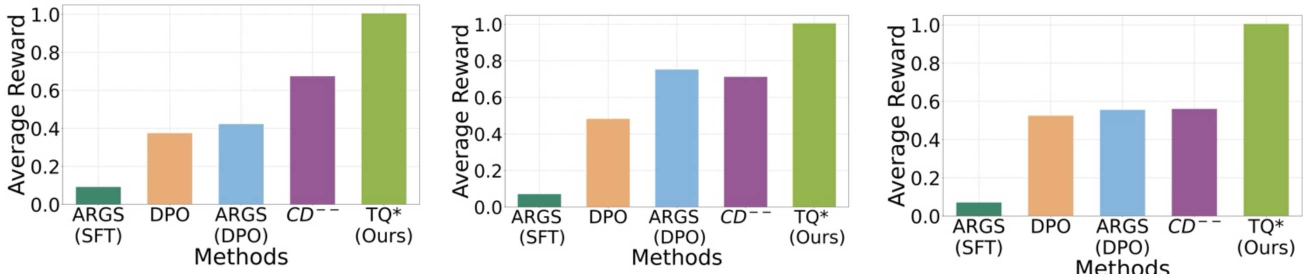
(1) Effect of Varying Output Length



(a) Output length=64

(b) Output length=128

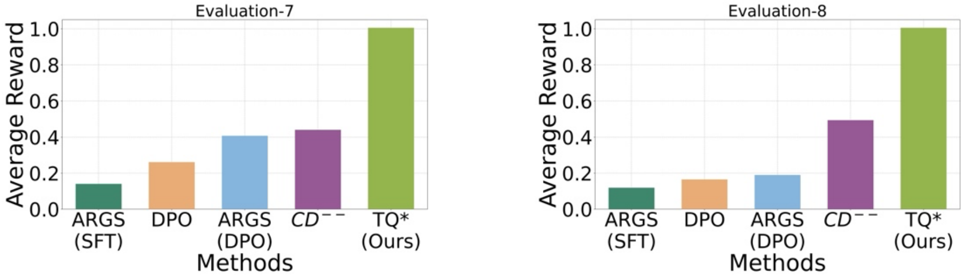
(c) Output length=512

Figure 1: We compare **TQ*** against all baselines varying the length of generated text. We observe that **TQ*** consistently outperforms all the compared baselines. The result is on the setup Evaluation-1.

(2) Additional Evaluations on Larger Models (12B and 13B parameters)

Table 1: Summary of the datasets and model architectures used for new experimental evaluations for Figure 2.

	Dataset	Model Architectures			Reward Preference
		SFT	DPO	Reward	
Evaluation-7	HH-RLHF [1]	Llama-2-13B [3]	Llama-2-13B [3]	Llama-2-13B [3]	Helpful and Harmless responses.
Evaluation-8	OpenAssistant Conversations Dataset	Pythia-12B [2]	Pythia-12B [2]	Pythia-6.9B [2]	Helpful Conversations.



(a) 13 B parameters

(b) 12 B parameters

Figure 2: In (a) and (b), we report the normalized average reward obtained by different decoding methods on setups Evaluation 7 and Evaluation 8 as described in Table 1 above, respectively. Consistent with our findings, our proposed **TQ*** significantly outperforms all the competitive baselines.

(3) Pareto Front Plot

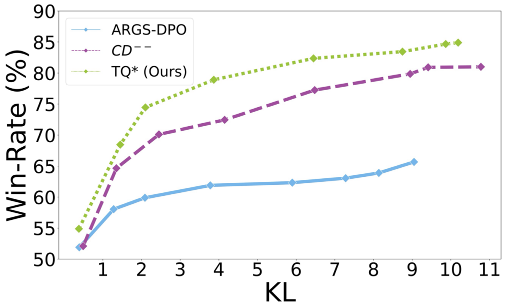


Figure 3: This figure compares the tradeoff between the win-rate and the KL divergence to the base reference SFT policy. Our proposed method **TQ*** performs better as compared to existing baselines.