GRPO for Factuality

Finetuning for Factual and Cautious Responses

Wenbo Pan

March 11, 2025

Table of Contents

Improve Factuality

- Project focused on improving language model factuality
- **Key Goals:**
  - Reduce hallucinated responses
  - Improve factual consistency
  - Train models to acknowledge uncertainty

Basic Idea of Finetuning

- We finetune qwen to avoid hallucinated responses
- **Generate more factual and consistent answer:**
  - If it can give correct answer 4 out of 10 times, we want it to output it 10 out of 10 times
- **Acknowledge if it's unable to answer correctly:**
  - If it gives incorrect answers no matter how many times, we want it to output "I don't know"

Setup

- **Training Dataset:**
  - As SimpleQA is too difficult, SimpleQA, PopQA, SelfAware and TriviaQA are mixed evenly
- **Model:**
  - Qwen32B 2.5 with LoRA Training
- **Predefined Instruction:**
  - A cold start prompt is used to induce reasoning and refuse

System Prompt

System prompt

A conversation between User and Assistant. The Assistant must think step by step. Then give a brief answer in boxed[] if sure about the answer, otherwise the Assistant can return boxed[Unknown] if not sure.

Rewards

- We use **reinforcement learning (RL)**, which train a model to maximize given rewards
- The reward function both encourage factual accuracy and reduce hallucination
- Consider at each training step, model generates 20 responses:
  - **If any of them got correct:** the correct response gets 1 reward, process-correct gets 0.5 and other gets 0
  - **If non of them got correct:** not-attempting response gets 1 score and other gets 0

Reward Function Visualization

$$R(x) = \begin{cases} \left.\begin{array}{ll} 1.0 & \text{if correct answer} \\ 0.5 & \text{if process correct} \\ 0.0 & \text{otherwise} \end{array}\right\} \text{if ANY response is correct} \\ \left.\begin{array}{ll} 1.0 & \text{if "Unknown"} \\ 0.0 & \text{if attempting wrong answer} \end{array}\right\} \text{if NO response is correct} \end{cases}$$

RL Goal

- We use GRPO (the algorithm training Deepseek R1 to reason)

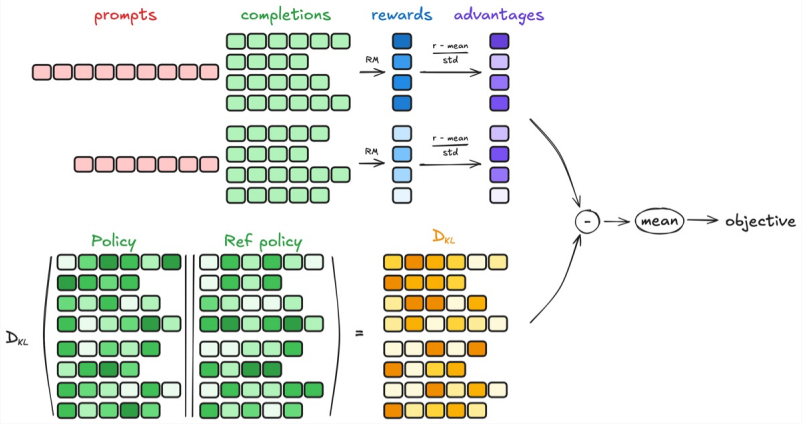

Figure 1: GRPO visualization showing the reinforcement learning approach

Training Progress

- Training is very noisy

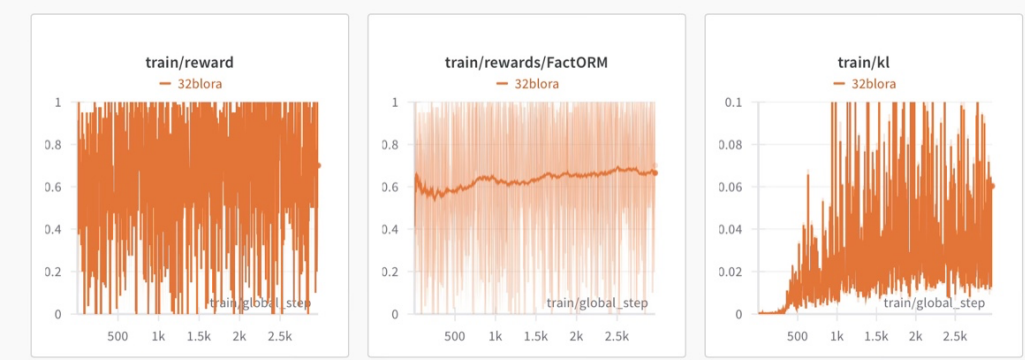

Figure 2: Noisy training metrics during the training process

Reward Progress

- But the rewards on eval dataset is increasing steadily

Figure 3: Steady increase in rewards on evaluation dataset

Eval Results Decomposition

Training Steps	Partial Correct	Partial NotAttempt	All NotAttempt	All Correct	All Attempt and Wrong
200	36.67	33.33	8.33	10.00	11.67
400	33.33	33.33	10.00	10.00	13.33
600	30.00	33.33	13.33	10.00	13.33
800	31.67	35.00	11.67	10.00	11.67
1000	26.67	28.33	20.00	13.33	11.67
1200	25.00	30.00	21.67	13.33	10.00

Table 1: Evaluation results breakdown by training steps

Observations

- The model is not deviate from original model much (small kl)
- Reward has high variance (many 0, many 1 across the training session)
- Model learned to not attempt questions in Partial Correct and Partial Not Attempt
  - We actually don't want Partial Correct to drop

Problems

- Too many training samples that are too easy or too hard
- Didn't see long reasoning or emerging behavior, although rewards are increasing
- Full training on 7B will encounter NaN
- The system prompt is too simple

Problems (Continued)

- Model will prioritize not attempting even for simple questions
- Training takes too much time (33 hours)

Improvement Strategies

- Try different system prompt for a better start point
- Filter the training set to exclude trivial or impossible samples
- Add kl clip to stablize 7B full training
- Apply fact and caution reward separately
  - To see how each one is working

Filtering Training Data

- Too many trivial or impossible samples (evident by lots of 0 and 1 rewards)
- A sample is just-hard-enough if the model only generates correct answers sometimes
- Use original model to predict each sample 16 times
- Filtering top 2500 samples with most varied outputs

Use Different System Prompts

- We find system prompts greatly affect model
- If we encourage model to be cautious, it will have more NotAttempt and less correct
- Experimenting different prompts gives a trade-off boundary
- We choose a balanced prompt for training

Prompt Selection Results

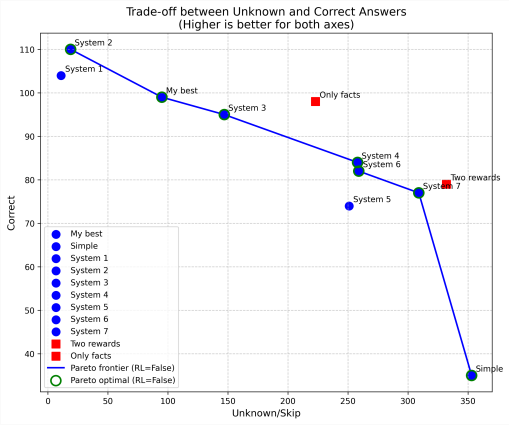

Figure 4: Trade-off between correctness and caution with different system prompts

Prompt Selection

- Show some example of prompts.

System 1

You are a hardworking assistant. When asked a question, you try your best to find the correct answer. [...] You provide a single word answer. First think through your reasoning in <think></think>tags, then give your answer in <answer></answer>format.

System 4

You are a hardworking assistant. [...] You make educated guesses when you don't have complete information. [...] First think through your reasoning in <think></think>tags, then give your answer in <answer></answer>format.

System 8

You are a hardworking assistant. [...] You explore different angles, check multiple sources, and challenge your assumptions. [...] First think through your reasoning, then give your answer in <think></think>and <answer></answer>format.

Use 7B Full Training

- 32B + Lora allows us to use stronger base model, but
  - Slower to train
  - Less optimization freedom (Lora only update few of the model parameters)
- Use 7B Full Training
  - In PPO methods like GRPO, we can enable policy clip to avoid model update too much in one step, stabilizing training
  - 7B takes 7 hours for one epoch

Second Round Experiment Design

- We test 4 settings
  - Only use fact reward
  - Only use caution reward
  - Use fact and caution reward at the same time
  - First fact reward, then caution
- Motivations:
  - (1) I find the training not stable if two rewards are applied
  - (2) Caution is easier than fact

Results

- The 7B models can also benefit from RL training
- The fact and caution rewards seems conflict
- The model will prioritize trying NotAttempt
- RL can push the frontier, but not that much

Next Step

- Make some reasonable efforts to improve the result
- Towards drafting the paper
  - Instead of focusing on better accuracy
  - I focus on balance between correct and caution
  - Propose a new metric and compare with existing baselines