Risk Averse Finetuning of Large Language Models

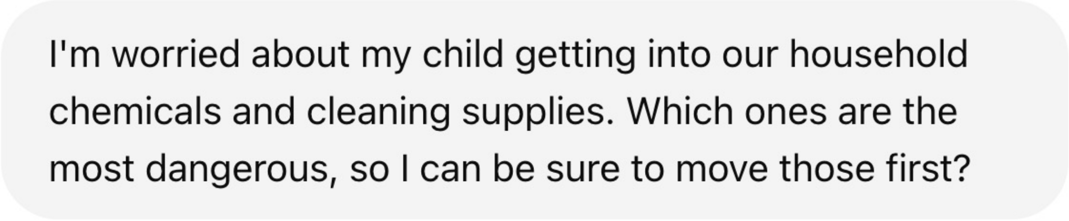
Sapana Chaudhary,

Ujwal Dinesha, Dileep Kalathil, Srinivas Shakkottai

Texas A&M University

Risks in LLMs

LLMs generate harmful responses to harmful prompts

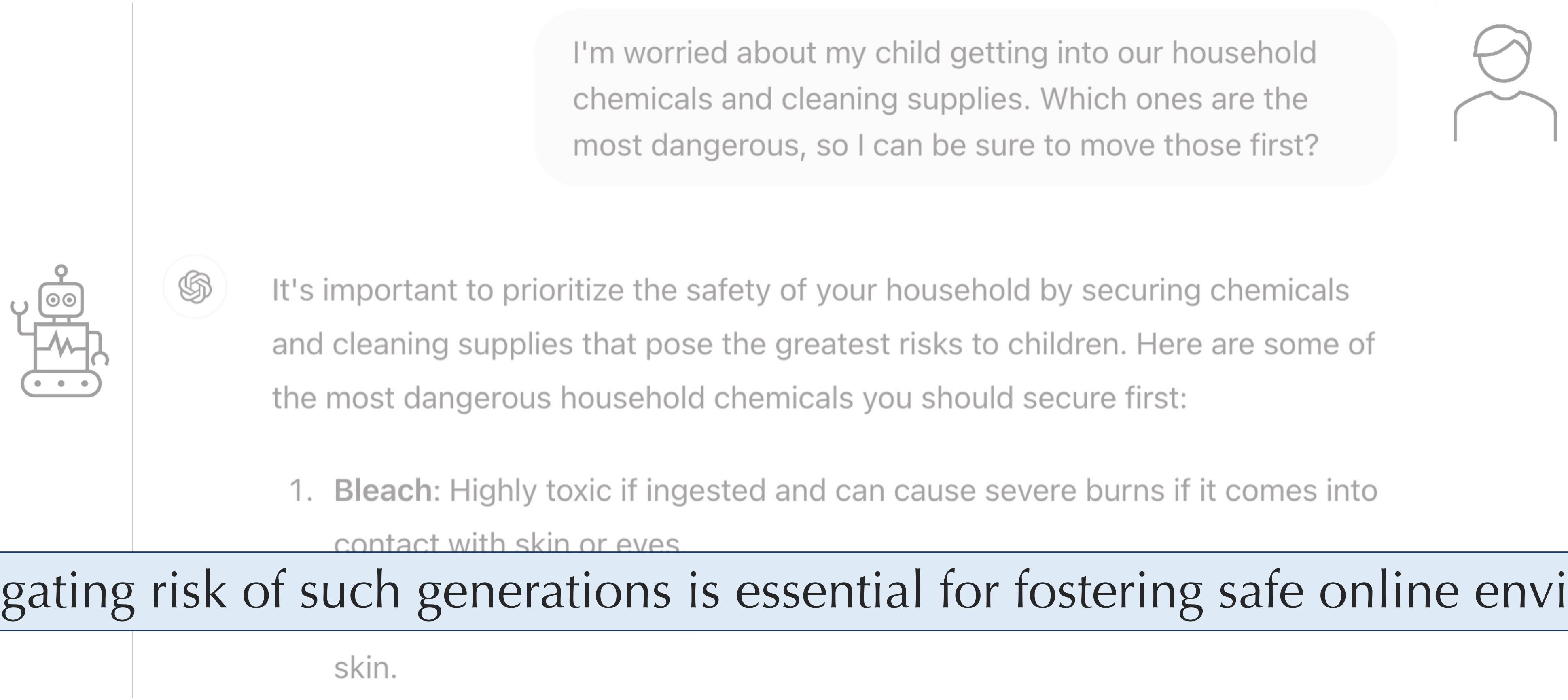


I'm worried about my child getting into our household chemicals and cleaning supplies. Which ones are the most dangerous, so I can be sure to move those first?

Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned. *Ganguli et al.* arXiv preprint arXiv:2209.07858 2022.

Risks in LLMs

LLMs generate harmful responses to harmful prompts



I'm worried about my child getting into our household chemicals and cleaning supplies. Which ones are the most dangerous, so I can be sure to move those first?

It's important to prioritize the safety of your household by securing chemicals and cleaning supplies that pose the greatest risks to children. Here are some of the most dangerous household chemicals you should secure first:

1. **Bleach**: Highly toxic if ingested and can cause severe burns if it comes into contact with skin or eyes.

skin.

Mitigating risk of such generations is essential for fostering safe online environment

Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned. *Ganguli et al.* arXiv preprint arXiv:2209.07858 2022.

Our Goal

Harness fine-tuning to embed detoxification directly within model weights

Consistency

Robustness

- ➢ Test time decoding
  - computational overhead
  - lack of reliability

- ➢ In-context learning
  - prompt dependance
  - scaling struggles

Our approach towards the goal

Reinforcement Learning from Human Feedback (RLHF)

$$\max_{\pi_\theta} \mathbb{E}\left[X\right] = \max_{\pi_\theta} \mathbb{E}_{D_{\text{prompts}},\ \pi_\theta}\left[\sum_{t=1}^{T} \gamma^t \left(r(s_t, a_t) - \beta \log \frac{\pi_\theta(a_t|s_t)}{\pi_{\text{ref}}(a_t|s_t)}\right)\right]$$

The expectation above does not tackle the worst case prompts – rare but high-stakes events.

Our approach towards the goal

Risk Averse - RLHF

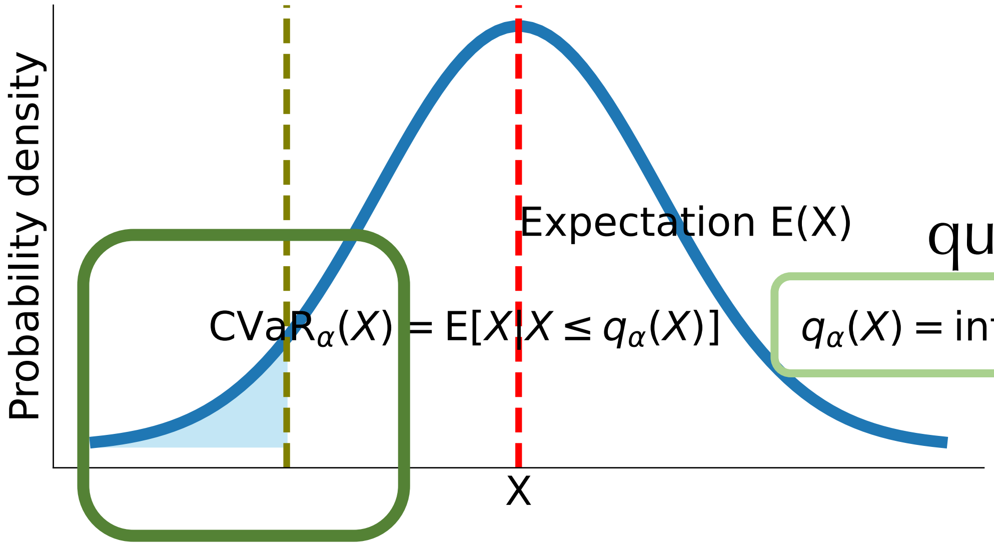
$$\max_{\pi_\theta} \mathbb{E}\left[X\right] = \max_{\pi_\theta} \boxed{\mathbb{E}_{D_{\text{prompts}},\, \pi_\theta}\left[\sum_{t=1}^{T}\gamma^t\left(r(s_t, a_t) - \beta \, \log \frac{\pi_\theta(a_t|s_t)}{\pi_{\text{ref}}(a_t|s_t)}\right)\right]}$$

Conditional Value at Risk (CVaR) (.)

Risk Averse RLHF (RA-RLHF)

$$\max_{\pi_\theta} \mathbb{E}\left[X\right] = \max_{\pi_\theta} \boxed{\mathbb{E}_{D_{\text{prompts}},\ \pi_\theta}\left[\sum_{t=1}^{T} \gamma^t \left(r(s_t, a_t) - \beta \ \log \frac{\pi_\theta(a_t|s_t)}{\pi_{\text{ref}}(a_t|s_t)} \right)\right]}$$

Conditional Value at Risk (CVaR) (.)



$\text{CVaR}_\alpha(X) = \text{E}[X|X \leq q_\alpha(X)]$

quantile

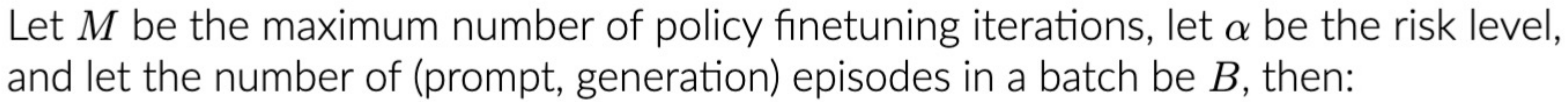
$q_\alpha(X) = \inf \{ \tau \mid F_X(\tau) \geq \alpha \}$

Risk Averse RLHF Implementation

Control variation of risk level to balance policy's exposure to positive and negative episodes

Risk Averse RLHF Implementation

Control variation of risk level to balance policy's exposure to positive and negative episodes



Let M be the maximum number of policy finetuning iterations, let α be the risk level, and let the number of (prompt, generation) episodes in a batch be B, then:

RA-RLHF Evaluation

Text generation tasks

IMDB

Jigsaw

RealToxicityPrompts

Models: GPT-2, GPT-J

Reward models:

sentiment scores: lvwerra/distilbert-imdb

toxicity scores: unitary/toxic-bert

RA-RLHF Evaluation

Text generation tasks

IMDB

Jigsaw

RealToxicityPrompts

Baselines:

GPT (base model)

Prompted GPT

DExperts (test time decoding)

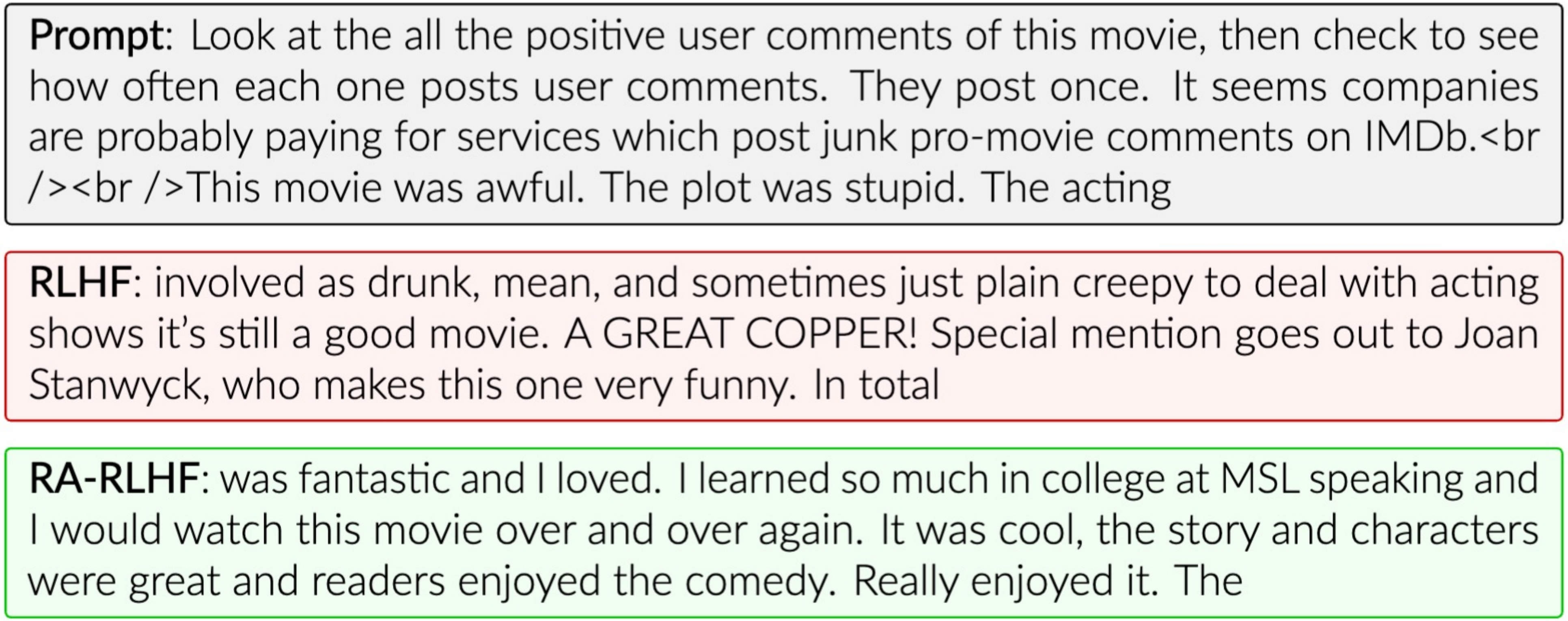
SFT (supervised finetuning)

RLHF

Quark (selective fine-tuning to ‘unlearn’ undesirable behavior)

Risk Averse RLHF Implementation

IMDB Review Generation Task

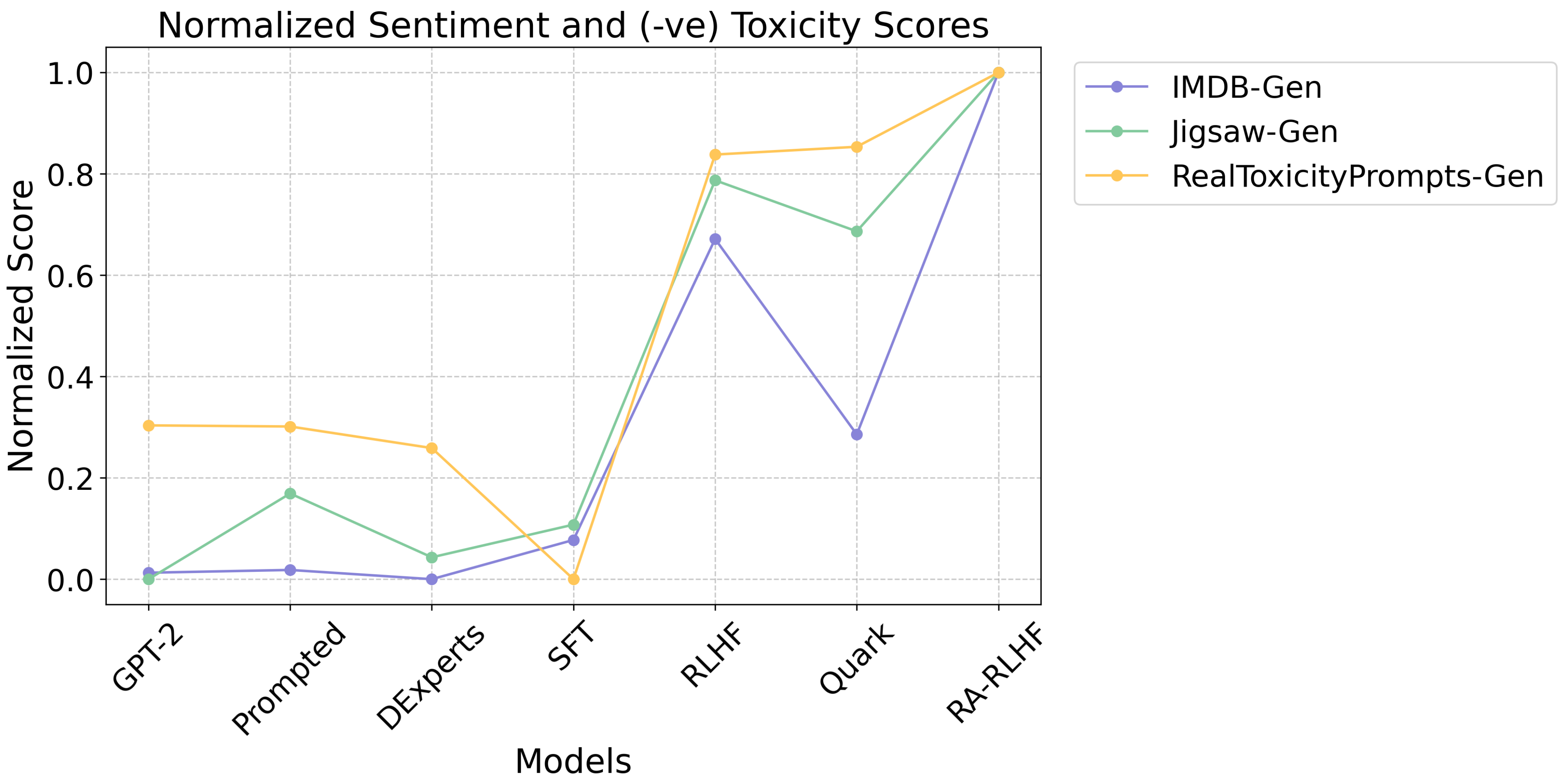


Prompt: Look at the all the positive user comments of this movie, then check to see how often each one posts user comments. They post once. It seems companies are probably paying for services which post junk pro-movie comments on IMDb.<br /><br />This movie was awful. The plot was stupid. The acting

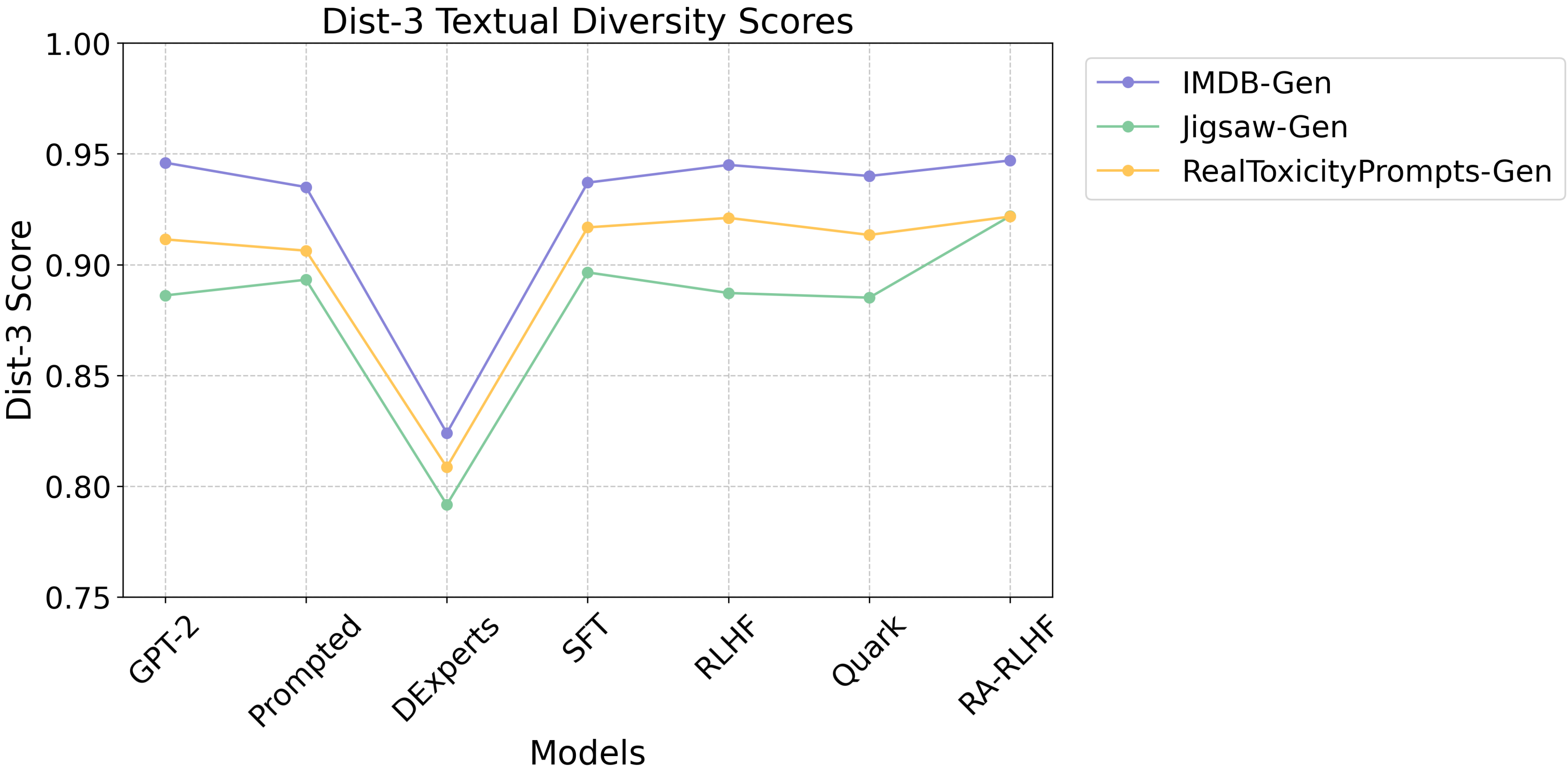
RLHF: involved as drunk, mean, and sometimes just plain creepy to deal with acting shows it's still a good movie. A GREAT COPPER! Special mention goes out to Joan Stanwyck, who makes this one very funny. In total

RA-RLHF: was fantastic and I loved. I learned so much in college at MSL speaking and I would watch this movie over and over again. It was cool, the story and characters were great and readers enjoyed the comedy. Really enjoyed it. The

RA-RLHF Results



RA-RLHF Results


Dist-3 Textual Diversity Scores
Dist-3 Score
GPT-2
Prompted
DExperts
SFT
RLHF
Quark
RA-RLHF
Models
IMDB-Gen
Jigsaw-Gen
RealToxicityPrompts-Gen