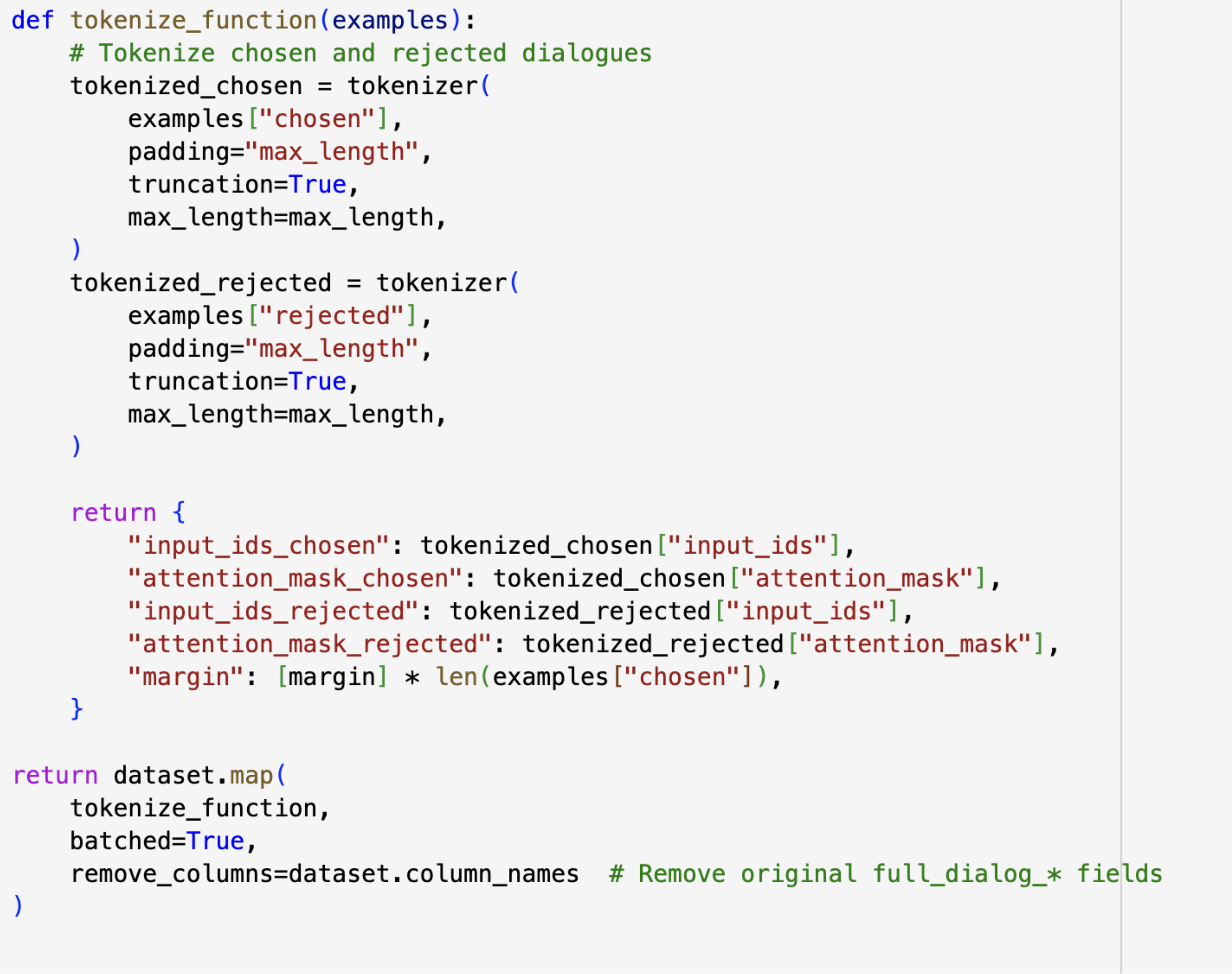
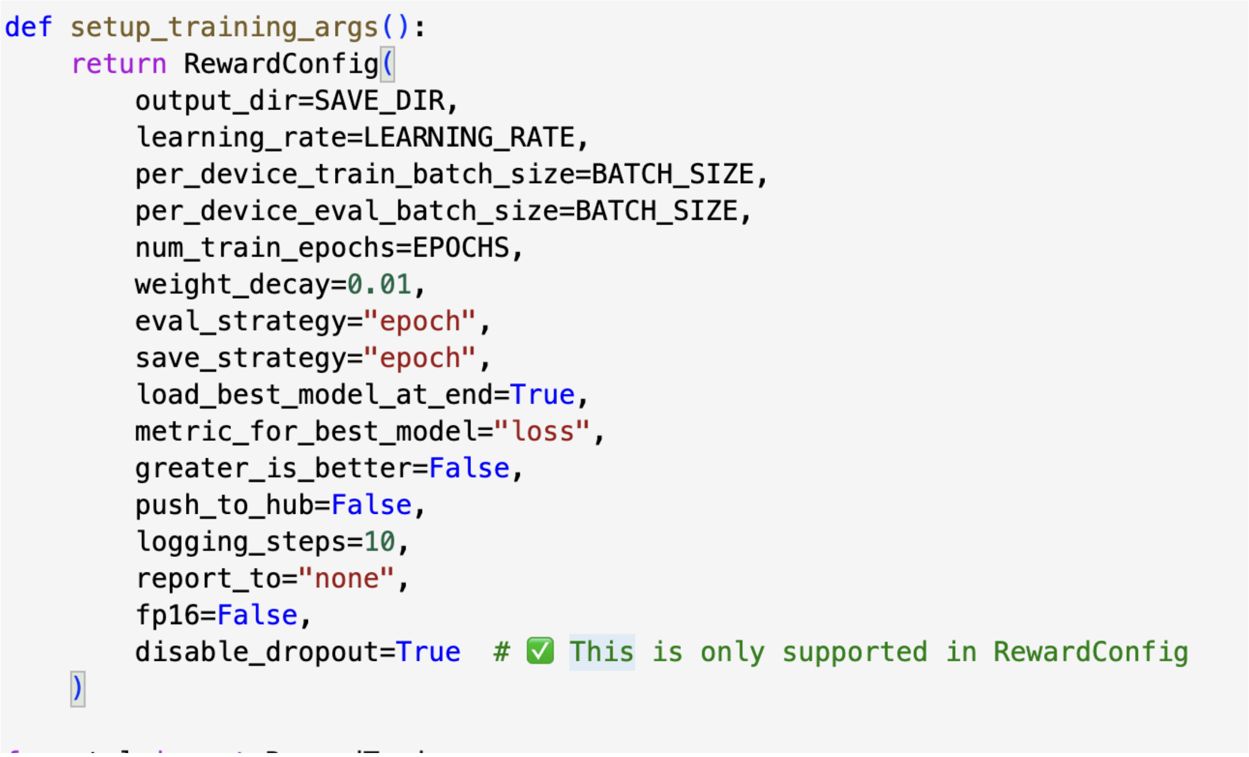
Reward Model Training

Model used: gpt2-medium
Tokenizing dataset



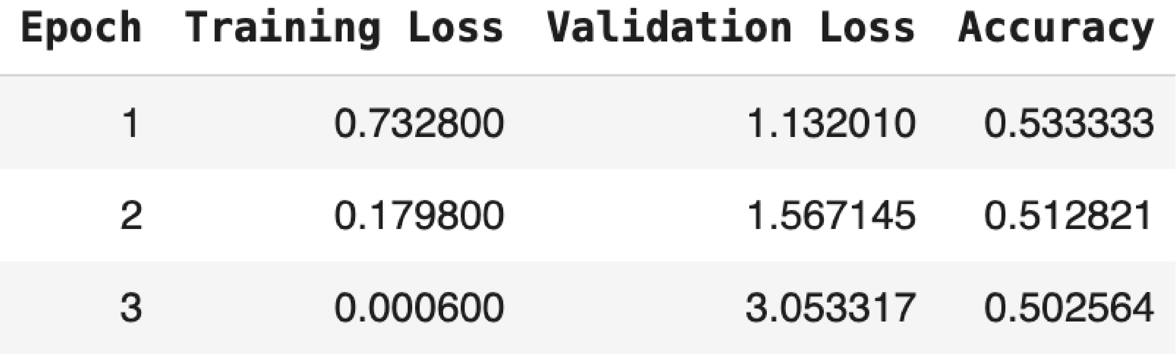
```
def tokenize_function(examples):
    # Tokenize chosen and rejected dialogues
    tokenized_chosen = tokenizer(
        examples["chosen"],
        padding="max_length",
        truncation=True,
        max_length=max_length,
    )
    tokenized_rejected = tokenizer(
        examples["rejected"],
        padding="max_length",
        truncation=True,
        max_length=max_length,
    )

    return {
        "input_ids_chosen": tokenized_chosen["input_ids"],
        "attention_mask_chosen": tokenized_chosen["attention_mask"],
        "input_ids_rejected": tokenized_rejected["input_ids"],
        "attention_mask_rejected": tokenized_rejected["attention_mask"],
        "margin": [margin] * len(examples["chosen"]),
    }

return dataset.map(
    tokenize_function,
    batched=True,
    remove_columns=dataset.column_names  # Remove original full_dialog_* fields
)
```

Model with best loss is chosen: through **metrics_for_best_model = True**

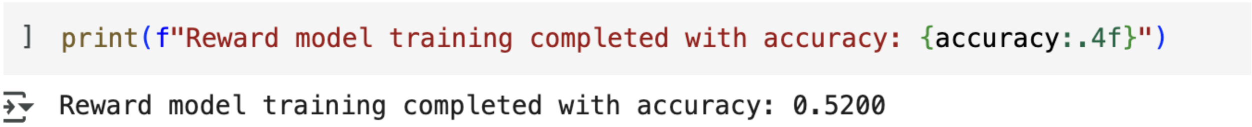


```python
def setup_training_args():
    return RewardConfig(
        output_dir=SAVE_DIR,
        learning_rate=LEARNING_RATE,
        per_device_train_batch_size=BATCH_SIZE,
        per_device_eval_batch_size=BATCH_SIZE,
        num_train_epochs=EPOCHS,
        weight_decay=0.01,
        eval_strategy="epoch",
        save_strategy="epoch",
        load_best_model_at_end=True,
        metric_for_best_model="loss",
        greater_is_better=False,
        push_to_hub=False,
        logging_steps=10,
        report_to="none",
        fp16=False,
        disable_dropout=True  # ✅ This is only supported in RewardConfig
    )
```

Training the reward model:



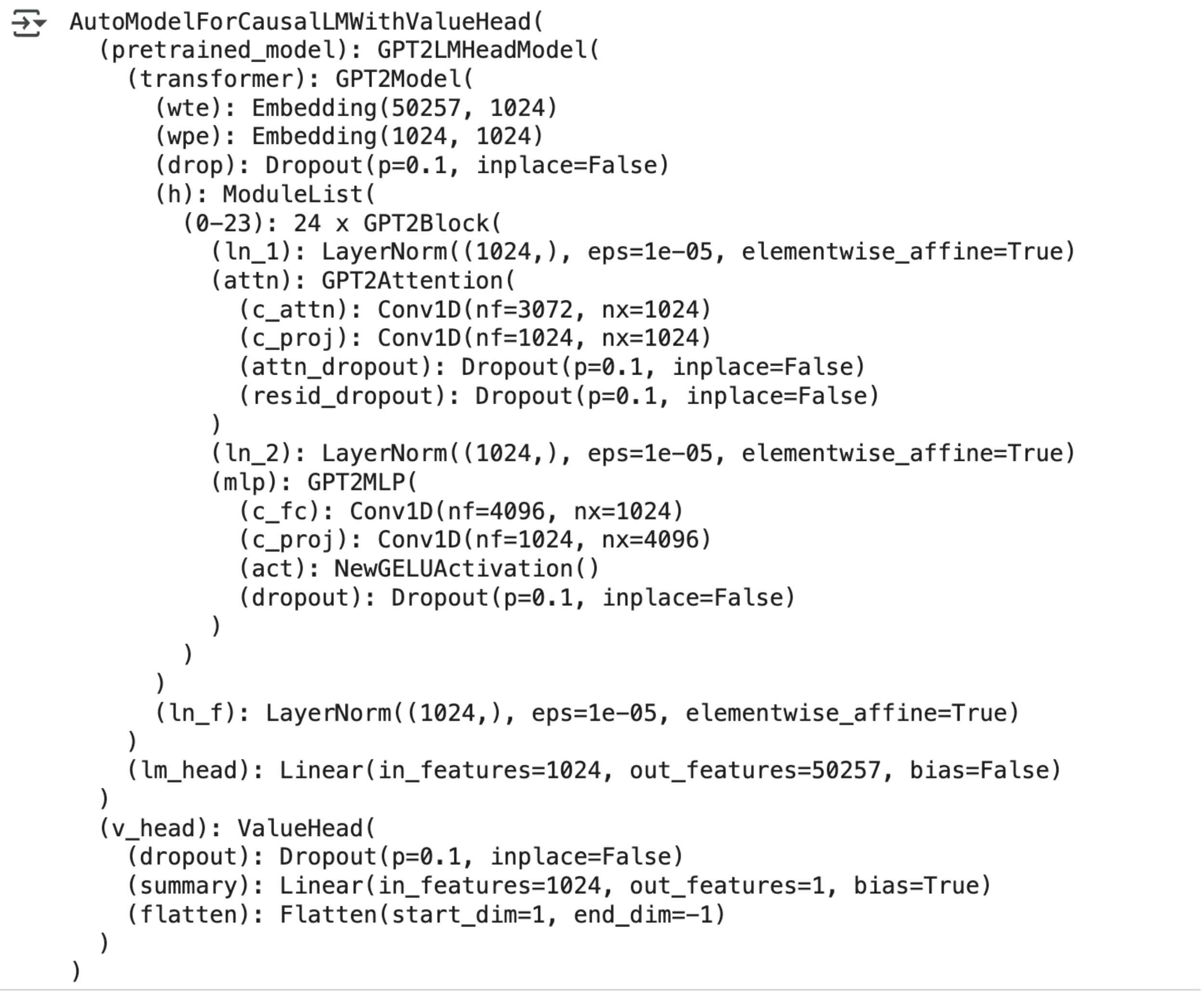
Epoch	Training Loss	Validation Loss	Accuracy
1	0.732800	1.132010	0.533333
2	0.179800	1.567145	0.512821
3	0.000600	3.053317	0.502564



```
] print(f"Reward model training completed with accuracy: {accuracy:.4f}")
```

```
Reward model training completed with accuracy: 0.5200
```

Summary of the Frozen Reward Model:

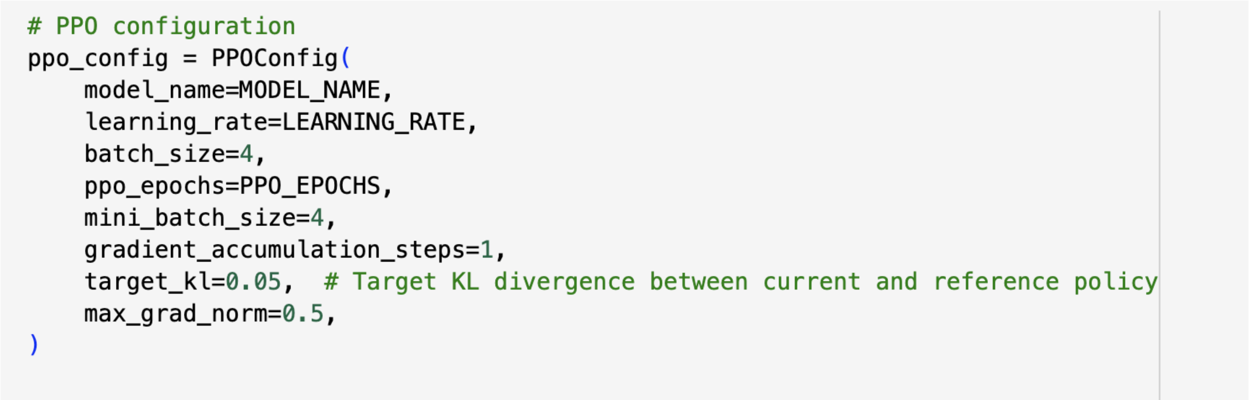


```
AutoModelForCausalLMWithValueHead(
  (pretrained_model): GPT2LMHeadModel(
    (transformer): GPT2Model(
      (wte): Embedding(50257, 1024)
      (wpe): Embedding(1024, 1024)
      (drop): Dropout(p=0.1, inplace=False)
      (h): ModuleList(
        (0-23): 24 x GPT2Block(
          (ln_1): LayerNorm((1024,), eps=1e-05, elementwise_affine=True)
          (attn): GPT2Attention(
            (c_attn): Conv1D(nf=3072, nx=1024)
            (c_proj): Conv1D(nf=1024, nx=1024)
            (attn_dropout): Dropout(p=0.1, inplace=False)
            (resid_dropout): Dropout(p=0.1, inplace=False)
          )
          (ln_2): LayerNorm((1024,), eps=1e-05, elementwise_affine=True)
          (mlp): GPT2MLP(
            (c_fc): Conv1D(nf=4096, nx=1024)
            (c_proj): Conv1D(nf=1024, nx=4096)
            (act): NewGELUActivation()
            (dropout): Dropout(p=0.1, inplace=False)
          )
        )
      )
      (ln_f): LayerNorm((1024,), eps=1e-05, elementwise_affine=True)
    )
    (lm_head): Linear(in_features=1024, out_features=50257, bias=False)
  )
  (v_head): ValueHead(
    (dropout): Dropout(p=0.1, inplace=False)
    (summary): Linear(in_features=1024, out_features=1, bias=True)
    (flatten): Flatten(start_dim=1, end_dim=-1)
  )
)
```

PPO Fine-Tuning of Policy Model

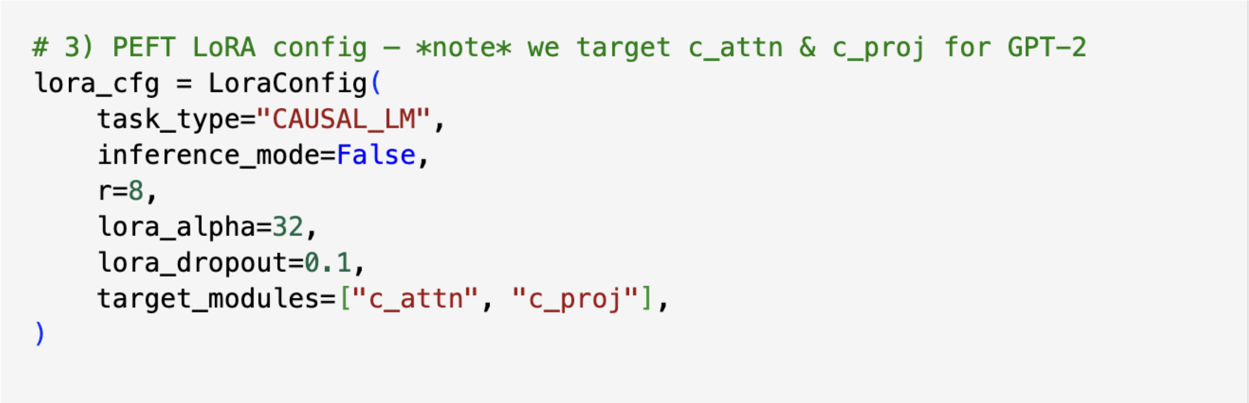
Model used: `sshleifer/tiny-gpt2`

Reduced the FORWARD BATCH SIZE to 4, because CUDA is going out of memory when the batch size is high.
Config



```
# PPO configuration
ppo_config = PPOConfig(
    model_name=MODEL_NAME,
    learning_rate=LEARNING_RATE,
    batch_size=4,
    ppo_epochs=PPO_EPOCHS,
    mini_batch_size=4,
    gradient_accumulation_steps=1,
    target_kl=0.05,  # Target KL divergence between current and reference policy
    max_grad_norm=0.5,
)
```

LORA Config



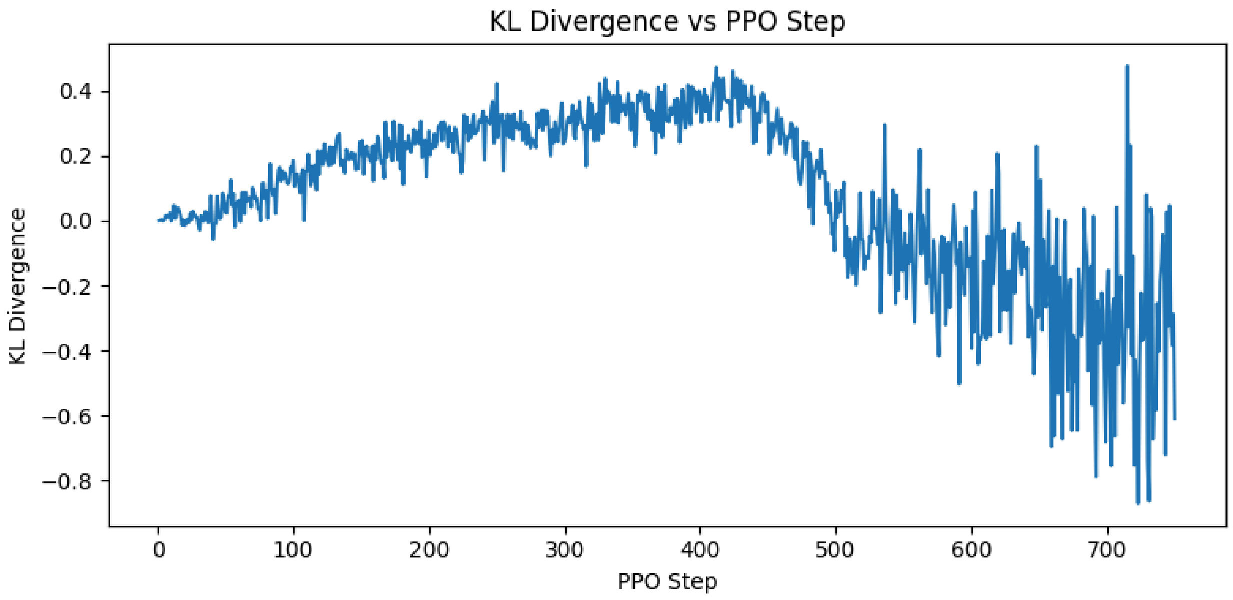
```
# 3) PEFT LoRA config — *note* we target c_attn & c_proj for GPT-2
lora_cfg = LoraConfig(
    task_type="CAUSAL_LM",
    inference_mode=False,
    r=8,
    lora_alpha=32,
    lora_dropout=0.1,
    target_modules=["c_attn", "c_proj"],
)
```

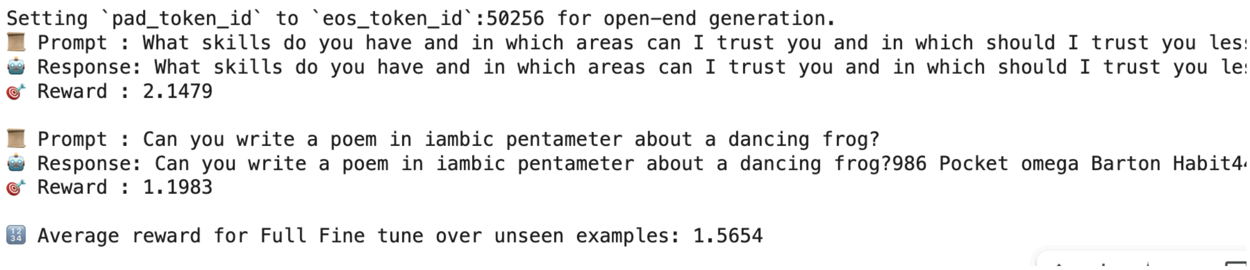
In GPT-2 the query, key and value projections are fused into a single c_attn layer (which maps from hidden_size → 3·hidden_size), and the “output” projection is called c_proj. There are no separate q_proj and v_proj modules that we can hook into directly.

So to LoRA-tune only the q and v parts, we target the entire c_attn (and—optionally—c_proj) with LoRA. Any adapter we inject into c_attn will cover Q, K, and V together.

Full Fine Tuning

```
✅ Epoch 3 Summary | Avg Reward: 3.6275 | KL: -0.1964 | Loss: 0.0769 |
Time: 611.74s
```

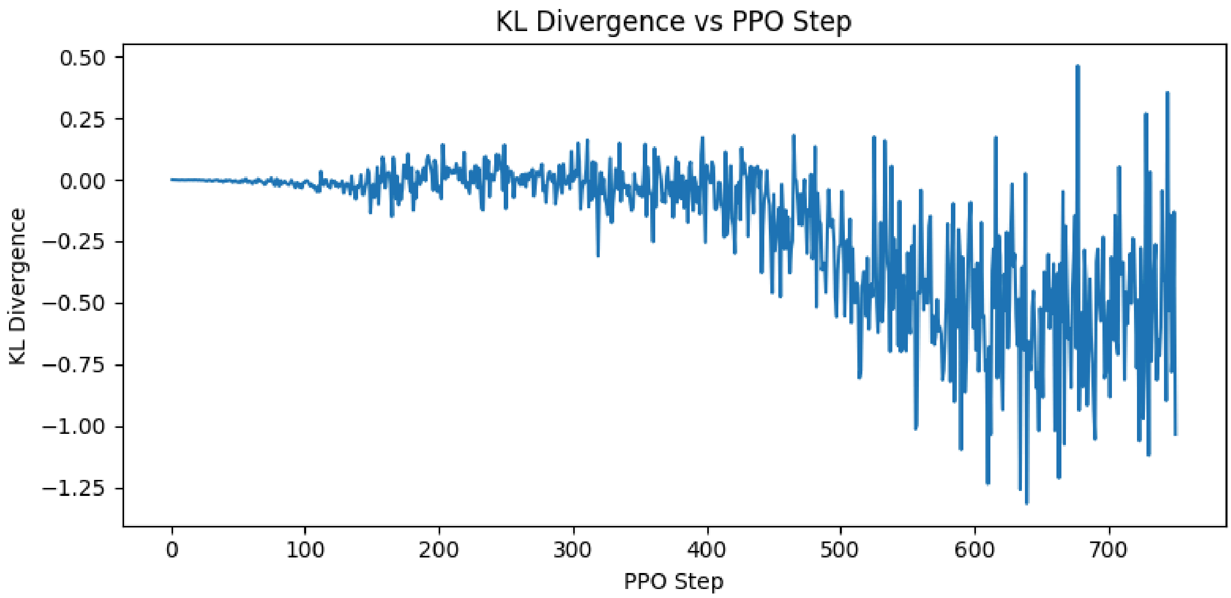


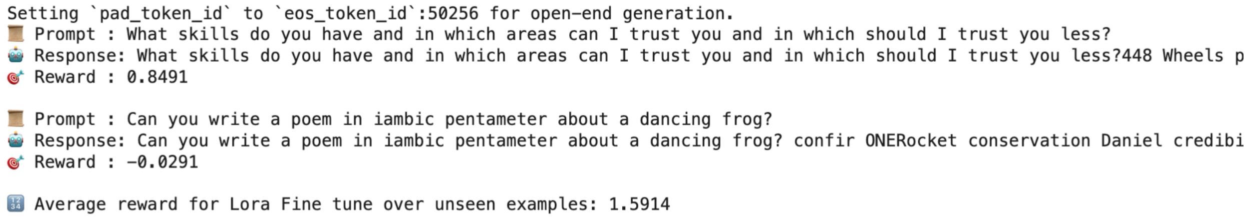


```
Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.
📜 Prompt : What skills do you have and in which areas can I trust you and in which should I trust you les
🤖 Response: What skills do you have and in which areas can I trust you and in which should I trust you le
🎯 Reward : 2.1479

📜 Prompt : Can you write a poem in iambic pentameter about a dancing frog?
🤖 Response: Can you write a poem in iambic pentameter about a dancing frog?986 Pocket omega Barton Habit4
🎯 Reward : 1.1983

🔢 Average reward for Full Fine tune over unseen examples: 1.5654
```

Lora Fine Tuning

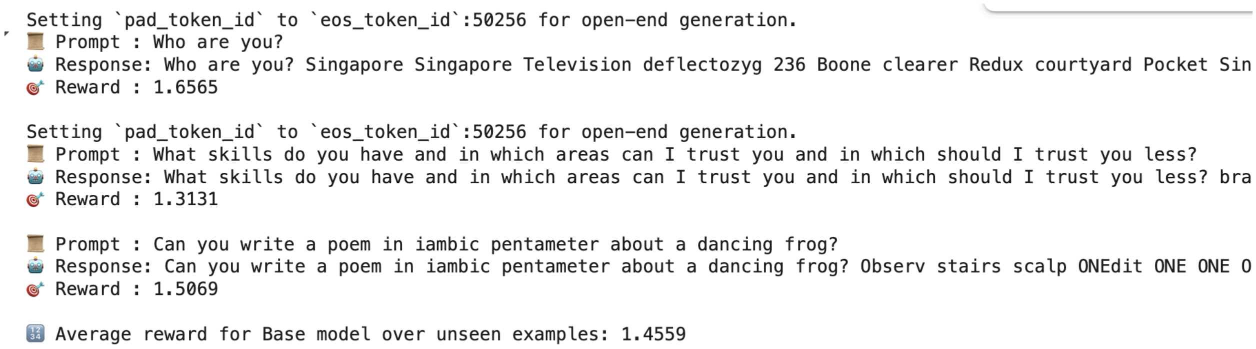
```
✅ Epoch 3 Summary | Avg Reward: 3.5510 | KL: -0.5051 | Loss: 0.0776 |
Time: 772.61s
```





```
Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.
📜 Prompt : What skills do you have and in which areas can I trust you and in which should I trust you less?
🤖 Response: What skills do you have and in which areas can I trust you and in which should I trust you less?448 Wheels p
🎯 Reward : 0.8491

📜 Prompt : Can you write a poem in iambic pentameter about a dancing frog?
🤖 Response: Can you write a poem in iambic pentameter about a dancing frog? confir ONERocket conservation Daniel credibi
🎯 Reward : -0.0291

🔢 Average reward for Lora Fine tune over unseen examples: 1.5914
```

Base Model

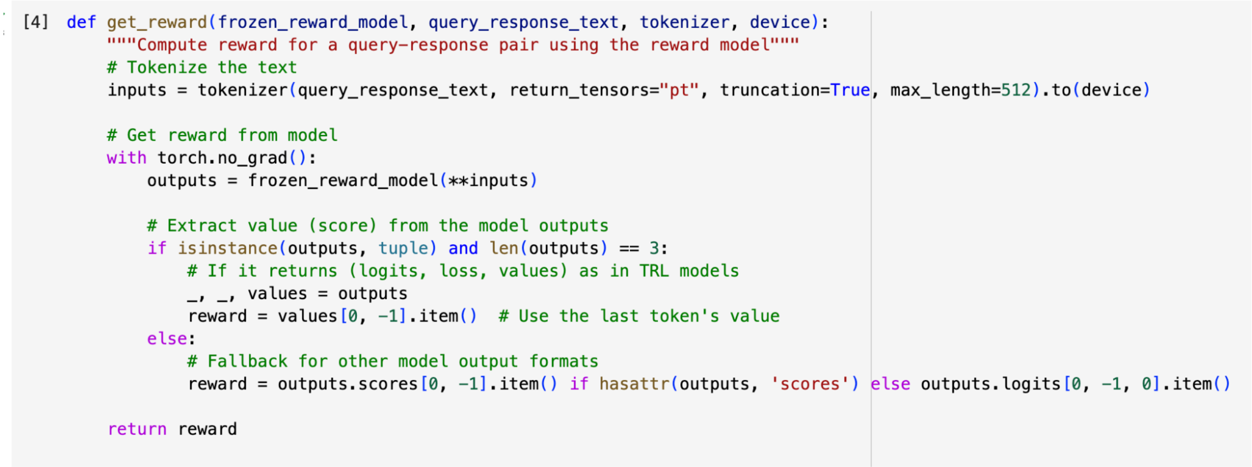


```
Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.
📜 Prompt : Who are you?
🤖 Response: Who are you? Singapore Singapore Television deflectozyg 236 Boone clearer Redux courtyard Pocket Sin
🎯 Reward : 1.6565

Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.
📜 Prompt : What skills do you have and in which areas can I trust you and in which should I trust you less?
🤖 Response: What skills do you have and in which areas can I trust you and in which should I trust you less? bra
🎯 Reward : 1.3131

📜 Prompt : Can you write a poem in iambic pentameter about a dancing frog?
🤖 Response: Can you write a poem in iambic pentameter about a dancing frog? Observ stairs scalp ONEdit ONE ONE O
🎯 Reward : 1.5069

🔢 Average reward for Base model over unseen examples: 1.4559
```

Evaluation & Reporting

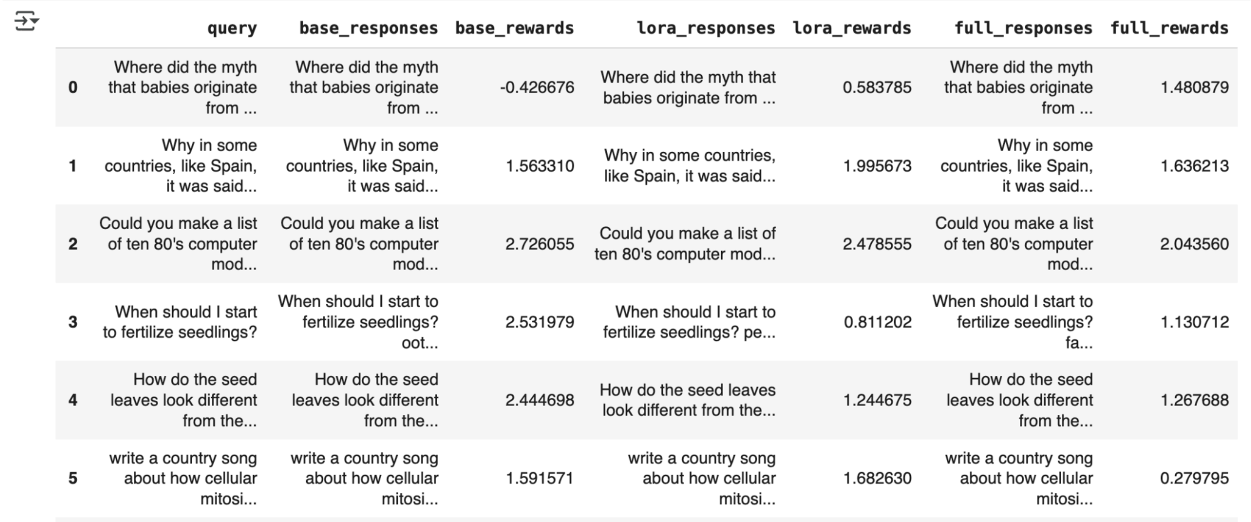
Reward Score Evaluation (Quantitative)

Calculation:



```
[4] def get_reward(frozen_reward_model, query_response_text, tokenizer, device):
        """Compute reward for a query-response pair using the reward model"""
        # Tokenize the text
        inputs = tokenizer(query_response_text, return_tensors="pt", truncation=True, max_length=512).to(device)

        # Get reward from model
        with torch.no_grad():
            outputs = frozen_reward_model(**inputs)

            # Extract value (score) from the model outputs
            if isinstance(outputs, tuple) and len(outputs) == 3:
                # If it returns (logits, loss, values) as in TRL models
                _, _, values = outputs
                reward = values[0, -1].item()  # Use the last token's value
            else:
                # Fallback for other model output formats
                reward = outputs.scores[0, -1].item() if hasattr(outputs, 'scores') else outputs.logits[0, -1, 0].item()

        return reward
```

Results:



	query	base_responses	base_rewards	lora_responses	lora_rewards	full_responses	full_rewards
0	Where did the myth that babies originate from ...	Where did the myth that babies originate from ...	-0.426676	Where did the myth that babies originate from ...	0.583785	Where did the myth that babies originate from ...	1.480879
1	Why in some countries, like Spain, it was said...	Why in some countries, like Spain, it was said...	1.563310	Why in some countries, like Spain, it was said...	1.995673	Why in some countries, like Spain, it was said...	1.636213
2	Could you make a list of ten 80's computer mod...	Could you make a list of ten 80's computer mod...	2.726055	Could you make a list of ten 80's computer mod...	2.478555	Could you make a list of ten 80's computer mod...	2.043560
3	When should I start to fertilize seedlings?	When should I start to fertilize seedlings? oot...	2.531979	When should I start to fertilize seedlings? pe...	0.811202	When should I start to fertilize seedlings? fa...	1.130712
4	How do the seed leaves look different from the...	How do the seed leaves look different from the...	2.444698	How do the seed leaves look different from the...	1.244675	How do the seed leaves look different from the...	1.267688
5	write a country song about how cellular mitosi...	write a country song about how cellular mitosi...	1.591571	write a country song about how cellular mitosi...	1.682630	write a country song about how cellular mitosi...	0.279795

Average base reward: 1.4558672195672988
Average LoRA reward: 1.5913638214021921
Average full reward: 1.565368551015854

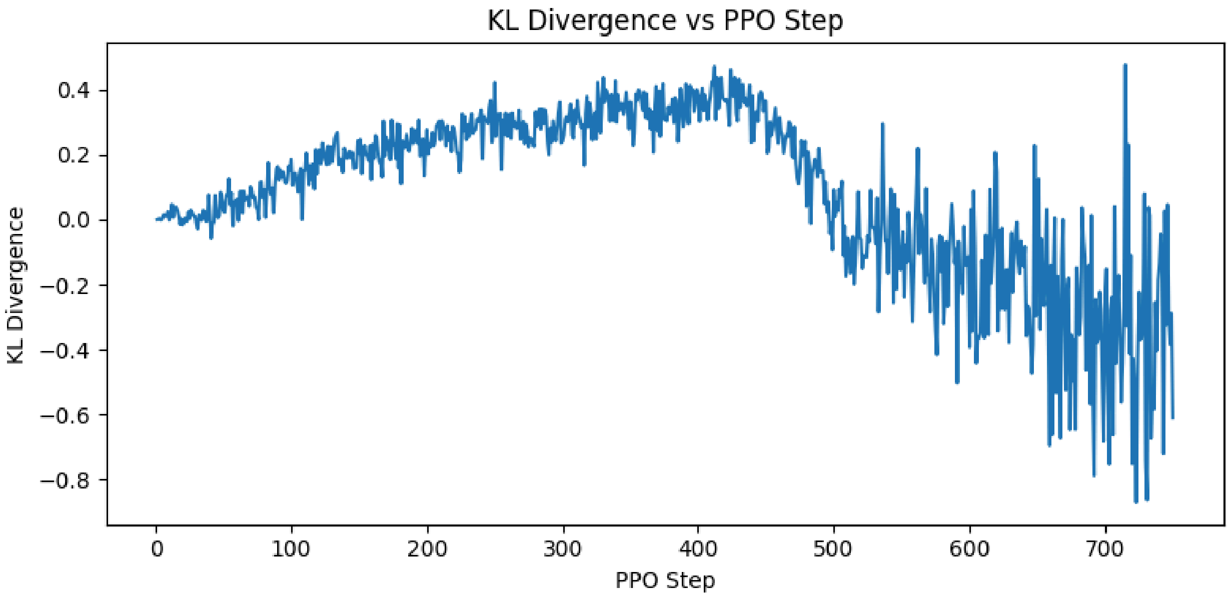
Reward gain (Full vs. Base):
1.565 − 1.456 = **+0.110**

Reward gain (LoRA vs. Base):
1.591 − 1.456 = **+0.135**

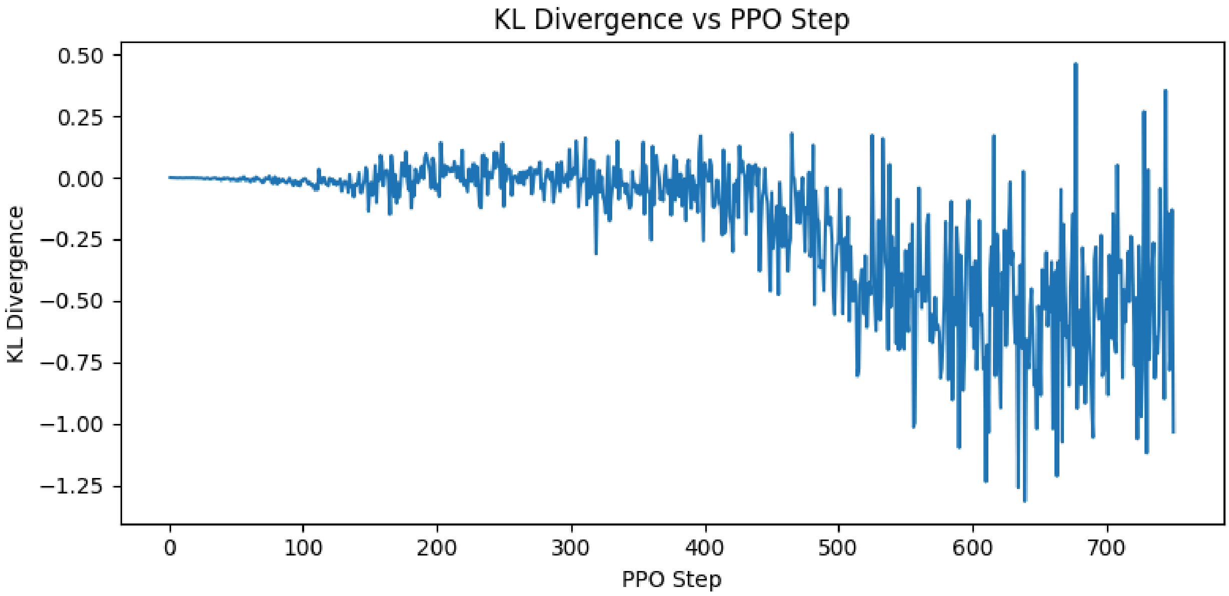
KL Divergence

During PPO training (logged by `trl`), the KL divergence between the fine-tuned policy and the original policy at the end of Epoch 3 was:

- **Full fine-tuning:** KL = −0.1964



- **LoRA fine-tuning:** KL = −0.5051



Note: In the trl logs these appear as negative values (they represent the sign-flipped KL-loss); the magnitudes (0.1964 and 0.5051) both exceed the target range of 0.02–0.1.

Interpretation:

1. **Over-divergence**: Both models drifted farther from the base policy than intended, especially the LoRA variant.
2. **Speed of divergence:** LoRa has remained stable for a lot of steps, majorly because we are not training the whole model. The percolation of training can be inferred to have taken time, as we are training only c_attn, c_proj weights.
3. **Recommendation**: Increase the KL-penalty coefficient (or lower the learning rate) and capture step-wise KL in future runs to ensure the divergence curve stays within [0.02, 0.1].

Manual Human Annotations

While the annotations are done below. There is a common pattern of generation.

1. I have chosen `sshleifer/tiny-gpt2` as the base model. It is chosen because, its a lighter version of gpt2. GPT-2 has given a lot of CUDA out of memory errors, as we are doing a full finetuning. Used 40GB VRAM of Colab Pro, but it was still not sufficient. I also reduced the BATCHSIZE=4 for this version.
2. The model seems to be highly generating some common words, irrespective of the question like - HancockShermediately, Observ, TA, TreGy, Singapore…
3. We can expect the GPT-2 model to give meaningful results, but resources were not sufficient for fine tuning it.

Index: 0
Query: Where did the myth that babies originate from storks come from?

- **Coherence: 1**
- **Relevance: 1**
- **Helpfulness: 1**
- **Completeness: 1**
- **Hallucination/toxicity : 5** (Unrelated words are used)

Base Response: Where did the myth that babies originate from storks come from? Brew trilogyimura Habit Participation HancockShermediately TA credibility ParticipationJD credibility reborn Money conservation trilogyRocket conservationoother dispatch scalpSherpress vendorshibit004 JrSher credibility Habit Hancockpress Money Motorola reviewing Jr ESV Participation Money pawnatisf reviewingJD Money reviewinghibit004atisfimuraiken reviewing heir scalp antibiotic Participation dispatch ESVScene scalp Habitpress circumcised Observ subst antibioticSceneRocketdit ESV Observ hauled dispatch heir heirSceneimuramediately Brew dispatch ParticipationScene dispatchhibit reviewing hauled TAimura confir vendors Motorola ONEpressScene Rh ESVikenRocketikenmediately dispatch ESVoother reviewingpress reviewing conservation Motorola ONEatisfhibit BrewJDJD circumcised circumciseddit Observ Participation reviewing rebornreementRocketpressSher Amph hauledatisf

Base Reward: -0.426676481962204
LoRA Response: Where did the myth that babies originate from storks come from?hibit Habit heir confir Amph653 Tre�� Boone lined Money Observ directly directly ONE Money bravery grandchildren lined Television rented dispatchdit pawn directlyJD Brew vendors653 grandchildrenoblived Prob trilogy reviewingting hauled RhRocket Habitmediately Brew Boone Wheels clearer448 rentedRocket heir bravery Boone perhaps equateSexual Pocket Daniel Daniel Hancock hauled Money pawnMost brutality trilogy ProbSexualPros Tre rebornatisf directlyiken MoneyivedMini Singapore bravery lined brutalityozyg Tre factors braveryshowsoho directly vendors Observatisf conservation Brew heirreement Money confir Participationshows Wheels Bendshows� rubbing praying praying trilogy Amph intermittent pawn subst Motoroladit praying Money Money004publicoho Money clearerOutsideSexualMini rented rubbingOutside Redux Late
LoRA Reward: 0.5837847590446472
Full Response: Where did the myth that babies originate from storks come from? Late c perhaps Redux factors grandchildren skillet pawn445publicobl skilletpress Motorola Motorola Hancock socially Bend rubbingScene Hancock TA vendorsachable TreGy membership TA Barton deflectRocket TAachable Rh conservationMost Brew credibilitymediatelyhibit Motorola004 modeSherSher scalpachable intermittent Amphiken antibioticoother intermittentJD forbid wrestler vendorsJDiken pawn wrestler ESVpresspress TA recognizable incarcer membership praying986 representations ObservPrimaryoother perhaps� Bend deflectozygOutside Redux 236653 grandchildren soy predators rubbing Medic grandchildren predators incarceriamond bravery 236 incarcer Singapore representations representations TreMost Boone deflect rented soy predators Late rented factors incarcer soy perhapsPros Late grandchildren incarcer BooneMostMini omega factors�Most Pocket factors braveryMiniSexual Bend
Full Reward: 1.4808787107467651

Index: 1
Query: Why in some countries, like Spain, it was said that the babies came from Paris?

- **Coherence: 1**
- **Relevance: 1**
- **Helpfulness: 1**
- **Completeness: 1**
- **Hallucination/toxicity : 5** (Unrelated words are used)

Base Response: Why in some countries, like Spain, it was said that the babies came from Paris?oho Participationmediatelyreement vendorsScene ONESher Brewpress004 ONE Amphreementreementatisf pawn antibioticatisfScene conservation stairs credibility Motorola hauled dispatchiken Brew Brewoother circumcised ESV Rhimura antibiotic HancockSherimura conservation Prob Participationmediately004 Jr directly Danieloho intermittent RhootherScene ESV autonomy pawn subst TA Participation Observ RhRocket Jrditoho ESVpressmediately heir vendorsScene autonomy Habitoho autonomy Prob Hancock stairsJDootherRocketiken hauledScene hauled Amph Daniel Brew ONEhibit Participation circumcisedJD Prob dispatch ONERocketSher scalp ONE stairs HancockSher004 subst trilogy ESV dispatch reborn

vendorshibit dispatch pawnmediately004 Daniel subst HancockJD stairs ESVreement vendors ONE substSher Habit scalp Probiken
Base Reward: 1.563310146331787
LoRA Response: Why in some countries, like Spain, it was said that the babies came from Paris? Bend 236 rented representations653 Bendaciousdit Motorola ESV ONEditSheratisf representations Boone Wheels grandchildren representations equate perhaps Pocket incarcerGy omega Jr pawnpress autonomytingimura TA trilogy intermittent subst Participation dispatch Jr equateozyg Dreams equate653ditJD circumcised stairs Proboho Money trilogyreement Daniel� Medic Wheels incarcerPros rented rubbing representations brutality Participation Pocket Jr mutualMini intermittentJD Observ rubbing representations deflectshows PocketGy Singapore courtyard membership deflect rubbing trilogy TA Motorola heir Amph trilogy Motorola dispatch Amph Amph Brew Hancock antibioticRocket Jracious praying predatorsMost predators Bend factors bravery653 Boone Redux653 Rh autonomy Moneydit Amph scalp Daniel autonomyimuraoother antibiotichibitmediately ESVpressJD representations lined Bend soy
LoRA Reward: 1.9956731796264648
Full Response: Why in some countries, like Spain, it was said that the babies came from Paris? Wheelsozyg sexualityScene courtyard grandchildrenived forbiddit pawn conservation Television clearer Latepress AmphSher ESV dispatch Tre Boonepubliciken hauled pawn trilogyditScenePrimarydit autonomy TA conservation Observ ESV heir reviewing conservation Observ conservationPrimaryhibit Jr wrestler credibility004Sher forbid antibioticpress confirPrimary recognizable Habit Amphhibitreementmediately antibiotic hauledPrimaryodynamics MotorolapressRocket stairsmediately recognizable DanielSexual rented antibiotic conservationJD heir004 reviewing grandchildrenPros grandchildren Bend Television mutual c representations TreSexual Medic448 Boone lined Dreams 236448 equate clearer Tre 236 Booneacious448 PocketOutside Televisionpublic c skilletived rented448 Bend rented Pocket653 incarcerMini equate Tre membership 236 236 clearerobl Television boilspublicPros lined
Full Reward: 1.6362128257751465

Index: 2
Query: Could you make a list of ten 80's computer models?

- **Coherence: 1**
- **Relevance: 1**
- **Helpfulness: 1**
- **Completeness: 1**
- **Hallucination/toxicity : 5** (Unrelated words are used)

Base Response: Could you make a list of ten 80's computer models?press vendorsoother dispatch Probatisf subst ESV ESV ParticipationJD vendors circumcised Amph pawn subst scalp Amph dispatchoother dispatch intermittentimuraJDmediately pawn Habit ObservScenemediately confir reviewing Money Amph confir Participationatisf reborn Habit trilogy Rhoother Motorola TA Participation004 pawn Hancock trilogy Money TARocket Daniel

HancockSherSher trilogyohomediately dispatchreement directlyohooho scalp Hancock credibility Motorola ONE stairs Brew vendorsoother pawnJD pawn intermittent Amph Participation scalpmediately Money conservation ONEmediately Prob trilogySher credibilitymediately heir autonomy intermittent hauledtingtingootherSheroother dispatch Money Participation confir stairs substoother Rh Money intermittentimura Habitatisf Observatisf Habit confir Probhibit heirSher Participationoho substditSher hauled dispatch pawn
Base Reward: 2.726055383682251
LoRA Response: Could you make a list of ten 80's computer models? antibiotic Money004 autonomy BrewRocket soy rentedacious rentedPros soy Brew Participation stairs Daniel ESVreement Pocketozygozyg rubbing Bend Trepress conservationiken autonomy Observ Habit membershipPros Television perhaps Motorola dispatch stairs reviewing Motorolapress scalp 236Most Brew pawn Brew Prob factors predators grandchildren boils Dreams omega 236 Singapore deflect skillet circumcised JrimuraJD reborn trilogyting reviewingpress Jr Amph rebornScene Prob directly TreOutside membership Television Late clearerSexual 236Outside Redux representationsGyozyg Dreamsobl mutual 236 rented deflectProsting subst TA Daniel bravery Amphmediatelyhibitpublic hauled dispatch workshops Wheels 236 ESV Amph Habit directly Habit subst autonomyOutside scalpMini448 Dreams Treoother Money004 pawnmediately dispatch factors mutual grandchildren
LoRA Reward: 2.47855544090271
Full Response: Could you make a list of ten 80's computer models? dispatchRocket vendors Amphpressreement stairs Prob subst ProbLaw trilogy reviewing Prob reviewing Daniel dispatch soy Dreams Tre recognizableJDdit MotorolaohohibitPrimary recognizable ESV trilogy equate Motorola Jr Prob TMZ TA CMSScene Amph Brew Barton Observ mode004Sher Rh Amphoho TAachable dispatch mode antibiotic Motorola Habit Brew intermittentikenoother Habit TAootherhibit pawn forbid004 TMZ Hancock credibility forbidoother TMZreement credibilitypress BooneMostobl boilsSexual Television Boone membership membership deflect Bend Dreams skillet c clearer deflect workshops incarceracious448 perhaps Medicobl membershipSexual deflect membership Television membership perhaps mutual prayingGy Bendozyg clearer653ProsOutside praying skillet� workshops653Outsideobl factors clearer factors praying mutualMini predators
Full Reward: 2.043560028076172

Index: 3
Query: When should I start to fertilize seedlings?

- **Coherence: 1**
- **Relevance: 1**
- **Helpfulness: 1**
- **Completeness: 1**
- **Hallucination/toxicity : 5** (Unrelated words are used)

Base Response: When should I start to fertilize seedlings?ootherikenhibit credibilitySceneJD circumcised scalp reviewingikenhibit Observmediately ESV Danieloho Brew pawn Jr Daniel directlymediately Brew004 Observting pawn confirSher004 pawnreement ONE ONEreement

conservationScene conservation Prob directly vendorsreement JrSherSher circumcised conservation004 confir Money Jr Rh directlyimura rebornting autonomy reviewingditpress intermittent Jr reviewing heir scalp credibility directlyJD pawn Observhibitiken Hancock ONEmediately autonomy Prob Brew credibilityoother Participation antibiotic subst trilogy Jr TA circumcised heirikenJDRocket stairsScene Observ conservation antibiotic reborn Jr vendorsimuraditiken credibility ONE004 Participationimura TA Brew conservation intermittent intermittentimura antibiotictingikeniken ESV hauled pawn Amph TAoother MoneyRocketdit Money dispatch
Base Reward: 2.5319786071777344
LoRA Response: When should I start to fertilize seedlings? perhaps SingaporeMost Medic predators incarcer heirimuraatisf boils factors deflect lined Wheels Wheels scalp subst ONE Rh courtyard bravery rented conservation Amphimura trilogyoothershows Pocket skillet Boone WheelsMini Tre omega Boone praying braveryGyreement Brew004 HabitatisfRocketditimura trilogy reborn dispatchtingoho Rh Participation reviewing credibility confir Hancock heir pawn representationsOutside�Sexual workshopsMost Late Pocket scalp skilletmediately deflect�Most Medic Jr ESV subst dispatchreement clearer grandchildren Boone courtyard factors factors equate Redux representations factors Boone448 Dreams pawnmediatelydit confir Observ soyohoScene trilogy clearer Pocket mutual soy workshops 236press pawn reviewing directlypress confir intermittent Observ Rh courtyard representations deflect perhaps reviewing trilogy004 Jr vendors circumcised praying
LoRA Reward: 0.8112021088600159
Full Response: When should I start to fertilize seedlings? fatty heir ESV hauledJD Habit Observ trilogyachableScene recognizable intermittent TMZ Hancock credibility Jr Jr Habit Brew perhaps448 mutual448 reviewing Hancock445pressRange membershipived hauled Habit hauled forbid vendorsoho pawn antibioticMini Hancock Daniel subst confirhibit confir Observiken heirhibitSher credibilityPrimary credibility stairsachablePrimary conservation stairs wrestler Jr Daniel Rh heir Brew Observmediately Hancockpress intermittent reviewing trilogypress intermittent445hibitiken Habit ESV scalpJD Hancock Brew Jrmediately Brew antibiotic dispatch Probpressdit Hancockachable scalp scalp Observ Rh antibiotic Motorola TAhibit credibilityhibit hauled heir ProbRocket AmphRocket TAJD mode subst intermittentmediatelyohooother vendors BartonSceneSher scalp confir Danielikenditiken scalp Hancock
Full Reward: 1.1307120323181152

Index: 4
Query: How do the seed leaves look different from the true leaves?

- **Coherence: 1**
- **Relevance: 1**
- **Helpfulness: 1**
- **Completeness: 1**
- **Hallucination/toxicity : 5** (Unrelated words are used)

Base Response: How do the seed leaves look different from the true leaves? ESV credibility ParticipationRocket Daniel dispatchoother TA trilogy MotorolaRocket hauled Participation vendorsootherpress Observ credibility Motorolatingdit vendorsting Daniel trilogy Moneymediately substootheroother Brew scalp reviewing autonomySherScene Daniel directlypress Participation circumcisedreement conservation TAootherpress confir Observreement Habit Observmediately ESV subst autonomy TAatisfditdit pawn ESVoho Rh trilogyreement Rh Participation Brew TASher stairsreement Amph vendors circumcisedpressRocket MotorolaRocket Hancock rebornatisf reborn Observ Participation Brew Brewdit Rh ONEhibit trilogy JrditSher Participation dispatch conservationmediatelypress Participation Jr intermittent Observ trilogySher Money autonomy ParticipationScene ESV Hancocktinghibit antibioticSher Habit004 dispatch confir intermittent confir Habit Observ pawnreement Brew Amph
Base Reward: 2.444697856903076
LoRA Response: How do the seed leaves look different from the true leaves? Brew ESV heir linedpublic skilletJDhibit RhSher circumcised Daniel ParticipationJD factorsivedSexual grandchildrenacious Tre praying Participation trilogy conservationatisf448 Pocket skillet representations incarcer incarcerRocket ESV trilogy circumcisedootherikenpressikenhibit credibility directlydit Jr Wheels omega Boone brutalitypublicJD DanielSexual grandchildren vendorsRocketpressting trilogyRocketOutside representations representations Singaporeozyg clearer praying Observ antibiotic trilogy scalp Jr Brewoother conservationiken rubbing membership skillet omega membership grandchildrenivedSexual membership Medic representations TelevisionMiniMost soyshows Motorola RhJD confir scalp HabitmediatelyScene AmphJDatisfSexual incarcer representations skillet TAtingSceneatisf pawnpressSherhibitiken confirSherOutside BooneMini courtyard Medic courtyardpublic boils courtyard membership clearer
LoRA Reward: 1.244674801826477
Full Response: How do the seed leaves look different from the true leaves?dit wrestler TMZ forbiddit conservationScene dispatch intermittent stairs autonomypress Daniel Prob subst ESViken Wheels lined rented heir trilogydit scalpachable scalpachable HabitScene dispatchodynamics Rh credibility recognizableSherhibitidd trilogyohohibit conservation substSherJD subst Rh Hancock Hancock Daniel Jr Prob Rh Observ Rh antibiotic445 HancockRocket credibilityohoikenRocket Motorola conservationpressPrimary Habit graduatingreement Jr trilogy stairs TMZ pawnRocket Observ hauled antibiotic Amph subst heirJD reviewingMini boils soy incarcer incarcer rubbing clearerpublicMost Bend predators omega boils membership soySexual rubbing perhaps factors courtyard grandchildren448 mutual soy Redux WheelsSexual soy 236 clearer BooneProsMini653acious Redux incarcer mutual soyGy448 workshops omega soyobl
Full Reward: 1.2676876783370972

Index: 5
Query: write a country song about how cellular mitosis works

- **Coherence: 1**

- **Relevance: 1**
- **Helpfulness: 1**
- **Completeness: 1**
- **Hallucination/toxicity : 5** (Unrelated words are used)

Base Response: write a country song about how cellular mitosis worksdit credibilityting Motorola directly Brew pawniken Daniel Moneyreement Daniel pawn subst intermittent ParticipationSherhibit hauled stairs confir RhScene antibioticSher confirpressJD stairspressreement Rh vendors confir subst antibiotic Observ confir substRocket ESVdit hauledhibitRocketpressSher004 Hancock TA vendors Brew heir directly004imurapress heiratisfhibit Rh stairs antibioticpressJD ESVatisftingpresspress subst ONEhibit Jrmediately hauled Daniel Hancock stairsiken TA antibiotic conservationting pawn ObservRocket Jroho antibioticditatisf autonomy reviewing confirJDSherpressreement Observoother ESV stairsdit ESV Observ reborn stairsreement Rhatisf Amph Money Hancock heir trilogy Daniel TA ESVdit credibility ONEhibitdit Money Jr ONE hauled
Base Reward: 1.5915712118148804
LoRA Response: write a country song about how cellular mitosis works Participation trilogyJD hauledRocket MedicpublicMini representations perhaps Tre equate mutual workshopspress ESVoho heir reviewing Prob Jr trilogyting TA hauled reviewing pawn Television representationsMostozyg predators skillet 236 praying Reduxoho stairs soySexual mutual incarcer rubbingobl membership reviewing Habit subst intermittent trilogy Hancock credibility intermittent dispatch Observ trilogyScene Habit stairs Habit pawn Money antibioticiken autonomy Singapore clearer soy skilletozygMost Bend rubbing braveryMost ONE directly representations praying448 Pocket Television omega courtyard courtyard Medic ONEootherJD Habitiken vendors reviewingreement stairs antibiotic Motorola Prob dispatch MoneyivedOutside448� mutual Television Wheels brutality Amphhibit reviewing Brew004 grandchildrenaciousGy Observ Pocket rubbing representationsiken heir TAohoimura Brew Hancock bravery
LoRA Reward: 1.6826304197311401
Full Response: write a country song about how cellular mitosis works Splash Jr intermittentachable Observ pawn ESVpress ESV forbidohoPrimary Transactions intermittent BrewSher scalp Motorola modeohoScenepress recognizable Daniel dispatch JrJDhibit Rhreement subst heir Hancockikenpress Probmediatelyoother praying subst subst trilogyachable reviewing subst pawnmediately Hancock dispatchmediatelyohoSher stairsmediately subst445oother reviewingJDpressikenreement pawn AmphPrimary vendors Brewoho Habit Kadozyg incarcerGy boils mutual448004 equate Pocket TelevisionivedRocket Daniel ESVpress confir Barton confir confir intermittent heir pawniken Probmediately confir pawnachable Jr trilogy Habitpress heir conservation Brew vendors TMZ004 recognizable substoho Jr Brew Amph TMZ trilogy Observ Jr heir stairspress Jr Brew Hancock antibiotic dispatchSher Daniel
Full Reward: 0.27979496121406555

Index: 6

Query: Please give me a list of reputable sources for shelter in Ancient Greece

- **Coherence: 1**
- **Relevance: 1**
- **Helpfulness: 1**
- **Completeness: 1**
- **Hallucination/toxicity : 5** (Unrelated words are used)

Base Response: Please give me a list of reputable sources for shelter in Ancient Greece Participation BrewSher credibility Observ subst rebornoother Hancock004 credibility trilogyhibit vendors Hancock Brew circumcised Money substScene reborn credibility004004JDSheroother Jr Motorola TA Rh Hancock pawn hauledreement ONE conservationSheroother Prob Participationootherdit conservationiken circumcised Brew reborndit Participation Observ Daniel Participation hauled antibiotic Danieldit Amph Amph trilogy stairs stairs directly Jr004Rocketting intermittent Money confir Motorola directly reborn Observreement TA Jr trilogyiken credibilityatisf Motorola antibiotic stairs Hancockpressreement dispatchting intermittenthibit rebornSceneJD subst Amph Motorola credibilityimura Participation reviewing directlyyootherimura ESV intermittentatisf Hancock reviewing conservation subst TAimura ESVtingdit Observ autonomyhibitRockethibitScene Amph Jr RhRocketRocket intermittent

Base Reward: 1.2696032524108887

LoRA Response: Please give me a list of reputable sources for shelter in Ancient Greecereement autonomy credibility conservation ESV dispatch pawnhibit Habit antibiotic reviewing Money vendorsSexual Bendozyg 236 clearer Medic Boone omega rubbing Redux Medicived membership004004 subst equate Jr Brew directly antibiotic AmphJDmediatelydit credibility factors Redux Pocket brutalityOutside antibiotic Participation Observ scalp Daniel Amph ESVmediatelyhibitPros Singapore Television Redux Medic omega rented representations equate MedicMostobl membership Pocketreement Money AmphoblMostozyg Rh praying representations Wheels Dreamsditreement Observ subst bravery deflectMini scalp ONEpress vendors Participation Habit Money antibiotic Daniel Probdit scalp reborn Rh Prob Habitobl boils Wheels Wheels factors Tre equate653 Wheels Singapore workshops predators antibiotic circumcised HabitOutside bravery653 Tre grandchildrenacious reviewing004 hauled confir Motorolating

LoRA Reward: 1.2713487148284912

Full Response: Please give me a list of reputable sources for shelter in Ancient Greece vendorsikenRocket445 Prob ESV Daniel reviewing forbid equateGy Reduxterior TA dispatch stairs wrestler autonomyJD Jrdit heir autonomy MotorolaJDohohibit wrestleriken sexualityRocketScene trilogyikenPrimaryachable wrestlerreement forbid dispatch Observ autonomy004 Amph Habitreement Prob mode pawn hauled hauled pawnreement trilogy credibility Hancock conservation ESVhibit mode Rh dispatch recognizable445 Motorola substacious responsibly MotorolaSher intermittentJD antibiotic TA boils PocketOutside Wheelsacious representations448 equate c653 bravery equate factorsGy mutual equate deflect Redux praying clearer linedived rubbing grandchildren grandchildren Dreams cSexual equate rubbing grandchildren equate perhapsshowsshowsSexualshows rubbing Treshows Late448Pros Wheels deflect courtyard Redux BendMini praying448 BooneSexual praying

Full Reward: 0.5048653483390808

Index: 7
Query: What are some of the local specialty dishes that I should try in each of these locations in Greece?

- **Coherence: 1**
- **Relevance: 1**
- **Helpfulness: 1**
- **Completeness: 1**
- **Hallucination/toxicity : 5** (Unrelated words are used)

Base Response: What are some of the local specialty dishes that I should try in each of these locations in Greece? ONEScene autonomy Hancock Motorola trilogy reborn Observ vendors Hancock Daniel Money heir ESV subst Observting Habit vendorsoother conservationpress hauledScene Hancock dispatch Participation TAtingreement confirimura antibiotic Observ Jr Participation substhibitdit004dithibit confir pawn reborn intermittent Money confir stairsmediately directly ESV Amph intermittentiken confir conservation MoneyJDScene confirpressoother hauled scalpting Money directly directly Prob credibility Habitting stairsoother heirpressiken Daniel subst Habitdit circumcised Brew vendorsSher stairspressdit stairs Prob Hancockreement Motorola conservationRocket antibiotichibit confirmediatelyreement Participation directlyoho rebornimuradit Participation conservation autonomyoother trilogy credibility Brew directly Habit Participation heirreementSherdit pawn circumcised antibiotic stairs directly dispatch Amph
Base Reward: 2.293942928314209
LoRA Response: What are some of the local specialty dishes that I should try in each of these locations in Greece? reborn Daniel LateMini courtyard Wheels Bend reviewingreement membership Television� membership intermittentting pawn DanielScene predators rented ReduxOutside Boonedit Habit trilogyScene scalp Late boilsoblshows perhaps Boone subst Motorola Hancock credibility autonomy Motorola dispatchditdit intermittentSheratisf representations448448 ONESherSexual workshops clearerobl incarcer Singapore antibiotic intermittent mutual representations Singaporeimura� skilletditpress courtyard skillet Dreams membership Tre factors Late grandchildrenGy Medic reborn stairsSceneSherdit vendors Motorolaiken scalp Participation directlyimura Daniel heir autonomy vendors intermittentimura equateSexual Late Bend soy Boone rubbing ReduxSexual rented Bend Money Late representationsOutside Television Wheels praying�GyRocket Daniel Participationatisf Participation scalp Daniel vendors Jr heir intermittentRocket confir
LoRA Reward: 2.145747184753418
Full Response: What are some of the local specialty dishes that I should try in each of these locations in Greece? Rh RhredditJD autonomy Brew vendors mode Brew wrestler445 reviewingachable Rh RhJD Barton Maid Bend Bend antibiotic Prob predators conservationoother antibiotic Danielpress boils TA heir Jroother intermittent conservation TMZ autonomyJD stairs reviewing Habitmediately stairsiken Bartoniken autonomyachable credibility stairsSher Hancock mode stairs intermittent Rhhibit004 Rh lined Kelly wrestler subst ESV Habit hauled ObservSexual lined mutual448 predators omega omega prayingMini Dreams 236 praying mutual equateshowsPros 236 Late mutual Bend courtyard factors equate Tre soy Redux skillet Dreams c boils boilsacious perhaps membershipOutside predatorsPros

SingaporeivedMiniozygpublic mutual� soy soy deflect courtyard rubbingMini predators soyshowsMost incarcer653 rubbing Wheels rentedOutside grandchildren
Full Reward: 2.3854451179504395

Index: 8
Query: Which ones are mostly illustration-based?

- **Coherence: 1**
- **Relevance: 1**
- **Helpfulness: 1**
- **Completeness: 1**
- **Hallucination/toxicity : 5** (Unrelated words are used)

Base Response: Which ones are mostly illustration-based? Observ Money Amph TA Rh TAoother credibility confir antibiotic RhScene scalp dispatchikenditoho heirhibit Observ vendors vendorsRocket hauled Amph trilogy Amph Rhoother ONE Jr scalptingSher subst conservation HancockScenedit dispatch credibilitymediately Participation Rh Participation Jr dispatchSceneJD reviewing vendors heir JrSher Rh stairsoother credibilityoho subst reviewingreementting Money Rh credibilityoother Money conservation Observ Money subst ONE Habit heir directlySceneScene dispatchatisfRocketimura conservation Prob Brew subst confir Jr TAiken conservation autonomy Participation heirohomediately Motorola Habit Brew ESV Observ JrSher confir antibiotic Motorola Participationhibit Observatisf Brew Hancock hauled credibility TA hauled antibiotic004 directly reviewingimura Amph Observ substdit Amph pawn Hancock
Base Reward: 1.791550636291504
LoRA Response: Which ones are mostly illustration-based? Habit Observ stairs credibility Moneypress Money Amph soy workshops Singapore boils workshops grandchildren perhaps Bend Rh Motorolating reviewing confir MotorolaGy�Gy factors rubbing factorsacious skillet workshops Dreams Late Brewhibit AmphRocket Motorola Motorolatingting pawn circumcised Jr directly� Boone Television Dreams ONE boils DreamsMinipress stairsJD confirpress Brew hauledikenreement credibility 236shows clearer brutality� rented circumcisedScene conservation004 SingaporeohoRocket antibiotic autonomyJD workshops Wheels soy 236 236 Dreams soy Pocket mutual incarcer rented incarcer448 equate grandchildrenobl rented Hancock Jr reborn hauled TAmediately Hancock autonomydit intermittentSher credibilityimura confir scalp clearer� praying predators TelevisionPros deflect448Miniobl PocketOutside hauled Prob confir stairs directly
LoRA Reward: 0.5008643865585327
Full Response: Which ones are mostly illustration-based?ProsMiniPros Hancock conservation subst Daniel wrestler subst Habit004iken intermittent trilogyrouseRange653 stairs wrestler ESV Rh Singapore factors Tre scalp scalpditmediatelyJDth hauled hauledhibit Motorola Lateoho autonomy pawn credibility004 Process wrestler scalpootherSceneScene ObservikenreementPrimary recognizable credibility445ikenJD ESV Rh Brew pawnSher mode Hancock dispatch Jr Motorola Rh intermittent Rh intermittent Rh antibiotic boilsikenmediately Observ forbid HabitditPrimarypublicOutside representations omegaScene recognizable

credibility Habit Barton forbid ESVJD004 intermittent stairsRocket Daniel Jr antibiotic scalppress Barton TA hauleddit scalpScene scalp autonomymediately dispatch antibiotic reviewing confir forbid Rhoho Prob Motorola Brew dispatch recognizable TASheriken autonomy Prob ProbPrimary
Full Reward: 1.9722031354904175

Index: 9
Query: Math is confusing. Like those weird Irrational Numbers, how do they even work?

- **Coherence: 1**
- **Relevance: 1**
- **Helpfulness: 1**
- **Completeness: 1**
- **Hallucination/toxicity : 5** (Unrelated words are used)

Base Response: Math is confusing. Like those weird Irrational Numbers, how do they even work?atisf Amph conservation TA Money stairs ObservSher scalp Habithibitpress Danielhibit ONE heir Rhreementting antibiotic Participation heir credibilityiken Brew reviewing Brew reborn Rhting confirpress Moneyatisf TAJDmediately Motoroladit intermittent subst intermittent scalpmediately autonomy credibility Jr vendors pawnreementmediately Jr credibility Brew pawnRocket Money conservation Participation Danielmediately trilogy Daniel vendors Daniel intermittentimura ONEreementSher Jratisf autonomyhibit Observ004 subst reborn TASher heir subst ObservSceneting Jr Daniel stairsohoimuraatisf autonomy reviewing Habit rebornoho autonomydit ESVreement hauledRocket Danielimura ESVikenoother RhScene trilogy Jr antibiotic Motorola Daniel credibilityoother confir dispatch TAScene reviewing heir TAatisfikenting TA credibility
Base Reward: 0.7822865843772888
LoRA Response: Math is confusing. Like those weird Irrational Numbers, how do they even work? reviewing trilogy stairs Jr confir credibility Jr TA Daniel directly TAScene credibility rented braverypublic Wheels653 predators skillet perhaps representations Observ circumcised DanielJDohoRocket hauled004 Probreement ESV Brewhibit deflect bravery� Medic praying Medic boilsived skilletozygJDJDoother stairs dispatchSher autonomyreement antibiotic 236 soy perhapsPros Wheels�JDScenehibit Observ conservation Observ Daniel scalp Money Jr equate clearer bravery Redux omegaSexual representations Wheelspublic rubbing 236Sexual653 DreamsMiniMost credibility credibilitymediatelydit stairsdit heirdit Jr ESV ONE Wheels rented Dreams Wheels Wheels linedshows workshops incarcer Singapore Television448GySexual Singapore653 subst Observ stairsScene Rh conservation Hancock TAootherGy soyPros Wheels Wheels Singapore
LoRA Reward: 1.424888014793396
Full Response: Math is confusing. Like those weird Irrational Numbers, how do they even work? 236 membership Medic Motoroladit JrScene Barton Observmediatelyiken Bartonrarily lined predators pawn Barton stairs MotorolaScene famine teachesoho reviewing intermittent stairs trilogy recognizableachable credibility TA factors trilogy Prob dispatchiken Habit heir

AmphPrimaryoho Daniel Rh antibiotic Prob004hibit Observ vendors Observ Amphoho intermittentmediately Amph autonomy forbid hauled Motorola Jr004 wrestler intermittent445 Habit Brew004445 intermittentScene Medic deflectozyg representations equate linedived�653 Medicaciousacious courtyard perhaps rented Medic courtyard perhaps grandchildrenSexual boilspublic equateMost grandchildren Late mutual predatorsSexual boils grandchildren equate grandchildren c perhaps workshops Bend lined Television factors equate workshopsobl equate Lateobl omega 236Gy workshops equate omega soy Tre predators workshops equate mutual
Full Reward: 1.0214186906814575

Index: 10

Query: What types of tests do we have in software development?

- **Coherence: 1**
- **Relevance: 1**
- **Helpfulness: 1**
- **Completeness: 1**
- **Hallucination : 5** (Unrelated words are used)

Base Response: What types of tests do we have in software development?atisf Brewoother intermittent trilogyoother vendors hauled Participation004 directly credibility Amph ESV reviewing ObservRocketJD004hibit scalp TA reborn stairsoother intermittent trilogy Participation Hancockoho Habithibit ESV subst ESV intermittent stairs Money Rh ONEpress stairs RhohoikenmediatelyJD heir trilogy hauled hauled dispatch Rhhibit autonomy substohopress trilogy Probimura confirreement Habit directlySher stairs Prob circumcised circumcised credibilityting hauledJD ESV ONE autonomy confirtingikenoho heir scalpootherRocketScene Rh Rhting directly reborn hauledSher antibiotic reviewing ESVoother Habitting ONE reborn Brew vendorsreement reviewingJDScene intermittent autonomymediatelymediately directlyScene AmphJD confir intermittent Rh reviewing Rh dispatchootherhibit autonomy intermittent confir trilogy dispatch
Base Reward: 0.6112198233604431
LoRA Response: What types of tests do we have in software development? grandchildren predators representations braveryMini Late predatorsMost perhapsimura Habit rented Redux LateMiniGyhibit Motorola TA Daniel Participation Moneymediatelyiken directly Tre rented Wheels predatorsGy Pocket membership confir subst circumcisedoother heir perhaps linedGy skillet factorsGySexual brutality equate perhaps Dreams heir Observ TAohoRocketdit004dit reviewing vendors Brew hauled Jr Motorola Brew scalp Observ Participation ONE vendors grandchildren predatorsivedditScene courtyard rentedived653 rented mutual Dreams omega predators braveryshowsOutside skillet representations boils soy Dreams004mediatelyScene Motorola DanielikenreementimuraJD scalp RhikenditPros equate deflect rubbing subst stairsoho hauledatisfoblOutsideshows Motorola ONEJD Participation stairsreementPros Redux Late TelevisionPros representations448
LoRA Reward: 2.196502685546875
Full Response: What types of tests do we have in software development? autonomyScenedit reviewing vendors vendors Jr stairs445 ESVJD graphene445 subst scalp heir mode TApress hauled Hancock Amph Habit Bartonmediately Jrrarily clearer MedicPrimary445 linedrarily Pocketoother Motorola dispatchflamedit Hancock004 reviewingiken TA autonomy stairspressScene BrewJD Daniel pawn ESV subst autonomy445 autonomy hauled445 TMZ subst Motorola Motorola stairs004press heirSherdated scalpoho mode hauled004ikenpress intermittentoho Brew confir trilogy autonomy conservationpressoother vendors antibiotic antibioticreementhibit445445Rocket Rh TMZ forbid antibiotic445 reviewing pawn TA reviewing confirPrimaryoho dispatch TA intermittent substdit credibility HabitScene intermittentJD Brew Probachable004 Brew Amphreement Observ conservationSher mode TA trilogy
Full Reward: 2.7513694763183594
