

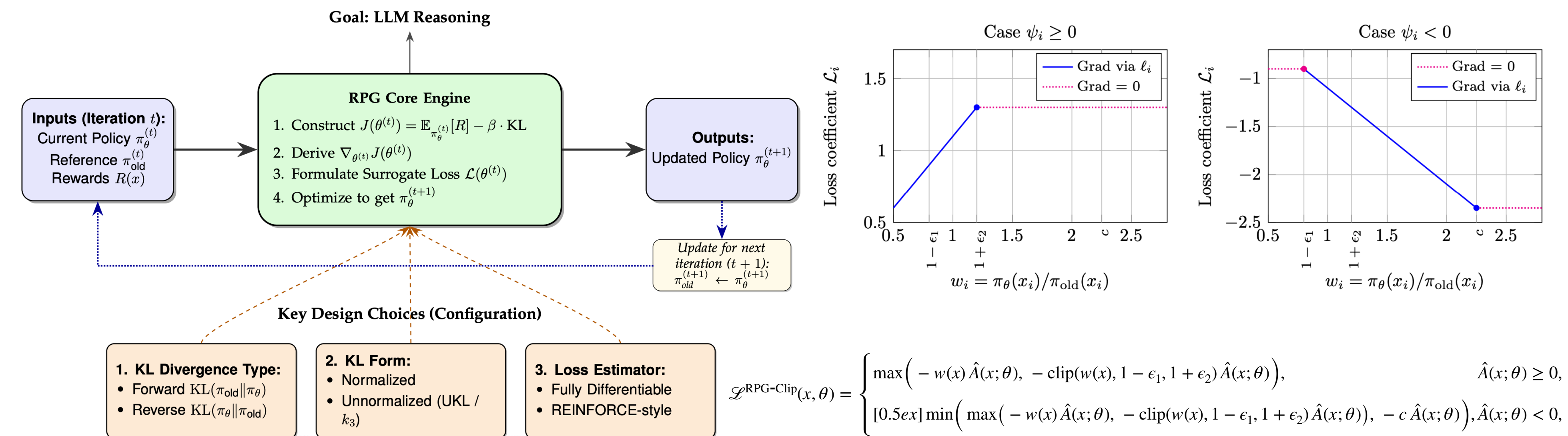
On the Design of KL-Regularized Policy Gradient Algorithms for LLM Reasoning

Yifan Zhang*, Yifeng Liu*, Huizhuo Yuan, Yang Yuan, Quanquan Gu†, Andrew C Yao†

*: equal contribution

<https://github.com/complex-reasoning/RPG>

RPG Framework



Regularized Policy Gradients

Regularization	Normalized ($\mathbb{E}$ w.r.t. π_{old})	Unnormalized ($\mathbb{E}$ w.r.t. $\tilde{\pi}_{\text{old}}$)
Forward KL	$\mathbb{E}[-w(x)R(x) - \beta \log \pi_{\theta}(x)]$	$Z_{\text{old}} \mathbb{E}[-w(x)R(x) + \beta(w(x) - \log w(x) - 1)]$
Reverse KL	$\mathbb{E}[w(x)(-R(x) + \beta \log w(x))]$	$Z_{\text{old}} \mathbb{E}[-w(x)R(x) + \beta(w(x) \log w(x) - w(x) + 1)]$

REINFORCE-style Regularized Policy Gradients

Regularization	Normalized ($\pi_{\text{sampling}} = \pi_{\text{old}}$)	Unnormalized ($\pi_{\text{sampling}} = \tilde{\pi}_{\text{old}}$)
Forward KL	$-\mathbb{E}[\text{SG}(w(x)R(x) + \beta) \log \pi_{\theta}(x)]$	$-\mathbb{E}[\text{SG}(Z_{\text{old}}(w(x)R(x) - \beta(w(x) - 1))) \log \pi_{\theta}(x)]$
Reverse KL	$-\mathbb{E}[\text{SG}(w(x)(R(x) - \beta \log w(x) - \beta)) \log \pi_{\theta}(x)]$	$-\mathbb{E}[\text{SG}(Z_{\text{old}}w(x)(R(x) - \beta \log w(x))) \log \pi_{\theta}(x)]$

$w(x) = \pi_{\theta}(x)/\pi_{\text{old}}(x)$, SG is stop-gradient

Experiments

4K context length

Method	AIME24		AIME25	
	Last	Best	Last	Best
GRPO	0.3458	0.3677	0.2896	0.3042
DAPO	0.4063	0.4479	0.3510	0.3938
RPG-UFKL	0.4031	0.4396	0.3625	0.3979
RPG-URKL	0.3990	0.4219	0.3438	0.3792
RPG-REINFORCE-UFKL	0.4281	0.4375	0.3771	0.4042
RPG-REINFORCE-URKL	0.4458	0.4531	0.4125	0.4313

2K context length

Method	AIME24		AIME25	
	Last	Best	Last	Best
GRPO	0.2563	0.2708	0.2323	0.2479
DAPO	0.3229	0.3281	0.2792	0.2844
RPG-UFKL	0.3427	0.3479	0.2833	0.2833
RPG-URKL	0.3260	0.3594	0.2677	0.2677
RPG-REINFORCE-UFKL	0.3396	0.3625	0.2927	0.3083
RPG-REINFORCE-URKL	0.3188	0.3417	0.2792	0.2938

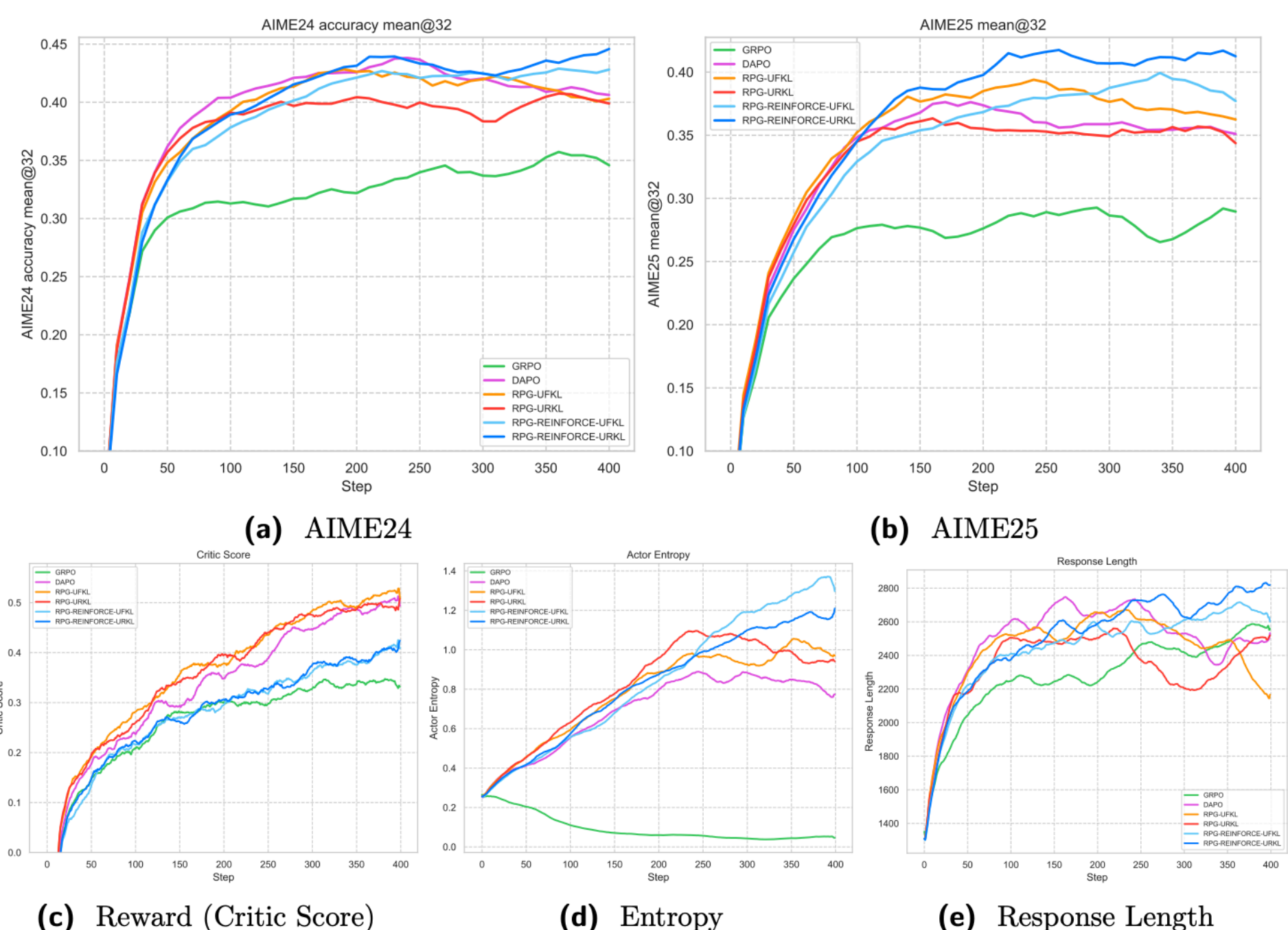


Figure 3 Performance of RPG and REINFORCE-Style Regularized Policy Gradient (RPG-REINFORCE) methods compared to baselines with 4k context length.