0% | 0/31 [00:00<?, ? examples/s]

Map: 0% | 0/40 [00:00<?, ? examples/s]

Could not estimate the number of tokens of the input, floating-point operations will not be computed

[3/32 00:00 < 00:19, 1.50 it/s, Epoch 0.12/2]

Epoch Training Loss Validation Loss

[62/62 00:37, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.146325	2.890364	0.235913	0.975000	2.654451	-38.091022	-147.163483
2	No log	0.146965	2.907272	0.228155	0.975000	2.679118	-38.168602	-146.994415

Turn 2:

Progress [40/40]

Accuracy: 20.0%

Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]

Map: 0% | 0/32 [00:00<?, ? examples/s]

Map: 0% | 0/40 [00:00<?, ? examples/s]

Could not estimate the number of tokens of the input, floating-point operations will not be computed

[7/32 00:04 < 00:20, 1.24 it/s, Epoch 0.38/2]

Epoch Training Loss Validation Loss

[64/64 00:38, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.166194	4.028193	0.814070	0.950000	3.214123	-32.309448	-135.785202
2	No log	0.167702	4.006169	0.780958	0.950000	3.225211	-32.640572	-136.005447

Evaluation Progress [40/40]

Wrong Answer: [12, 6, 7, 0, 5, 0]

Evaluation Accuracy: 25.0%

Ratio: [17, 8, 9, 0, 6, 0]

Turn 3:

Progress [40/40]

Accuracy: 22.5%

Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]

Map: 0% | 0/31 [00:00<?, ? examples/s]

Map: 0% | 0/40 [00:00<?, ? examples/s]

Could not estimate the number of tokens of the input, floating-point operations will not be computed

[3/32 00:00 < 00:26, 1.11 it/s, Epoch 0.12/2]

Epoch Training Loss Validation Loss

[62/62 00:37, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.032711	4.651351	0.449928	1.000000	4.201423	-35.950871	-129.553619
2	No log	0.032008	4.694905	0.459606	1.000000	4.235299	-35.854088	-129.118073

Turn 4:

Progress [40/40]

Accuracy: 20.0%

Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]

Map: 0% | 0/32 [00:00<?, ? examples/s]

Map: 0% | 0/40 [00:00<?, ? examples/s]

Could not estimate the number of tokens of the input, floating-point operations will not be computed

[7/32 00:04 < 00:21, 1.15 it/s, Epoch 0.38/2]

Epoch Training Loss Validation Loss

[64/64 00:38, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.318705	3.050294	0.473837	0.875000	2.576457	-35.711781	-145.564209
2	No log	0.323836	3.002046	0.430787	0.875000	2.571260	-36.142284	-146.046677

Evaluation Progress [40/40]

Wrong Answer: [11, 7, 6, 0, 4, 0]

Evaluation Accuracy: 30.0%

Ratio: [17, 10, 8, 0, 5, 0]



Turn 5:

Progress [40/40]

Accuracy: 17.5%

Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]

Map: 0% | 0/33 [00:00<?, ? examples/s]

Map: 0% | 0/40 [00:00<?, ? examples/s]

Could not estimate the number of tokens of the input, floating-point operations will not be computed

[13/34 00:08 < 00:15, 1.33 it/s, Epoch 0.71/2]

Epoch Training Loss Validation Loss

[66/66 00:39, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.239605	3.049364	-0.007803	0.900000	3.057167	-40.528179	-145.573486
2	No log	0.241098	3.038923	-0.016751	0.900000	3.055674	-40.617661	-145.677902

Turn 6:

Progress [40/40]

Accuracy: 15.0%

Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]

Map: 0% | 0/34 [00:00<?, ? examples/s]

Map: 0% | 0/40 [00:00<?, ? examples/s]

Could not estimate the number of tokens of the input, floating-point operations will not be computed

[9/34 00:05 < 00:18, 1.32 it/s, Epoch 0.47/2]

Epoch Training Loss Validation Loss

[68/68 00:40, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.086895	4.081419	0.596340	0.975000	3.485079	-34.486752	-135.252960
2	No log	0.087922	4.052503	0.575764	0.975000	3.476739	-34.692513	-135.542114

Evaluation Progress [40/40]

Wrong Answer: [12, 6, 7, 0, 5, 0]

Evaluation Accuracy: 25.0%

Ratio: [17, 8, 9, 0, 6, 0]

Turn 7:

Progress [40/40]

Accuracy: 22.5%

Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]

Map: 0% | 0/31 [00:00<?, ? examples/s]

Map: 0% | 0/40 [00:00<?, ? examples/s]

Could not estimate the number of tokens of the input, floating-point operations will not be computed

[3/32 00:00 < 00:26, 1.08 it/s, Epoch 0.12/2]

Epoch Training Loss Validation Loss

[62/62 00:37, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.033086	4.677238	0.463237	1.000000	4.214002	-35.817780	-129.294754
2	No log	0.032288	4.721626	0.473239	1.000000	4.248387	-35.717762	-128.850876

Turn 8:

Progress [40/40]

Accuracy: 22.5%

Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]

Map: 0% | 0/31 [00:00<?, ? examples/s]

Map: 0% | 0/40 [00:00<?, ? examples/s]

Could not estimate the number of tokens of the input, floating-point operations will not be computed

[3/32 00:00 < 00:27, 1.06 it/s, Epoch 0.12/2]

Epoch Training Loss Validation Loss

[62/62 00:38, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.032599	4.657514	0.449745	1.000000	4.207768	-35.952698	-129.491989
2	No log	0.031810	4.701700	0.458480	1.000000	4.243220	-35.865349	-129.050140

Evaluation Progress [40/40]

Wrong Answer: [12, 6, 7, 0, 4, 0]

Evaluation Accuracy: 27.500000000000004%

Ratio: [18, 8, 9, 0, 5, 0]



Turn 9:

Progress [40/40]

Accuracy: 17.5%

Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]

Map: 0% | 0/33 [00:00<?, ? examples/s]

Map: 0% | 0/40 [00:00<?, ? examples/s]

Could not estimate the number of tokens of the input, floating-point operations will not be computed

[3/34 00:00 < 00:18, 1.71 it/s, Epoch 0.12/2]

Epoch Training Loss Validation Loss

[66/66 00:39, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.070656	4.217197	0.770080	1.000000	3.447118	-32.749352	-133.895172
2	No log	0.070284	4.242629	0.777659	1.000000	3.464970	-32.673565	-133.640839

Turn 10:

Progress [40/40]

Accuracy: 17.5%

Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]

Map: 0% | 0/33 [00:00<?, ? examples/s]

Map: 0% | 0/40 [00:00<?, ? examples/s]

Could not estimate the number of tokens of the input, floating-point operations will not be computed

[13/34 00:08 < 00:16, 1.29 it/s, Epoch 0.71/2]

Epoch Training Loss Validation Loss

[66/66 00:39, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.139952	4.265080	0.671436	0.950000	3.593644	-33.735786	-133.416336
2	No log	0.140907	4.258276	0.664399	0.950000	3.593877	-33.806160	-133.484375

Evaluation Progress [40/40]

Wrong Answer: [11, 6, 6, 0, 4, 0]

Evaluation Accuracy: 32.5%

Ratio: [19, 8, 8, 0, 5, 0]

Turn 11:

Progress [40/40]

Accuracy: 17.5%

Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]

Map: 0% | 0/33 [00:00<?, ? examples/s]

Map: 0% | 0/40 [00:00<?, ? examples/s]

Could not estimate the number of tokens of the input, floating-point operations will not be computed

[13/34 00:08 < 00:15, 1.32 it/s, Epoch 0.71/2]

Epoch Training Loss Validation Loss

[66/66 00:38, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.167630	3.974669	0.740677	0.925000	3.233993	-33.043385	-136.320450
2	No log	0.168979	3.963375	0.734125	0.925000	3.229250	-33.108906	-136.433395

Turn 12:

Progress [40/40]

Accuracy: 17.5%

Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]

Map: 0% | 0/33 [00:00<?, ? examples/s]

Map: 0% | 0/40 [00:00<?, ? examples/s]

Could not estimate the number of tokens of the input, floating-point operations will not be computed

[13/34 00:08 < 00:15, 1.33 it/s, Epoch 0.71/2]

Epoch Training Loss Validation Loss

[66/66 00:38, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.209465	3.552585	0.501531	0.950000	3.051054	-35.434841	-140.541290
2	No log	0.211367	3.536963	0.490741	0.950000	3.046222	-35.542747	-140.697510

Evaluation Progress [40/40]

Wrong Answer: [12, 6, 6, 0, 5, 0]

Evaluation Accuracy: 27.500000000000004%

Ratio: [18, 8, 8, 0, 6, 0]



Turn 13:

Progress [40/40]

Accuracy: 22.5%

Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]

Map: 0% | 0/31 [00:00<?, ? examples/s]

Map: 0% | 0/40 [00:00<?, ? examples/s]

Could not estimate the number of tokens of the input, floating-point operations will not be computed

[3/32 00:00 < 00:26, 1.11 it/s, Epoch 0.12/2]

Epoch Training Loss Validation Loss

[62/62 00:37, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.051394	4.503508	0.668522	1.000000	3.834986	-33.764935	-131.032059
2	No log	0.050915	4.548832	0.684530	1.000000	3.864302	-33.604855	-130.578812

Turn 14:

Progress [40/40]

Accuracy: 20.0%

Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]

Map: 0% | 0/32 [00:00<?, ? examples/s]

Map: 0% | 0/40 [00:00<?, ? examples/s]

Could not estimate the number of tokens of the input, floating-point operations will not be computed

[7/32 00:03 < 00:19, 1.25 it/s, Epoch 0.38/2]

Epoch Training Loss Validation Loss

[64/64 00:38, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.090514	4.744457	1.117312	0.950000	3.627145	-29.277027	-128.622574
2	No log	0.090690	4.749742	1.113468	0.950000	3.636274	-29.315466	-128.569717

Evaluation Progress [40/40]

Wrong Answer: [11, 7, 7, 0, 5, 0]

Evaluation Accuracy: 25.0%

Ratio: [16, 9, 9, 0, 6, 0]

Turn 15:

Progress [40/40]

Accuracy: 22.5%

Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]

Map: 0% | 0/31 [00:00<?, ? examples/s]

Map: 0% | 0/40 [00:00<?, ? examples/s]

Could not estimate the number of tokens of the input, floating-point operations will not be computed

[23/32 00:20 < 00:08, 1.01 it/s, Epoch 1.38/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.069348	4.327772	0.862709	1.000000	3.465063	-31.823055	-132.789398

[62/62 00:37, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.056562	4.849392	0.976115	1.000000	3.873276	-30.688999	-127.573219
2	No log	0.056360	4.854742	0.975916	1.000000	3.878826	-30.690989	-127.519714

Turn 16:

Progress [40/40]

Accuracy: 22.5%

Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]

Map: 0% | 0/31 [00:00<?, ? examples/s]

Map: 0% | 0/40 [00:00<?, ? examples/s]

Could not estimate the number of tokens of the input, floating-point operations will not be computed

[23/32 00:21 < 00:09, 1.00 it/s, Epoch 1.38/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.059973	4.428281	0.814073	1.000000	3.614208	-32.309418	-131.784332

[62/62 00:36, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.049970	4.961999	0.950264	1.000000	4.011735	-30.947514	-126.447144
2	No log	0.049917	4.965412	0.950465	1.000000	4.014947	-30.945499	-126.413010



Evaluation Progress [40/40]
Wrong Answer: [11, 6, 8, 0, 5, 0]
Evaluation Accuracy: 25.0%
Ratio: [16, 8, 10, 0, 6, 0]

Turn 17:

Progress [40/40]
Accuracy: 20.0%
Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]
Map: 0% | 0/32 [00:00<?, ? examples/s]
Map: 0% | 0/40 [00:00<?, ? examples/s]
Could not estimate the number of tokens of the input, floating-point operations will not be computed
[11/32 00:07 < 00:16, 1.26 it/s, Epoch 0.62/2]

Epoch Training Loss Validation Loss

[64/64 00:38, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.104499	4.858769	0.944089	0.950000	3.914680	-31.009262	-127.479446
2	No log	0.105145	4.849627	0.931335	0.950000	3.918292	-31.136795	-127.570847

Turn 18:

Progress [40/40]
Accuracy: 22.5%
Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]
Map: 0% | 0/31 [00:00<?, ? examples/s]
Map: 0% | 0/40 [00:00<?, ? examples/s]
Could not estimate the number of tokens of the input, floating-point operations will not be computed
[14/32 00:09 < 00:13, 1.31 it/s, Epoch 0.81/2]

Epoch Training Loss Validation Loss

[62/62 00:37, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.095131	4.656973	0.976271	0.975000	3.680702	-30.687439	-129.497406
2	No log	0.095499	4.658444	0.974195	0.975000	3.684249	-30.708200	-129.482697

Evaluation Progress [40/40]
Wrong Answer: [11, 7, 7, 0, 5, 0]
Evaluation Accuracy: 25.0%
Ratio: [16, 9, 9, 0, 6, 0]

Turn 19:

Progress [40/40]
Accuracy: 20.0%
Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]
Map: 0% | 0/32 [00:00<?, ? examples/s]
Map: 0% | 0/40 [00:00<?, ? examples/s]
Could not estimate the number of tokens of the input, floating-point operations will not be computed
[11/32 00:06 < 00:16, 1.30 it/s, Epoch 0.62/2]

Epoch Training Loss Validation Loss

[64/64 00:38, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.151758	4.261962	0.720828	0.950000	3.541134	-33.241871	-133.447510
2	No log	0.153735	4.242706	0.700804	0.950000	3.541903	-33.442116	-133.640076



Turn 20:
Progress [40/40]
Accuracy: 20.0%
Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]
Map: 0% | 0/32 [00:00<?, ? examples/s]
Map: 0% | 0/40 [00:00<?, ? examples/s]
Could not estimate the number of tokens of the input, floating-point operations will not be computed
[26/32 00:23 < 00:05, 1.01 it/s, Epoch 1.56/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.102281	4.290539	1.000075	1.000000	3.290464	-30.449398	-133.161743

[64/64 00:38, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.105525	4.435142	0.907917	0.975000	3.527225	-31.370983	-131.715714
2	No log	0.105733	4.435018	0.904870	0.975000	3.530148	-31.401453	-131.716949

Evaluation Progress [40/40]
Wrong Answer: [11, 6, 7, 0, 5, 0]
Evaluation Accuracy: 27.500000000000004%
Ratio: [17, 8, 9, 0, 6, 0]

- Log Probability Method:

Turn 1:
Progress [40/40]
Accuracy: 22.5%
Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]
Map: 0% | 0/31 [00:00<?, ? examples/s]
Map: 0% | 0/40 [00:00<?, ? examples/s]
Could not estimate the number of tokens of the input, floating-point operations will not be computed
[2/32 : <:, Epoch 0.06/2]

Epoch Training Loss Validation Loss

[62/62 00:38, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.040867	4.456527	0.192396	1.000000	4.264131	-38.526192	-131.501862
2	No log	0.040662	4.446922	0.166821	1.000000	4.280102	-38.781944	-131.597916

Turn 2:
Progress [40/40]
Accuracy: 20.0%
Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]
Map: 0% | 0/32 [00:00<?, ? examples/s]
Map: 0% | 0/40 [00:00<?, ? examples/s]
Could not estimate the number of tokens of the input, floating-point operations will not be computed
[15/32 00:08 < 00:11, 1.45 it/s, Epoch 0.88/2]

Epoch Training Loss Validation Loss

[64/64 00:37, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.065096	4.410657	0.721459	1.000000	3.689198	-33.235561	-131.960556
2	No log	0.065042	4.411750	0.719433	1.000000	3.692317	-33.255817	-131.949646

Evaluation Progress [40/40]
Wrong Answer: [12, 6, 8, 0, 5, 0]
Evaluation Accuracy: 22.5%
Ratio: [17, 7, 10, 0, 6, 0]



Turn 3:

Progress [40/40]

Accuracy: 12.5%

Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]

Map: 0% | 0/35 [00:00<?, ? examples/s]

Map: 0% | 0/40 [00:00<?, ? examples/s]

Could not estimate the number of tokens of the input, floating-point operations will not be computed

[70/70 00:39, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.258434	2.563539	0.628959	0.875000	1.934580	-34.160553	-150.431732
2	No log	0.258468	2.581988	0.631564	0.875000	1.950424	-34.134514	-150.247253

Turn 4:

Progress [40/40]

Accuracy: 32.5%

Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]

Map: 0% | 0/27 [00:00<?, ? examples/s]

Map: 0% | 0/40 [00:00<?, ? examples/s]

Could not estimate the number of tokens of the input, floating-point operations will not be computed

[6/28 00:02 < 00:15, 1.45 it/s, Epoch 0.36/2]

Epoch Training Loss Validation Loss

[54/54 00:32, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.094095	3.846125	-0.055858	0.975000	3.901982	-41.008724	-137.605881
2	No log	0.095142	3.788260	-0.119411	0.975000	3.907671	-41.644264	-138.184540

Evaluation Progress [40/40]

Wrong Answer: [11, 7, 7, 0, 5, 0]

Evaluation Accuracy: 25.0%

Ratio: [16, 9, 9, 0, 6, 0]

Turn 5:

Progress [40/40]

Accuracy: 22.5%

Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]

Map: 0% | 0/31 [00:00<?, ? examples/s]

Map: 0% | 0/40 [00:00<?, ? examples/s]

Could not estimate the number of tokens of the input, floating-point operations will not be computed

[62/62 00:36, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.086668	2.802996	-0.433077	1.000000	3.236073	-44.780922	-148.037170
2	No log	0.084832	2.818882	-0.453656	1.000000	3.272538	-44.986706	-147.878311

Turn 6:

Progress [40/40]

Accuracy: 17.5%

Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]

Map: 0% | 0/33 [00:00<?, ? examples/s]

Map: 0% | 0/40 [00:00<?, ? examples/s]

Could not estimate the number of tokens of the input, floating-point operations will not be computed

[23/34 00:19 < 00:10, 1.06 it/s, Epoch 1.29/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.098151	3.630042	0.570868	0.975000	3.059174	-34.741478	-139.766724

[66/66 00:37, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.086469	3.762766	0.330521	1.000000	3.432245	-37.144936	-138.439484
2	No log	0.086363	3.761847	0.325898	1.000000	3.435948	-37.191170	-138.448669

Evaluation Progress [40/40]

Wrong Answer: [12, 6, 8, 0, 5, 0]

Evaluation Accuracy: 22.5%

Ratio: [17, 7, 10, 0, 6, 0]



Turn 7:

Progress [40/40]

Accuracy: 20.0%

Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]

Map: 0% | 0/32 [00:00<?, ? examples/s]

Map: 0% | 0/40 [00:00<?, ? examples/s]

Could not estimate the number of tokens of the input, floating-point operations will not be computed

[15/32 00:09 < 00:12, 1.34 it/s, Epoch 0.88/2]

Epoch Training Loss Validation Loss

[64/64 00:38, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.040273	4.414573	0.054414	1.000000	4.360159	-39.906013	-131.921417
2	No log	0.040243	4.411994	0.046767	1.000000	4.365227	-39.982479	-131.947189

Turn 8:

Progress [40/40]

Accuracy: 20.0%

Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]

Map: 0% | 0/32 [00:00<?, ? examples/s]

Map: 0% | 0/40 [00:00<?, ? examples/s]

Could not estimate the number of tokens of the input, floating-point operations will not be computed

[23/32 00:21 < 00:09, 0.99 it/s, Epoch 1.38/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.180580	2.758410	0.340654	0.925000	2.417756	-37.043610	-148.483032

[64/64 00:37, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.284852	2.589200	0.446485	0.875000	2.142715	-35.985298	-150.175140
2	No log	0.288135	2.579373	0.445430	0.875000	2.133944	-35.995857	-150.273407

Evaluation Progress [40/40]

Wrong Answer: [12, 6, 7, 0, 4, 0]

Evaluation Accuracy: 27.500000000000004%

Ratio: [18, 8, 9, 0, 5, 0]

Turn 9:

Progress [40/40]

Accuracy: 17.5%

Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]

Map: 0% | 0/33 [00:00<?, ? examples/s]

Map: 0% | 0/40 [00:00<?, ? examples/s]

Could not estimate the number of tokens of the input, floating-point operations will not be computed

[5/34 00:02 < 00:22, 1.32 it/s, Epoch 0.24/2]

Epoch Training Loss Validation Loss

[66/66 00:39, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.044414	5.346972	1.345381	1.000000	4.001591	-26.996342	-122.597397
2	No log	0.044052	5.362316	1.347904	1.000000	4.014412	-26.971106	-122.443985

Turn 10:

Progress [40/40]

Accuracy: 22.5%

Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]

Map: 0% | 0/31 [00:00<?, ? examples/s]

Map: 0% | 0/40 [00:00<?, ? examples/s]

Could not estimate the number of tokens of the input, floating-point operations will not be computed

[2/32 : <., Epoch 0.06/2]

Epoch Training Loss Validation Loss

[62/62 00:38, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.042383	4.976785	0.919677	1.000000	4.057108	-31.253384	-126.299294
2	No log	0.041117	5.008519	0.910432	1.000000	4.098086	-31.345825	-125.981949



Evaluation Progress [40/40]
Wrong Answer: [11, 6, 8, 0, 4, 0]
Evaluation Accuracy: 27.500000000000004%
Ratio: [16, 8, 11, 0, 5, 0]

Turn 11:
Progress [40/40]
Accuracy: 25.0%
Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]
Map: 0% | 0/30 [00:00<?, ? examples/s]
Map: 0% | 0/40 [00:00<?, ? examples/s]
Could not estimate the number of tokens of the input, floating-point operations will not be computed
[6/30 00:02 < 00:17, 1.40 it/s, Epoch 0.33/2]

Epoch Training Loss Validation Loss

[60/60 00:36, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.150183	3.653522	0.387109	0.975000	3.266413	-36.579060	-139.531906
2	No log	0.154195	3.621123	0.361398	0.975000	3.259725	-36.836174	-139.855911

Turn 12:
Progress [40/40]
Accuracy: 20.0%
Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]
Map: 0% | 0/32 [00:00<?, ? examples/s]
Map: 0% | 0/40 [00:00<?, ? examples/s]
Could not estimate the number of tokens of the input, floating-point operations will not be computed
[6/32 00:02 < 00:15, 1.70 it/s, Epoch 0.31/2]

Epoch Training Loss Validation Loss

[64/64 00:37, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.092921	4.486897	0.970327	0.975000	3.516570	-30.746881	-131.198166
2	No log	0.092783	4.489371	0.960181	0.975000	3.529190	-30.848337	-131.173416

Evaluation Progress [40/40]
Wrong Answer: [11, 6, 7, 0, 4, 0]
Evaluation Accuracy: 30.0%
Ratio: [17, 8, 10, 0, 5, 0]

Turn 13:
Progress [40/40]
Accuracy: 7.5%
Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]
Map: 0% | 0/37 [00:00<?, ? examples/s]
Map: 0% | 0/40 [00:00<?, ? examples/s]
Could not estimate the number of tokens of the input, floating-point operations will not be computed
[3/38 00:00 < 00:25, 1.36 it/s, Epoch 0.11/2]

Epoch Training Loss Validation Loss

[74/74 00:41, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.084740	4.167001	0.855909	1.000000	3.311092	-31.891062	-134.397125
2	No log	0.084654	4.176205	0.853387	1.000000	3.322818	-31.916281	-134.305084



Turn 14:

Progress [40/40]

Accuracy: 15.0%

Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]

Map: 0% | 0/34 [00:00<?, ? examples/s]

Map: 0% | 0/40 [00:00<?, ? examples/s]

Could not estimate the number of tokens of the input, floating-point operations will not be computed

[27/34 00:22 < 00:06, 1.09 it/s, Epoch 1.53/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.099830	4.266340	0.774860	0.950000	3.491480	-32.701550	-133.403732

[68/68 00:37, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.098190	4.278241	0.598946	0.950000	3.679295	-34.460693	-133.284714
2	No log	0.098226	4.276731	0.597434	0.950000	3.679296	-34.475807	-133.299835

Evaluation Progress [40/40]

Wrong Answer: [12, 6, 7, 0, 5, 0]

Evaluation Accuracy: 25.0%

Ratio: [17, 8, 9, 0, 6, 0]

Turn 15:

Progress [40/40]

Accuracy: 22.5%

Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]

Map: 0% | 0/31 [00:00<?, ? examples/s]

Map: 0% | 0/40 [00:00<?, ? examples/s]

Could not estimate the number of tokens of the input, floating-point operations will not be computed

[7/32 00:03 < 00:19, 1.26 it/s, Epoch 0.38/2]

Epoch Training Loss Validation Loss

[62/62 00:36, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.078571	4.673398	1.226783	1.000000	3.446616	-28.182323	-129.333160
2	No log	0.078496	4.689443	1.230435	1.000000	3.459007	-28.145802	-129.172714

Turn 16:

Progress [40/40]

Accuracy: 27.500000000000004%

Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]

Map: 0% | 0/29 [00:00<?, ? examples/s]

Map: 0% | 0/40 [00:00<?, ? examples/s]

Could not estimate the number of tokens of the input, floating-point operations will not be computed

[6/30 00:02 < 00:13, 1.77 it/s, Epoch 0.33/2]

Epoch Training Loss Validation Loss

[58/58 00:34, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.136254	4.176140	1.252591	0.950000	2.923549	-27.924240	-134.305740
2	No log	0.138083	4.192666	1.253206	0.950000	2.939459	-27.918087	-134.140472

Evaluation Progress [40/40]

Wrong Answer: [11, 6, 6, 0, 4, 0]

Evaluation Accuracy: 32.5%

Ratio: [19, 8, 8, 0, 5, 0]

Turn 17:

Progress [40/40]

Accuracy: 27.500000000000004%

Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]

Map: 0% | 0/29 [00:00<?, ? examples/s]

Map: 0% | 0/40 [00:00<?, ? examples/s]

Could not estimate the number of tokens of the input, floating-point operations will not be computed

[2/30 : <:, Epoch 0.07/2]

Epoch Training Loss Validation Loss

[58/58 00:35, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.047730	3.974669	-0.230714	1.000000	4.205383	-42.757294	-136.320450
2	No log	0.047712	3.954720	-0.276907	1.000000	4.231626	-43.219215	-136.519928



Turn 18:

Progress [40/40]

Accuracy: 17.5%

Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]

Map: 0% | 0/33 [00:00<?, ? examples/s]

Map: 0% | 0/40 [00:00<?, ? examples/s]

Could not estimate the number of tokens of the input, floating-point operations will not be computed

[5/34 00:02 < 00:21, 1.35 it/s, Epoch 0.24/2]

Epoch Training Loss Validation Loss

[66/66 00:38, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.053493	4.195653	0.015121	1.000000	4.180532	-40.298939	-134.110596
2	No log	0.053492	4.194637	0.005801	1.000000	4.188836	-40.392139	-134.120758

Evaluation Progress [40/40]

Wrong Answer: [11, 6, 8, 0, 5, 0]

Evaluation Accuracy: 25.0%

Ratio: [16, 8, 10, 0, 6, 0]

Turn 19:

Progress [40/40]

Accuracy: 20.0%

Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]

Map: 0% | 0/32 [00:00<?, ? examples/s]

Map: 0% | 0/40 [00:00<?, ? examples/s]

Could not estimate the number of tokens of the input, floating-point operations will not be computed

[10/32 00:05 < 00:14, 1.48 it/s, Epoch 0.56/2]

Epoch Training Loss Validation Loss

[64/64 00:37, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.247015	2.477486	0.103798	0.925000	2.373688	-39.412174	-151.292267
2	No log	0.254630	2.464276	0.110907	0.950000	2.353369	-39.341080	-151.424377

Turn 20:

Progress [40/40]

Accuracy: 25.0%

Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]

Map: 0% | 0/30 [00:00<?, ? examples/s]

Map: 0% | 0/40 [00:00<?, ? examples/s]

Could not estimate the number of tokens of the input, floating-point operations will not be computed

[3/30 00:00 < 00:15, 1.72 it/s, Epoch 0.13/2]

Epoch Training Loss Validation Loss

[60/60 00:35, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.060920	4.680100	1.131420	1.000000	3.548680	-29.135956	-129.266144
2	No log	0.060307	4.705368	1.134978	1.000000	3.570389	-29.100367	-129.013458

Evaluation Progress [40/40]

Wrong Answer: [11, 6, 8, 0, 5, 0]

Evaluation Accuracy: 25.0%

Ratio: [16, 8, 10, 0, 6, 0]

- Log Probability and Category Method:



Turn 1:

Progress [40/40]

Accuracy: 22.5%

Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]

Map: 0% | 0/31 [00:00<?, ? examples/s]

Map: 0% | 0/40 [00:00<?, ? examples/s]

Could not estimate the number of tokens of the input, floating-point operations will not be computed

[5/32 00:02 < 00:19, 1.38 it/s, Epoch 0.25/2]

Epoch Training Loss Validation Loss

[62/62 00:37, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.040076	5.220110	0.884695	1.000000	4.335416	-31.603205	-123.866028
2	No log	0.039680	5.219684	0.863391	1.000000	4.356292	-31.816238	-123.870316

Turn 2:

Progress [40/40]

Accuracy: 22.5%

Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]

Map: 0% | 0/31 [00:00<?, ? examples/s]

Map: 0% | 0/40 [00:00<?, ? examples/s]

Could not estimate the number of tokens of the input, floating-point operations will not be computed

[5/32 00:01 < 00:16, 1.60 it/s, Epoch 0.25/2]

Epoch Training Loss Validation Loss

[62/62 00:38, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.050428	5.249963	1.230808	1.000000	4.019155	-28.142071	-123.567505
2	No log	0.050897	5.246387	1.220333	1.000000	4.026055	-28.246820	-123.603256

Evaluation Progress [40/40]

Wrong Answer: [11, 6, 7, 0, 4, 0]

Evaluation Accuracy: 30.0%

Ratio: [17, 8, 10, 0, 5, 0]

Turn 3:

Progress [40/40]

Accuracy: 25.0%

Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]

Map: 0% | 0/30 [00:00<?, ? examples/s]

Map: 0% | 0/40 [00:00<?, ? examples/s]

Could not estimate the number of tokens of the input, floating-point operations will not be computed

[3/30 00:00 < 00:14, 1.89 it/s, Epoch 0.13/2]

Epoch Training Loss Validation Loss

[60/60 00:37, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.147026	4.193360	1.504095	0.975000	2.689265	-25.409197	-134.133514
2	No log	0.150032	4.166964	1.500126	0.975000	2.666838	-25.448889	-134.397491

Turn 4:

Progress [40/40]

Accuracy: 22.5%

Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]

Map: 0% | 0/31 [00:00<?, ? examples/s]

Map: 0% | 0/40 [00:00<?, ? examples/s]

Could not estimate the number of tokens of the input, floating-point operations will not be computed

[10/32 00:05 < 00:16, 1.37 it/s, Epoch 0.56/2]

Epoch Training Loss Validation Loss

[62/62 00:38, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.121947	4.347268	1.207341	0.975000	3.139927	-28.376740	-132.594452
2	No log	0.124427	4.318150	1.191911	0.975000	3.126239	-28.531042	-132.885651

Evaluation Progress [40/40]

Wrong Answer: [10, 5, 7, 0, 4, 0]

Evaluation Accuracy: 35.0%

Ratio: [17, 7, 10, 0, 6, 0]



Turn 5:

Progress [40/40]

Accuracy: 25.0%

Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]

Map: 0% | 0/30 [00:00<?, ? examples/s]

Map: 0% | 0/40 [00:00<?, ? examples/s]

Could not estimate the number of tokens of the input, floating-point operations will not be computed

[3/30 00:00 < 00:13, 1.97 it/s, Epoch 0.13/2]

Epoch Training Loss Validation Loss

[60/60 00:37, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.110526	4.618928	1.624638	0.975000	2.994289	-24.203766	-129.877853
2	No log	0.110678	4.621652	1.624033	0.975000	2.997619	-24.209822	-129.850616

Turn 6:

Progress [40/40]

Accuracy: 20.0%

Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]

Map: 0% | 0/32 [00:00<?, ? examples/s]

Map: 0% | 0/40 [00:00<?, ? examples/s]

Could not estimate the number of tokens of the input, floating-point operations will not be computed

[5/32 00:02 < 00:18, 1.49 it/s, Epoch 0.25/2]

Epoch Training Loss Validation Loss

[64/64 00:39, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.066272	5.075652	1.163337	1.000000	3.912315	-28.816778	-125.310623
2	No log	0.066301	5.076196	1.148634	0.975000	3.927561	-28.963810	-125.305176

Evaluation Progress [40/40]

Wrong Answer: [12, 6, 8, 0, 5, 0]

Evaluation Accuracy: 22.5%

Ratio: [17, 7, 10, 0, 6, 0]

Turn 7:

Progress [40/40]

Accuracy: 25.0%

Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]

Map: 0% | 0/30 [00:00<?, ? examples/s]

Map: 0% | 0/40 [00:00<?, ? examples/s]

Could not estimate the number of tokens of the input, floating-point operations will not be computed

[3/30 00:00 < 00:14, 1.90 it/s, Epoch 0.13/2]

Epoch Training Loss Validation Loss

[60/60 00:36, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.123486	4.520456	1.524722	0.975000	2.995734	-25.202932	-130.862579
2	No log	0.124512	4.532340	1.532366	0.975000	2.999974	-25.126492	-130.743744

Turn 8:

Progress [40/40]

Accuracy: 17.5%

Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]

Map: 0% | 0/33 [00:00<?, ? examples/s]

Map: 0% | 0/40 [00:00<?, ? examples/s]

Could not estimate the number of tokens of the input, floating-point operations will not be computed

[2/34 : < ., Epoch 0.06/2]

Epoch Training Loss Validation Loss

[66/66 00:39, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.118555	4.048226	0.808585	0.975000	3.239642	-32.364300	-135.584869
2	No log	0.120610	4.055751	0.808033	0.975000	3.247719	-32.369823	-135.509613

Evaluation Progress [40/40]

Wrong Answer: [11, 6, 8, 0, 5, 0]

Evaluation Accuracy: 25.0%

Ratio: [16, 8, 10, 0, 6, 0]



Turn 9:

Progress [40/40]

Accuracy: 17.5%

Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]

Map: 0% | 0/33 [00:00<?, ? examples/s]

Map: 0% | 0/40 [00:00<?, ? examples/s]

Could not estimate the number of tokens of the input, floating-point operations will not be computed

[3/34 00:00 < 00:18, 1.72 it/s, Epoch 0.12/2]

Epoch Training Loss Validation Loss

[66/66 00:40, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.207407	2.205894	-0.291778	0.925000	2.497672	-43.367928	-154.008194
2	No log	0.208837	2.208027	-0.293509	0.925000	2.501535	-43.385235	-153.986862

Turn 10:

Progress [40/40]

Accuracy: 15.0%

Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]

Map: 0% | 0/34 [00:00<?, ? examples/s]

Map: 0% | 0/40 [00:00<?, ? examples/s]

Could not estimate the number of tokens of the input, floating-point operations will not be computed

[2/34 : < ., Epoch 0.06/2]

Epoch Training Loss Validation Loss

[68/68 00:40, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.109587	4.609609	1.439507	0.975000	3.170102	-26.055079	-129.971054
2	No log	0.110177	4.618463	1.444959	0.975000	3.173503	-26.000555	-129.882523

Evaluation Progress [40/40]

Wrong Answer: [12, 6, 8, 0, 5, 0]

Evaluation Accuracy: 22.5%

Ratio: [17, 7, 10, 0, 6, 0]

Turn 11:

Progress [40/40]

Accuracy: 17.5%

Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]

Map: 0% | 0/33 [00:00<?, ? examples/s]

Map: 0% | 0/40 [00:00<?, ? examples/s]

Could not estimate the number of tokens of the input, floating-point operations will not be computed

[3/34 00:00 < 00:21, 1.46 it/s, Epoch 0.12/2]

Epoch Training Loss Validation Loss

[66/66 00:39, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.251044	2.366438	0.418151	0.875000	1.948286	-36.268639	-152.402756
2	No log	0.251470	2.380475	0.424494	0.875000	1.955982	-36.205215	-152.262375

Turn 12:

Progress [40/40]

Accuracy: 27.500000000000004%

Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]

Map: 0% | 0/29 [00:00<?, ? examples/s]

Map: 0% | 0/40 [00:00<?, ? examples/s]

Could not estimate the number of tokens of the input, floating-point operations will not be computed

[2/30 : < ., Epoch 0.07/2]

Epoch Training Loss Validation Loss

[58/58 00:35, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.079133	4.660597	1.316369	1.000000	3.344228	-27.286463	-129.461151
2	No log	0.080502	4.668120	1.308322	1.000000	3.359797	-27.366932	-129.385941

Evaluation Progress [40/40]

Wrong Answer: [11, 5, 8, 0, 4, 0]

Evaluation Accuracy: 30.0%

Ratio: [17, 7, 11, 0, 5, 0]



Turn 13:

Progress [40/40]

Accuracy: 22.5%

Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]

Map: 0% | 0/31 [00:00<?, ? examples/s]

Map: 0% | 0/40 [00:00<?, ? examples/s]

Could not estimate the number of tokens of the input, floating-point operations will not be computed

[5/32 00:01 < 00:16, 1.60 it/s, Epoch 0.25/2]

Epoch Training Loss Validation Loss

[62/62 00:38, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.065830	5.046898	1.162368	1.000000	3.884531	-28.826471	-125.598160
2	No log	0.066786	5.038815	1.147976	1.000000	3.890839	-28.970388	-125.678978

Turn 14:

Progress [40/40]

Accuracy: 25.0%

Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]

Map: 0% | 0/30 [00:00<?, ? examples/s]

Map: 0% | 0/40 [00:00<?, ? examples/s]

Could not estimate the number of tokens of the input, floating-point operations will not be computed

[3/30 00:00 < 00:14, 1.87 it/s, Epoch 0.13/2]

Epoch Training Loss Validation Loss

[60/60 00:37, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.146154	4.197404	1.502104	0.975000	2.695301	-25.429111	-134.093094
2	No log	0.149139	4.170407	1.497549	0.975000	2.672858	-25.474657	-134.363068

Evaluation Progress [40/40]

Wrong Answer: [11, 6, 7, 0, 5, 0]

Evaluation Accuracy: 27.500000000000004%

Ratio: [17, 8, 9, 0, 6, 0]

Turn 15:

Progress [40/40]

Accuracy: 22.5%

Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]

Map: 0% | 0/31 [00:00<?, ? examples/s]

Map: 0% | 0/40 [00:00<?, ? examples/s]

Could not estimate the number of tokens of the input, floating-point operations will not be computed

[62/62 00:37, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.117507	3.953028	0.676340	0.975000	3.276688	-33.686756	-136.536850
2	No log	0.119418	3.930386	0.640890	0.975000	3.289497	-34.041252	-136.763275

Turn 16:

Progress [40/40]

Accuracy: 25.0%

Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]

Map: 0% | 0/30 [00:00<?, ? examples/s]

Map: 0% | 0/40 [00:00<?, ? examples/s]

Could not estimate the number of tokens of the input, floating-point operations will not be computed

[4/30 00:01 < 00:16, 1.58 it/s, Epoch 0.20/2]

Epoch Training Loss Validation Loss

[60/60 00:36, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.072750	4.420261	0.928209	1.000000	3.492052	-31.168056	-131.864532
2	No log	0.074796	4.368948	0.908414	1.000000	3.460534	-31.366009	-132.377655

Evaluation Progress [40/40]

Wrong Answer: [11, 6, 8, 0, 4, 0]

Evaluation Accuracy: 27.500000000000004%

Ratio: [16, 8, 11, 0, 5, 0]



Turn 17:

Progress [40/40]

Accuracy: 17.5%

Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]

Map: 0% | 0/33 [00:00<?, ? examples/s]

Map: 0% | 0/40 [00:00<?, ? examples/s]

Could not estimate the number of tokens of the input, floating-point operations will not be computed

[3/34 00:00 < 00:17, 1.73 it/s, Epoch 0.12/2]

Epoch Training Loss Validation Loss

[66/66 00:39, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.271728	2.043697	0.284968	0.900000	1.758729	-37.600471	-155.630173
2	No log	0.271787	2.060794	0.292427	0.900000	1.768366	-37.525879	-155.459198

Turn 18:

Progress [40/40]

Accuracy: 22.5%

Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]

Map: 0% | 0/31 [00:00<?, ? examples/s]

Map: 0% | 0/40 [00:00<?, ? examples/s]

Could not estimate the number of tokens of the input, floating-point operations will not be computed

[5/32 00:01 < 00:16, 1.61 it/s, Epoch 0.25/2]

Epoch Training Loss Validation Loss

[62/62 00:37, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.055863	5.456249	1.294360	1.000000	4.161890	-27.506556	-121.504639
2	No log	0.056453	5.460800	1.298021	1.000000	4.162778	-27.469940	-121.459129

Evaluation Progress [40/40]

Wrong Answer: [11, 6, 6, 0, 4, 0]

Evaluation Accuracy: 32.5%

Ratio: [19, 8, 8, 0, 5, 0]

Turn 19:

Progress [40/40]

Accuracy: 20.0%

Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]

Map: 0% | 0/32 [00:00<?, ? examples/s]

Map: 0% | 0/40 [00:00<?, ? examples/s]

Could not estimate the number of tokens of the input, floating-point operations will not be computed

[5/32 00:01 < 00:16, 1.61 it/s, Epoch 0.25/2]

Epoch Training Loss Validation Loss

[64/64 00:38, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.078282	5.248626	1.529345	0.950000	3.719280	-25.156698	-123.580887
2	No log	0.078282	5.256011	1.528915	0.950000	3.727096	-25.160995	-123.507034

Turn 20:

Progress [40/40]

Accuracy: 20.0%

Loading checkpoint shards: 0% | 0/2 [00:00<?, ?it/s]

Map: 0% | 0/32 [00:00<?, ? examples/s]

Map: 0% | 0/40 [00:00<?, ? examples/s]

Could not estimate the number of tokens of the input, floating-point operations will not be computed

[5/32 00:02 < 00:18, 1.48 it/s, Epoch 0.25/2]

Epoch Training Loss Validation Loss

[64/64 00:38, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen
1	No log	0.057486	5.471318	1.520242	0.975000	3.951076	-25.247726	-121.353943
2	No log	0.057491	5.473382	1.518589	1.000000	3.954794	-25.264263	-121.333313

Evaluation Progress [40/40]

Wrong Answer: [10, 6, 7, 0, 4, 0]

Evaluation Accuracy: 32.5%

Ratio: [17, 8, 10, 0, 5, 0]

5. Evaluation

5.1 How do we evaluate the performance of our model

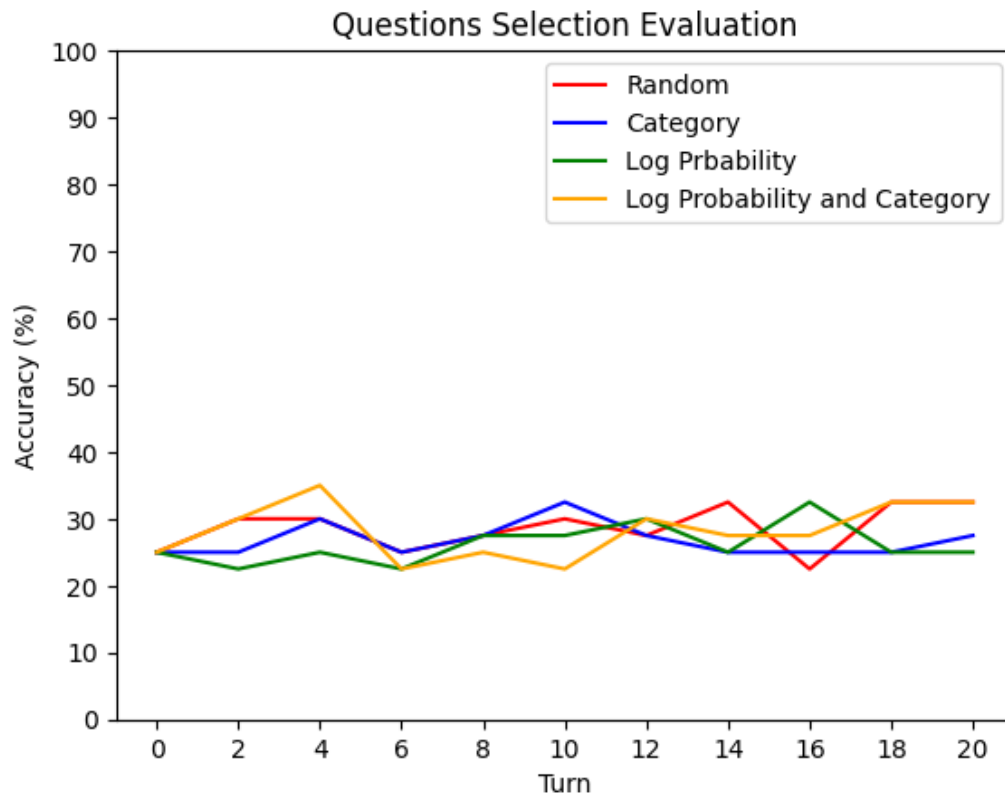
Evaluating the performance of our model involves a systematic process to gauge its effectiveness in solving the task at hand. In our case, where we are training a student model, assessing its proficiency is pivotal to ensure its efficacy in answering multiple-choice math questions. To accomplish this, we employ a line graph visualization method, which provides a clear and intuitive representation of the model's performance across different evaluation iterations.

The line graph serves as a dynamic visual tool, illustrating the model's accuracy over time as it undergoes training and refinement. It features an x-axis denoting the turn of our evaluation loop, effectively tracking the progression of the training process. Meanwhile, the y-axis quantifies the accuracy of our model after each evaluation turn, providing a measurable indication of its performance.

This visual representation allows us to observe trends, patterns, and fluctuations in the model's accuracy throughout the training process. By analyzing the trajectory of the line graph, we can identify periods of improvement, plateaus, or potential areas for further optimization. Additionally, it enables us to compare the performance of different methods or iterations, facilitating informed decision-making and model iteration.

In essence, the line graph encapsulates the evolution of our student model's proficiency, offering valuable insights into its learning dynamics and effectiveness. Through this comprehensive evaluation approach, we can iteratively refine and enhance the model's capabilities, ultimately striving for optimal performance in answering multiple-choice math questions.

5.2 Final result



Picture 13
Question Selection Evaluation

The visual representation showcases a line graph featuring four categories used for retraining the model: random (red line), category-based (blue line), log probability-based (green line), and a combination of log probability and category (yellow line). The x-axis denotes the number of training iterations, while the y-axis represents the accuracy achieved by the models. Throughout the graph, the accuracy values for all four categories fluctuate, ranging between approximately 20% and 38%. These fluctuations highlight the dynamic nature of model training and underscore the varying degrees of effectiveness across different retraining approaches.

5.3 Conclusion

In a broader context, the method employed for selecting questions not only influences the learning performance of our model but also has far-reaching implications for its applicability in real-world scenarios, particularly in assisting educators in devising effective teaching strategies. By carefully curating the questions presented to the model, we not only shape its learning trajectory but also empower it to serve as a valuable tool for educators in various educational settings.

One significant real-world application lies in the realm of personalized learning. By tailoring the selection of questions to cater to individual student needs, preferences, and learning styles, our model can facilitate personalized learning experiences. This customization enables educators to address the diverse needs of students, providing targeted support and guidance to optimize their learning outcomes.

Moreover, the method of question selection can greatly impact the efficacy of formative assessment practices. Formative assessment, which involves evaluating student understanding and progress throughout the learning process, is essential for identifying areas of strength and areas needing improvement. By selecting questions that align with specific learning objectives and competencies, our model can assist educators in conducting formative assessments more effectively, enabling them to make data-driven decisions to guide instruction and support student learning.

Furthermore, our model's ability to simulate student interaction with a diverse range of questions can aid educators in designing differentiated instruction strategies. By selecting questions that span various difficulty levels, cognitive domains, and instructional objectives, educators can differentiate instruction to meet the diverse needs of students within the classroom. This approach promotes inclusive learning environments where all students are challenged and supported at their individual levels.

Additionally, our model's role in selecting questions can extend to supporting the development of inquiry-based learning experiences. By posing open-ended and exploratory questions, educators can foster critical thinking, problem-solving skills, and deep conceptual understanding among students. Our model can assist educators in crafting inquiry-based learning activities that promote active engagement and inquiry-driven exploration, empowering students to construct their own knowledge and meaning.

Overall, the method of selecting questions holds significant implications for the learning performance of our model and its real-world applications in educational contexts. By leveraging our model's capabilities to curate questions effectively, educators can enhance teaching strategies, personalize learning experiences, conduct formative assessments, differentiate instruction, and promote inquiry-based learning. This collaborative approach between educators and AI-powered models has the potential to transform teaching and learning practices, ultimately empowering students to achieve academic success and reach their full potential.

6. References

1. "google/gemma-2b-it" - Hugging Face. Available at: <https://huggingface.co/google/gemma-2b-it>
2. "google/flan-t5-large" - Hugging Face. Available at: <https://huggingface.co/google/flan-t5-large>
3. "math_qa" - Datasets at Hugging Face. Available at: https://huggingface.co/datasets/math_qa
4. Hugging Face. "Fine-Tuning Gemma Models in Hugging Face." Available at: <https://huggingface.co/blog/gemma-peft>
5. Hugging Face. "Fine-tune Llama 2 with DPO." Available at: <https://huggingface.co/blog/dpo-trl>
6. Hugging Face. "DPO Trainer." Retrieved from https://huggingface.co/docs/trl/main/en/dpo_trainer
7. Maxime Labonne. "Fine-tune a Mistral-7b model with Direct Preference Optimization." Towards Data Science. Available at:



<https://towardsdatascience.com/fine-tune-a-mistral-7b-model-with-direct-preference-optimization-708042745aac>