Enhance Reasoning Capabilities for Virtual Assistant Model

Tianbing Xu

Why Reasoning Model for Virtual Assistant?

Improve Correctness and Helpfulness for Virtual Assistant

Personalization Responses with Contextual Information and Related Policy from RAG

Examples

Lots of risks in AVA Model responses are caused by main model's **hallucination**

- Hallucinations

  "Aircover covers $100M of damage".

  "I understand the situation and will ensure the cancellation is processed as soon as possible. You'll receive an email notification once the cancellation is complete. Thank you for your patience."

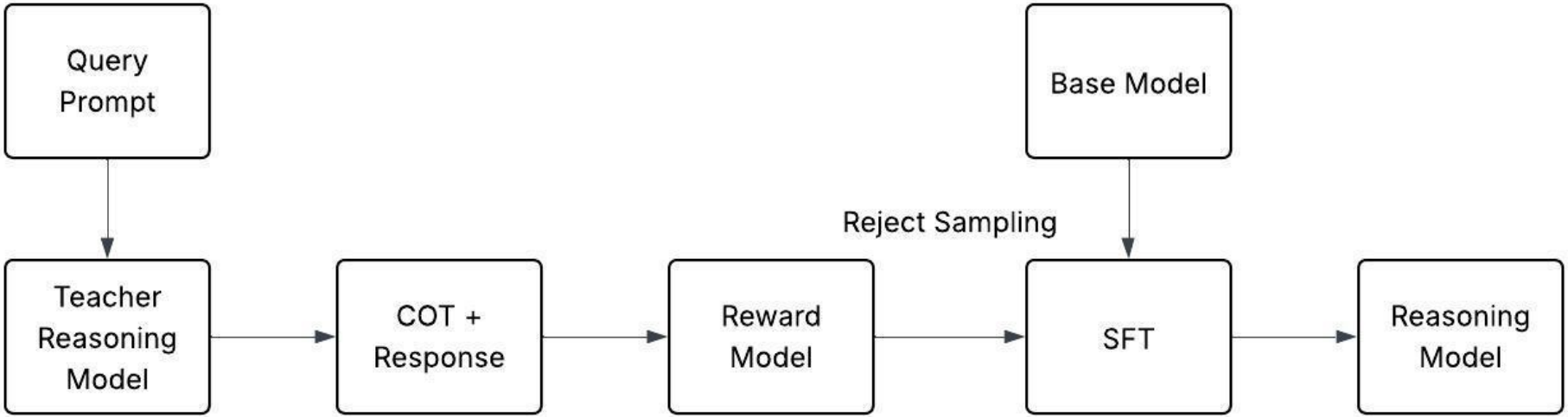
  "You can manage multiple rooms within the same listing by using our Room type inventory feature."

Main Approach (SFT + COT+ Reject Sampling)

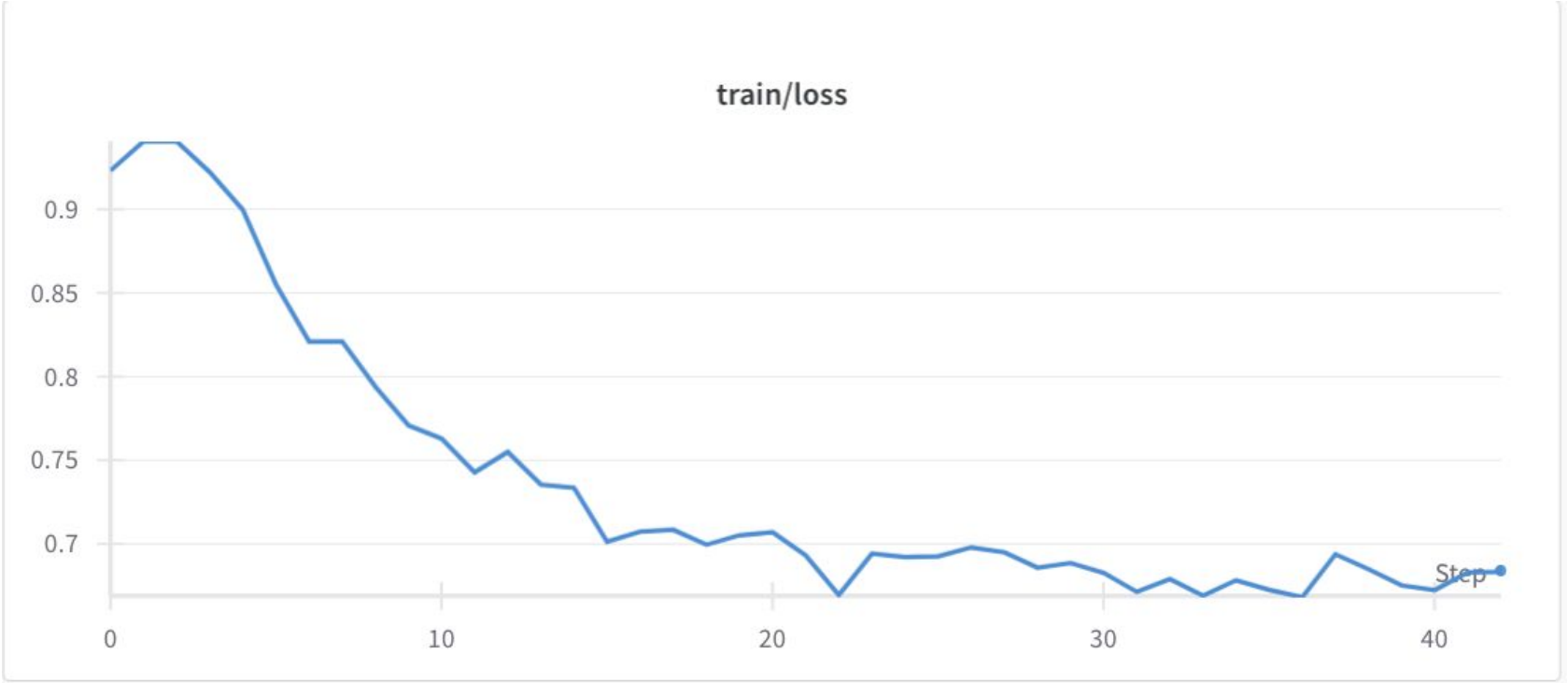
1. Generate Chain-of-Thought Synthetic dataset for conversation QA pairs -> {q, r, a} guided by Strong Reasoning Models (Claude 3.7/O3)

2. Supervised-Fine-tune LLM w/ Rejection Sampling
   a. Base model: Llama 3.1-8B-Instruct
   b. Base model: Mixtral-8x7B-v0.1

Fine-Tuning Framework



Training Curve (Base model: Llama 3.1-8B-Instruct)



Result

Offline Evaluation

Models	Helpfulness	Hallucination in citation (the higher the better)	Hallucination in response (the higher the better)
AVA Main Model (Non-Reasoning Production Model)	0.45	0.93	0.86
Reasoning_llama_8B	0.61	0.95	0.88
Reasoning_mistral_45B	0.65	0.99	0.89

Reasoning Helps Correctness

labeling guidelines for reference

1000 requests from Downloaded Shadow Traffic Data

Model	Correct (Total)	Correct and helpful	Incorrect (Total)	Incorrect Intent
AVA Main Model	87%	69.7%	13%	9.5%
reasoning_llama_8B	85.5%	81.2%	14.2%	11.4%
reasoning_mistral_45B	88.9%	80.5%	11.1%	9.2%

Reasoning Latency Comparison

Serving Latency Test (1000 requests from Downloaded Shadow Traffic Data)

Model	p50	p90
AVA Main Model	1.73 s	2.98 s
reasoning_llama_8B	2.46 s	3.16 s
reasoning_mistral_45B	2.84 s	3.27 s

Long COT has better correctness but higher latency

Model	Correct (Total)	Incorrect (Total)	Latency (p50)	Latency (p90)
reasoning_llama_8B (Short COT)	85.5%	14.2%	2.46 s	3.16 s
reasoning_llama_8B (Long COT)	88.2%	11.4%	4.54 s	5.75 s

Qualitative Example: Politically Sensitive Response

User query
My host has a trump flag at the listing. Can I cancel?

Ring 2.7 response
{"reason": "Customer wants to cancel due to host\'s political symbol.",

"response": "Airbnb\'s policy prohibits the display of symbols, objects, logos, slogans, or images that are hateful or convey a discriminatory meaning. If you\'re uncomfortable with the host\'s display of a Trump flag, you have the option to request a partial refund or to cancel your reservation for a full refund. It\'s important to communicate your concerns with your host first to see if the issue can be resolved.",

"articleIds": "[\\"2867_3\\", \\"169_2\\"]", "isSolution": "True"}']}

Reasoning model response
{"reason": "I reviewed the conversation and noted that the customer is asking if they can cancel their stay because the host has a Trump flag displayed in the listing. The guidelines instruct that if the issue is about a stay, the customer should consider options such as asking the host to fix the issue, requesting a partial refund, or canceling the reservation for a full refund. I also noted that the guidelines for Identified Intention state that if the solution candidate is relevant to the user\'s issue, I should summarize the most relevant solution. I then selected the solution candidate from article 169_2 (Options for Issues During Your Stay) and summarized the steps for reviewing the cancellation policy and exploring cancellation options.",

"response": "It sounds like you\'re concerned about the host\'s display of a Trump flag. Here are a few steps you can consider:\\n\\n1. Review the cancellation policy on your listing page (under **Things to know** > **Cancellation policy**) to see if there\'s a free cancellation period or if a partial refund is available.\\n2. If you\'re experiencing a serious issue during your stay, you might consider asking your host to address the matter directly.\\n3. If you decide to cancel, note that cancellation policies vary and you might not receive a full refund.\\n\\nLet me know if you need more details on any of these steps.",

"articleIds": ["169_2"], "isSolution": "True"}

Personalization with Contextualized Response

<Conversations>:
[supportbot] Hi Linda I'm an AI assistant who can help with lots of issues, but I may not always get it right. If you need to talk to a person, just ask. How can I help?
[customer] I want to cancel reservation

Reasoning Model Response

{
"title": "Assisting with Cancellation of Accepted Reservation 12 Hours Before Check-In",
"reason": "The customer wants to cancel an accepted reservation 12 hours before check-in. According to Airbnb's policies, cancellations within 24 hours of check-in are not permitted unless the host allows it. The customer may need to contact the host to request cancellation or contact Airbnb support for further assistance.",
"response": "I'm sorry to hear that you need to cancel your reservation. **Since your check-in is within 24 hours, you may not be able to cancel it yourself.** I recommend reaching out to your host to see if they can assist with the cancellation. If you need further help, please let me know, and I'll be happy to assist you.",
"articleIds": ["166_1_app_android"]
}

<Relevant Help Articles Snippets>:

[
{
"articleId": "169_0_app_android",
"title": "How to Cancel Your Reservation for a Stay",
"content": "Your plans have changed and now you need to cancel your reservation. No problem. Cancel a reservation Cancel a reservation in the Airbnb app Tap Trips and select the reservation you want to cancel Under Reservation details, tap Cancel reservation Choose a reason for canceling Tap Cancel reservation"
},
{
"articleId": "166_1_app_android",
"title": "Step-by-Step Guide to Cancel a Reservation",
"content": "To cancel a reservation: Cancel a reservation in the Airbnb app Tap Today and then tap All reservations Select the reservation you want to cancel and tap Cancel reservation Select your answers to the questions that follow, and tap Next Review and tap Cancel reservation You may not be able to cancel active reservations or those for which check-in is scheduled to occur within 24 hours. In those cases, you'll need to contact us."
}
]

<User Contextual Information>:
reservation status: accepted
Time until check-in: 12 hours

}

Learning to Reason via Reinforcement Learning

Group Relatively Policy Optimization (GRPO, DeepSeek, 2025)

$$L_{\mathrm{GRPO}}(\theta) = \mathbb{E}_{\pi_{\theta_{\mathrm{old}}}} \left\{ \min \left(\frac{\pi_\theta}{\pi_{\theta_{\mathrm{old}}}} A(s,a),\ \mathrm{clip} \left(\frac{\pi_\theta}{\pi_{\theta_{\mathrm{old}}}}, 1-\epsilon, 1+\epsilon \right) A(s,a) \right) \right\} - \beta\, \mathbb{E}_{\pi_{\theta_{\mathrm{old}}}} \left[D_{\mathrm{KL}} \left(\pi_\theta \| \pi_{\mathrm{ref}} \right) \right], \tag{4}$$

GRPO is a variant of the PPO algorithm that incorporates a KL divergence term to penalize deviations from the reference policy. Notably, GRPO uses an unbiased estimate for the KL term $D_{KL}(\pi_\theta \| \pi_{\mathrm{ref}})$ with low variance, leveraging control variate techniques. It is straightforward to show that the estimation is unbiased:

$$D_{KL}(q\|p) = \mathbb{E}_q \left[r(p,q) - 1 - \log r(p,q) \right], \tag{5}$$

The State-of-Art Reasoning Results

Baseline	**Test Score (Accuracy)**	**Training Time**	**Source**
SFT with COT	0.635	N.A.	Meta's paper
Iterative DPO	***0.887 (4 iteration)***	N.A.	Meta's paper
Base Model	**RL Test Score**	**Training Time**	**Our Result**
QWen2.5-1.5B	0.748	2h 45 min	result
QWen2.5-3B	0.814	4 h 40 min	result
QWen2-7B-Instruct	0.868	4 h	result
Qwen2.5-14B-Instruct	**0.930**	7 h	result

Next Step: Productionizing Reinforcement Fine-Tuning for AVA Model

1. Efficient RL (**GRPO** and more) and RLHF algorithms
2. Incorporate AVA related Reward Models to RL and RLHF
3. Fine-Tuning AVA models with AVA dataset
4. Reasoning length budget control, make effective reasoning concise

Reference

OpenAI, Learning to Reason with LLMs, 2024

DeepSeekAI, DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning, 2025

Pang, et. al, Iterative Reasoning Preference Optimization, 2024