BRiTE: Bootstrapping Reinforced Thinking Process to Enhance Language Model Reasoning

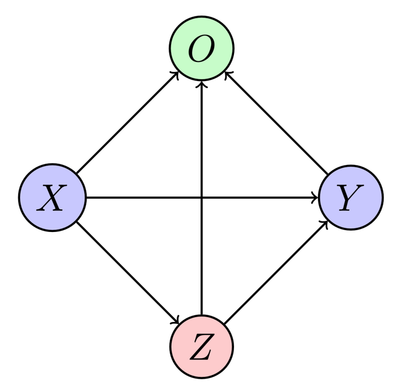
ICML
International Conference On Machine Learning

Han Zhong*[1], Yutong Yin*[2], Shenao Zhang*[2], Xiaojun Xu*[3], Yuanxin Liu*[2], Yifei Zuo*[2], Zhihan Liu*[2], Boyi Liu[3], Sirui Zheng[2], Hongyi Guo[2], Liwei Wang[1], Mingyi Hong[4], Zhaoran Wang[2]

[1]Peking University, [2]Northwestern University, [3]ByteDance, [4]University of Minnesota

Reasoning as a Graphical Model

```
Q: What is reasoning in large language models?
A: Okay, so I need to figure out what reasoning
   is in large language models (LLMs) is. Let
   me start by breaking down the question. The
   user is asking about reasoning ...
Reasoning in large language models (LLMs) refers
to their ability to generate responses that mimic
structuredlogical thought processes to solve
problems or answer questions.
```



$$\mathbb{P}(y, o \mid x, \theta) = \mathbb{P}(z \mid x, \theta) \cdot \mathbb{P}(y \mid x, z, \theta) \cdot \mathbb{P}(o \mid x, z, \theta)$$

Bootstrap Reinforced Thinking Process

$$\begin{aligned}\mathscr{L}(\theta) &= \log \sum_{(z,y,o)\in\mathscr{Z}\times\mathscr{Y}\times\mathscr{O}} \mathbb{P}(z, y, o \mid x, \theta) \\ &= \max_{\mathbb{Q}\in\Delta}\Bigg\{ \underbrace{\sum_{(z,y,o)} \log \mathbb{P}(z, y, o \mid x, \theta)\mathbb{Q}(z, y, o \mid x, \psi) - \sum_{(z,y,o)} \log \mathbb{Q}(z, y, o \mid x, \psi)\mathbb{Q}(z, y, o \mid x, \psi)}_{:=\mathscr{L}_\psi(\theta)} \Bigg\}\end{aligned}$$

Maximize $\mathscr{L}(\theta)$ (difficult) $\Longrightarrow$ Maximize evidence lower bound $\mathscr{L}_\psi(\theta)$ (easy)

BRiTE — An EM-type Algorithm

E

$$\mathbb{Q}(z, y, o \mid x, \psi_{t+1}) \leftarrow \mathrm{argmax}_{\mathbb{Q}} \mathscr{L}_\psi(\theta_t) = \frac{\mathbb{P}(z, y, o \mid x, \theta_t)}{\sum_{(z,y,o)} \mathbb{P}(z, y, o \mid x, \theta_t)}$$

Thought proposer

M

$$\begin{aligned}\theta_{t+1} &= \mathrm{argmax}_\theta \mathscr{L}_{\psi_{t+1}}(\theta) \\ &= \mathrm{argmax}_\theta \Bigg\{ \sum_{z,y,o\in\mathscr{Z}\times\mathscr{Y}\times\mathscr{O}} \log \mathbb{P}(z, y \mid x, \theta) \cdot \mathbb{Q}(z, y, o \mid x, \psi_{t+1}) \Bigg\}\end{aligned}$$

Assumptions: $f_\theta \in \mathscr{H}$ for a certain RKHS; $\mathbb{P}(z, y \mid x, \theta) \propto \exp(f_\theta(x, z, y))$

Theorem: convergence to optima

$$\min_{1\le t\le T} \left\{ \log \frac{\mathbb{P}(x \in \mathscr{X}, y \in \mathscr{Y}, o \in \mathscr{O} \mid x, \theta^*)}{\mathbb{P}(x \in \mathscr{X}, y \in \mathscr{Y}, o \in \mathscr{O} \mid x, \theta_t)} \right\} \le \frac{\mathbb{D}_{\mathrm{KL}}\big(\mathbb{P}(\cdot \mid x, \theta_1) \| \mathbb{P}(\cdot \mid x, \theta^*)\big)}{T}$$

Concrete Examples of BRiTE

Scope:

- $o \in \{0,1\}$, $\mathscr{O} = \{1\}$
- $\mathscr{Y}$ is the response space
- $\mathscr{Z}$ is the latent space

✦ $\mathbb{P}(o = 1 \mid x, z, y) := \exp(R(x, z, y)/\beta)$

✦ $\mathbb{P}(z, y, o = 1 \mid x, \theta) = \mathbb{P}(z, y \mid x, \theta)\mathbb{P}(o = 1 \mid x, z, y)$

✦ $\mathbb{Q}(z, y \mid x, \psi) := \mathbb{Q}(z, y, o = 1 \mid x, \psi)$

Example (PPO)

$$\begin{aligned}\mathscr{L}_\psi(\theta) &= \sum_{(z,y)} \log \mathbb{P}(z, y, o = 1 \mid x, \theta)\mathbb{Q}(z, y \mid x, \psi) \\ &\quad - \sum_{(z,y)} \log \mathbb{Q}(z, y \mid x, \psi)\mathbb{Q}(z, y \mid x, \psi) \\ &= \mathbb{E}_{(z,y)\sim\mathbb{Q}}\left[R(x, z, y)/\beta - \log \frac{\mathbb{Q}(z, y \mid x, \psi)}{\mathbb{P}(z, y \mid x, \theta)}\right]\end{aligned}$$

Scope:

- $o \in \{0,1\}$, $\mathscr{O} = \{1\}$
- $\mathscr{Y}$ is the response space
- $\mathscr{Z}$ is the latent space

✦ $\mathbb{P}(o = 1 \mid x, z, y) := \mathbb{I}(y \text{ is correct for } x)$ or $\exp(R(x, y)/\beta)$

✦ $\mathbb{P}(z, y, o = 1 \mid x, \theta) = \mathbb{P}(z, y \mid x, \theta)\mathbb{P}(o = 1 \mid x, z, y)$

✦ $\mathbb{Q}(z, y \mid x, \psi) := \mathbb{Q}(z, y, o = 1 \mid x, \psi)$

$$\max_{\mathbb{P}} \left\{ \mathbb{E}_{(z,y)\sim\mathbb{P}(\cdot,\cdot\mid x,\theta_t)}\Big[\log \mathbb{P}(z, y \mid x, \theta) \cdot \mathbb{I}(y \text{ is correct for } x)\Big] \right\}$$

$$\max_{\mathbb{P}} \left\{ \mathbb{E}_{(z,y)\sim\mathbb{P}(\cdot,\cdot\mid x,\theta_t)}\Big[\log \mathbb{P}(z, y \mid x, \theta) \cdot \exp\big(R(x, y)/\beta\big)\Big] \right\}$$

If $\mathscr{Z} = \varnothing$, then it recovers STaR and Reject Sampling Finetuning or RestEM

Learning Intractable Posterior via RL

$$\mathbb{Q}(z, y, o \mid x, \psi) \leftarrow \mathrm{argmax}_{\mathbb{Q}} \mathscr{L}_\psi(\theta) = \frac{\mathbb{P}(z, y, o \mid x, \theta)}{\sum_{(z,y,o)} \mathbb{P}(z, y, o \mid x, \theta)}$$

Intractable

Lemma: the optimal policy for an entropy-regularized token-level MDP

$$\pi^\star(a_h \cup \{(s_i, a_i)\}_{i=h+1}^H \mid s_h) \propto \exp\left(\frac{1}{\beta}\sum_{i=h}^H r(s_i, a_i)\right)$$

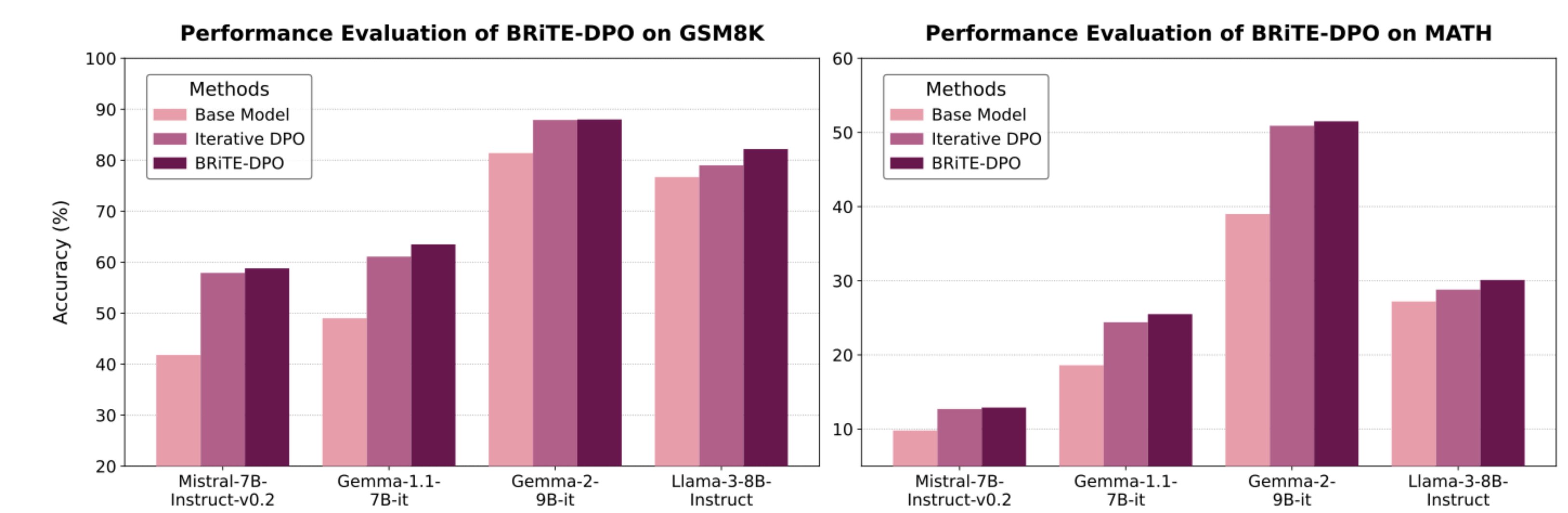
Set $\frac{1}{\beta}\sum_{i=0}^{H} r(s_i, a_i) = \log \mathbb{P}(z, y, o \mid x, \theta)$! Then $\pi^\star(s_H \mid s_0) = \mathbb{Q}(z, y, o \mid x, \psi)$

Experiments

1. BRiTE Significantly Improves Existing Rejection Sampling Algorithms.
2. BRiTE ≥ SFT with Human-Annotated Thinking Process.

Algorithm	Mistral-7B-Instruct-v0.2		Gemma-1.1-7B-it		Gemma-2-9B-it		Llama-3-8B-Instruct	
	GSM8K	MATH	GSM8K	MATH	GSM8K	MATH	GSM8K	MATH
—	41.8	9.8	49.0	18.8	81.3	37.3	79.2	28.3
SFT	**52.8**	**13.6**	57.5	19.6	80.1	41.5	72.6	27.1
RS	47.7	10.3	58.4	18.7	87.6	47.5	79.5	28.9
BRiTE	52.2	11.2	**59.2**	**23.7**	**89.7**	**50.5**	**81.0**	**30.0**

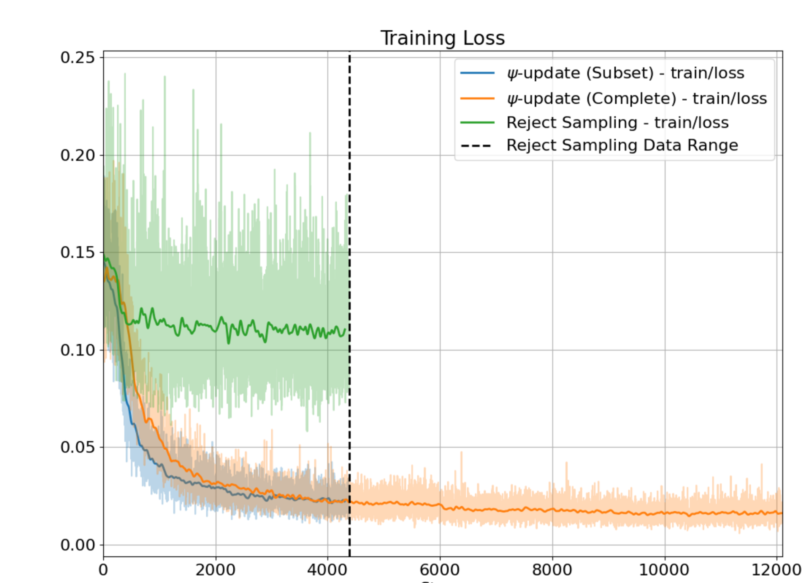
3. BRiTE Enhances the Reasoning Capacity in RLHF Stage.



4. BRiTE Improves Code Generation Ability.

Algorithm	HumanEval		BCB (Instruct)	
	Basic (%)	Plus (%)	Hard (%)	Full (%)
—	78.0	70.7	10.1	35.5
SFT	78.0	67.7	11.5	**37.2**
RS	79.3	**73.2**	11.5	35.6
BRiTE	**81.7**	**72.6**	**15.5**	36.3

5. BRiTE Generates High Quality Trajectories for Distillation.



Scan to read our paper!