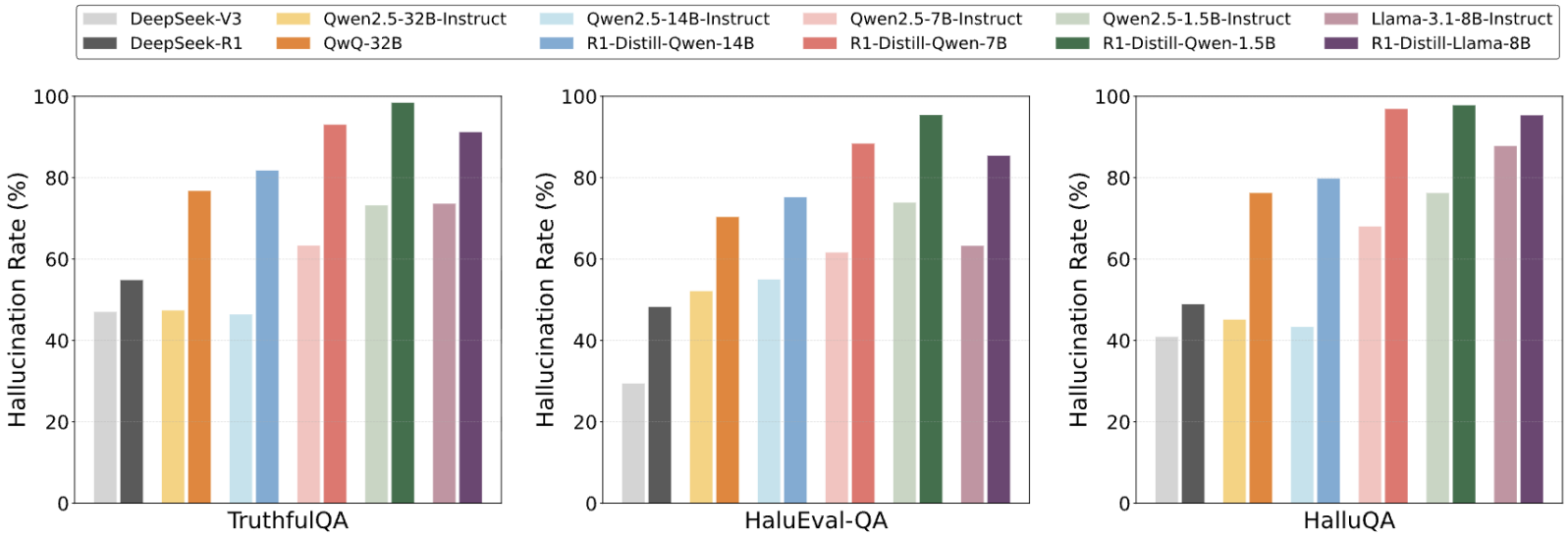
Reasoning Models Hallucinate More: Factuality-Aware Reinforcement Learning for Large Reasoning Models

Junyi Li, Hwee Tou Ng

Department of Computer Science
National University of Singapore

Background

- Research question
  - To what extent do large reasoning models maintain factuality after reinforcement learning (RL) fine-tuning?
- Preliminary experiments
  - 6 groups of 12 LLMs with or without RL fine-tuning



Theoretical Analysis

- Three key factors to hallucinations
  - High-variance gradient

$$p\,\|\nabla_\theta \log \pi_\theta(y^*|x)\|^2 - \|p\,\nabla_\theta \log \pi_\theta(y^*|x)\|^2 = p(1-p)\,\|\nabla_\theta \log \pi_\theta(y^*|x)\|^2$$

  - Entropy-induced randomness

$$H_\theta(x) = -\sum \pi(y|x) \log \pi(y|x) \geq H_{\min}(\epsilon) > 0.$$

  - Susceptibility to spurious local optima

Approach

- Reward function design
  - Step-wise factuality reward

$$\mathcal{R}_{\text{factuality}}(z_j) = \begin{cases} 1, & \text{if the sentence } z_j \text{ can be entailed from the evidence } \mathcal{K} \\ 0, & \text{if the sentence } z_j \text{ is neutral to the evidence } \mathcal{K} \\ -1, & \text{if the sentence } z_j \text{ contradicts the evidence } \mathcal{K} \end{cases}$$

  - Answer correctness reward

$$\mathcal{R}_{\text{answer}}(y) = \begin{cases} 1, & \text{if the final answer fully matches the ground truth} \\ 0, & \text{if the final answer does not match the ground truth} \end{cases}$$

  - Final reward

$$\mathcal{R}_{\text{final}}(y) = \mathcal{R}_{\text{answer}}(y) + \frac{1}{N}\sum_{j=1}^{N} \mathcal{R}_{\text{factuality}}(z_j)$$

Approach

- Factuality-aware policy optimization
  - Advantage adjustment

$$\hat{A}_{i,t} = \begin{cases} A_i, & \text{if } A_i > 0 \wedge \mathcal{R}_{\text{factuality}}(z_j) = 1 \text{ or } A_i < 0 \wedge \mathcal{R}_{\text{factuality}}(z_j) = -1 \\ -A_i, & \text{if } A_i > 0 \wedge \mathcal{R}_{\text{factuality}}(z_j) = -1 \text{ or } A_i < 0 \wedge \mathcal{R}_{\text{factuality}}(z_j) = 1 \\ A_i, & \text{if } A_i = 0 \text{ or } \mathcal{R}_{\text{factuality}}(z_j) = 0 \end{cases}$$

  - Policy optimization objective

$$\mathcal{J}_{\text{FSPO}}(\theta) = \mathbb{E}_{y \sim \pi(\cdot|x)}$$

$$\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|y_i|}\sum_{t=1}^{|y|}\left\{\min\left[\frac{\pi_\theta(o_{i,t}|x, y_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}|x, y_{i,<t})}\hat{A}_{i,t}, \text{clip}\left(\frac{\pi_\theta(o_{i,t}|x, y_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}|x, y_{i,<t})}, 1-\epsilon, 1+\epsilon\right)\hat{A}_{i,t}\right]\right\}$$

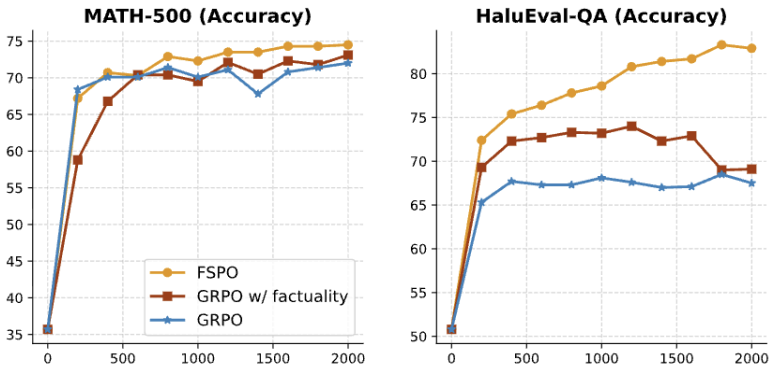
Experiments

- Main results

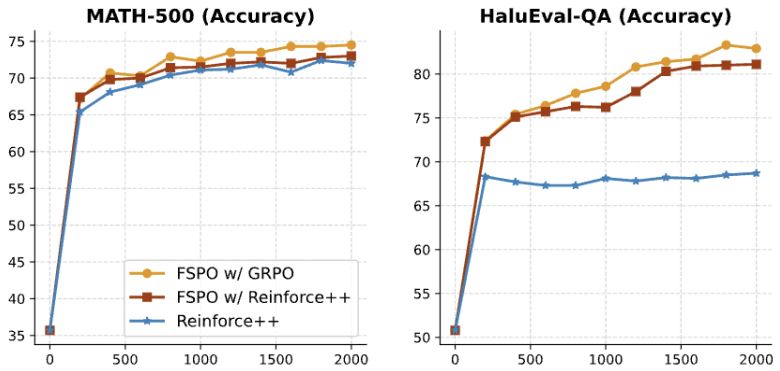
Model	GSM8K	MATH 500	AIME24 (Pass@1)	AIME25 (Pass@1)	TruthfulQA↑	HaluEval-QA↑	HalluQA↑
API-based Models							
DeepSeek-V3	89.3	90.2	39.2	26.6	53.0	70.6	59.1
DeepSeek-R1	92.0	97.3	79.8	66.7	45.2	51.8	51.1
GPT-4o-0513	90.8	74.6	9.3	13.3	59.0	62.6	53.1
GPT-o1-1217	92.3	96.4	79.2	76.7	62.4	70.2	56.0
Reasoning Models							
QwQ-32B	88.6	89.8	79.5	56.7	23.2	29.6	23.7
R1-Distill-Qwen-32B	87.4	94.3	72.6	53.3	19.7	33.5	26.9
R1-Distill-Qwen-14B	85.1	93.9	69.7	50.0	18.2	24.8	20.2
R1-Distill-Qwen-7B	84.3	92.8	55.5	33.3	6.9	11.6	3.1
R1-Distill-Llama-8B	82.1	89.1	50.4	26.7	8.8	14.6	4.6
Open-source Models							
Qwen2.5-7B-Base	65.2	35.7	3.3	3.3	38.2	48.0	39.5
Qwen2.5-7B-Instruct	73.2	51.6	6.7	3.3	36.7	38.4	32.0
Llama3.1-8B-Instruct	77.5	33.1	6.7	0.0	26.4	36.7	12.2
FSPO (Qwen-Base)	**89.5**	**75.5**	16.7	**13.3**	**58.4**	**83.0**	**52.0**
FSPO (Qwen-Instruct)	89.4	74.7	**20.0**	**13.3**	54.0	64.7	50.0
FSPO (Llama-Instruct)	86.2	68.3	13.3	6.7	41.1	67.1	42.0

Experiments

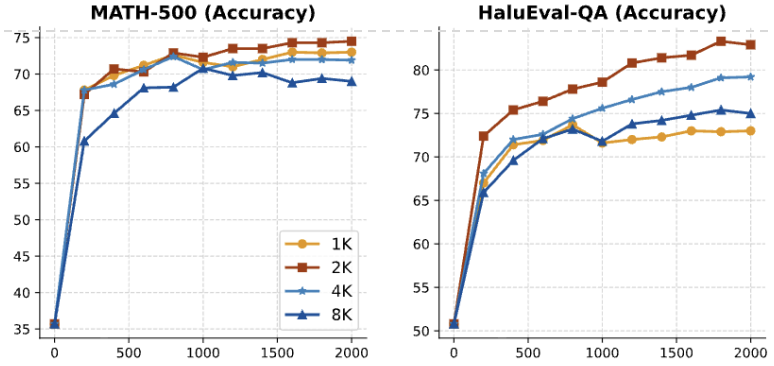
- Further Analysis



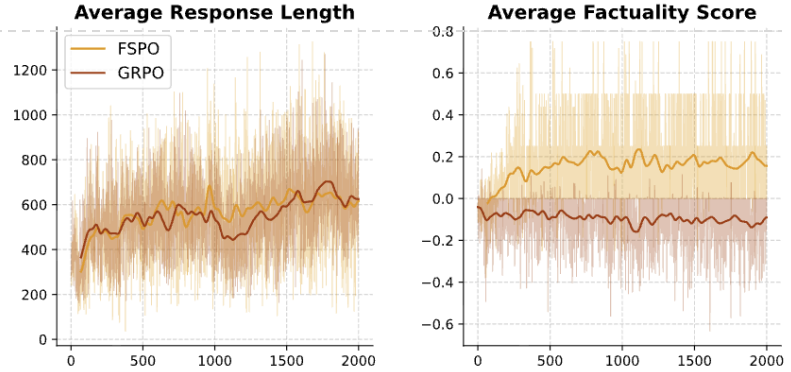
Ablation Analysis



RL Algorithm Analysis



Number of Training Samples



Factuality Improvement Analysis