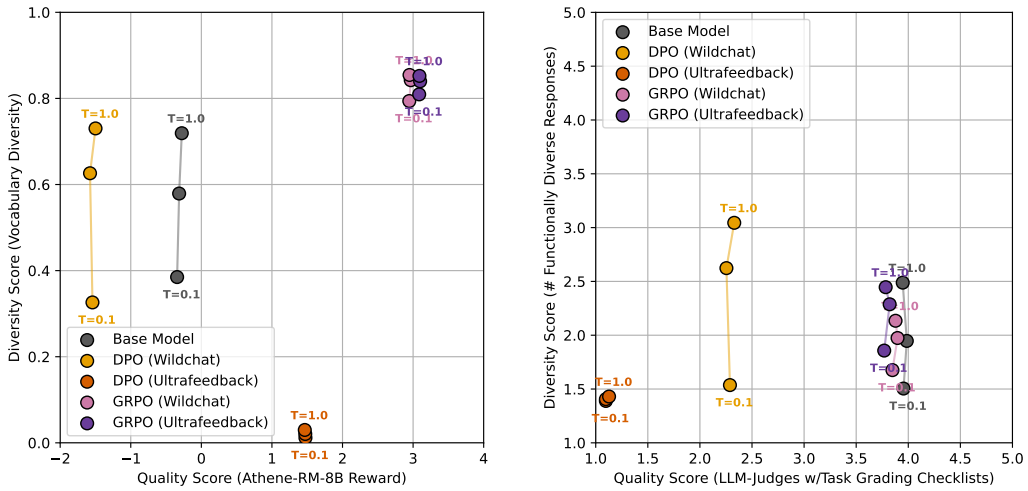
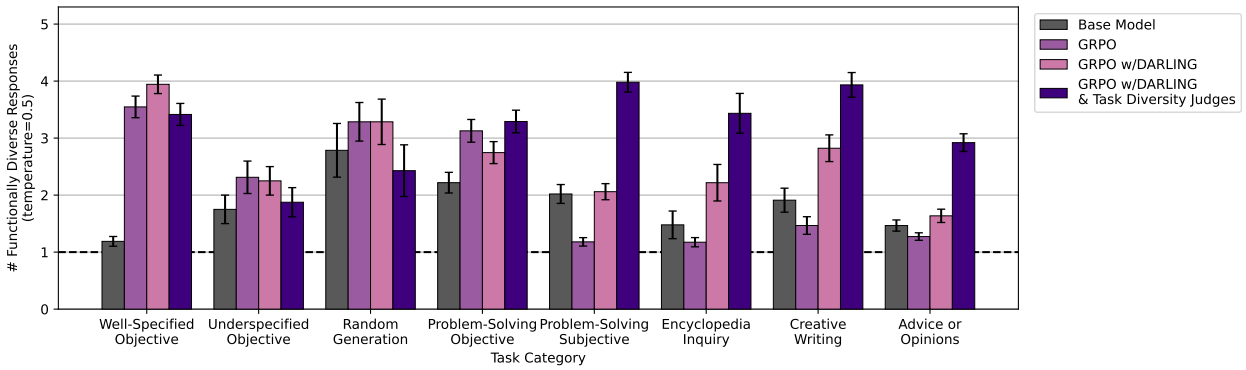
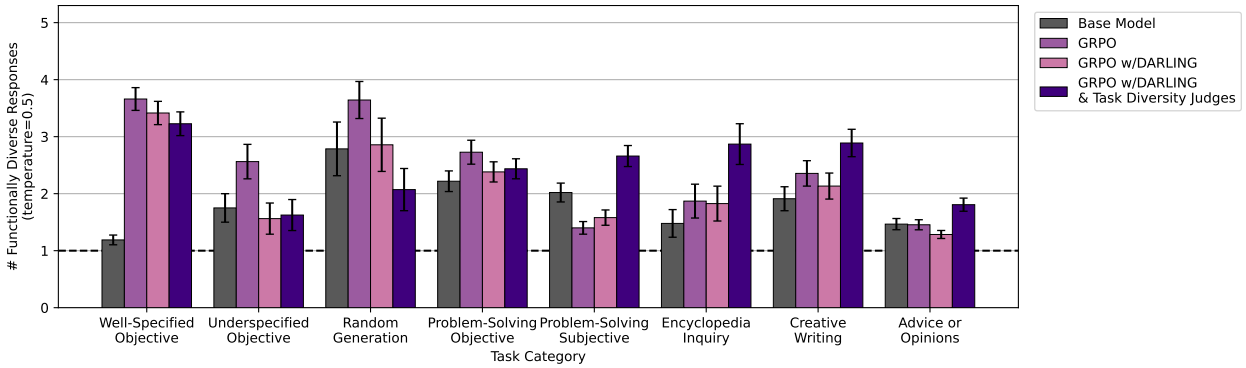


(a) General Metrics **(b)** Task-Based Metrics

Figure 17 Diversity-quality tradeoff under general vs task-based metrics for Llama-3.1-8B-Instruct, with and without preference alignment. DPO and GRPO results based on 1000 training steps and $\beta = 0.01$ and $\beta = 0.001$, respectively. While DPO and GRPO generally improve reward quality, they do not always improve checklist-based quality. Metrics avg. across all task categories except category A, where homogenization is desired.

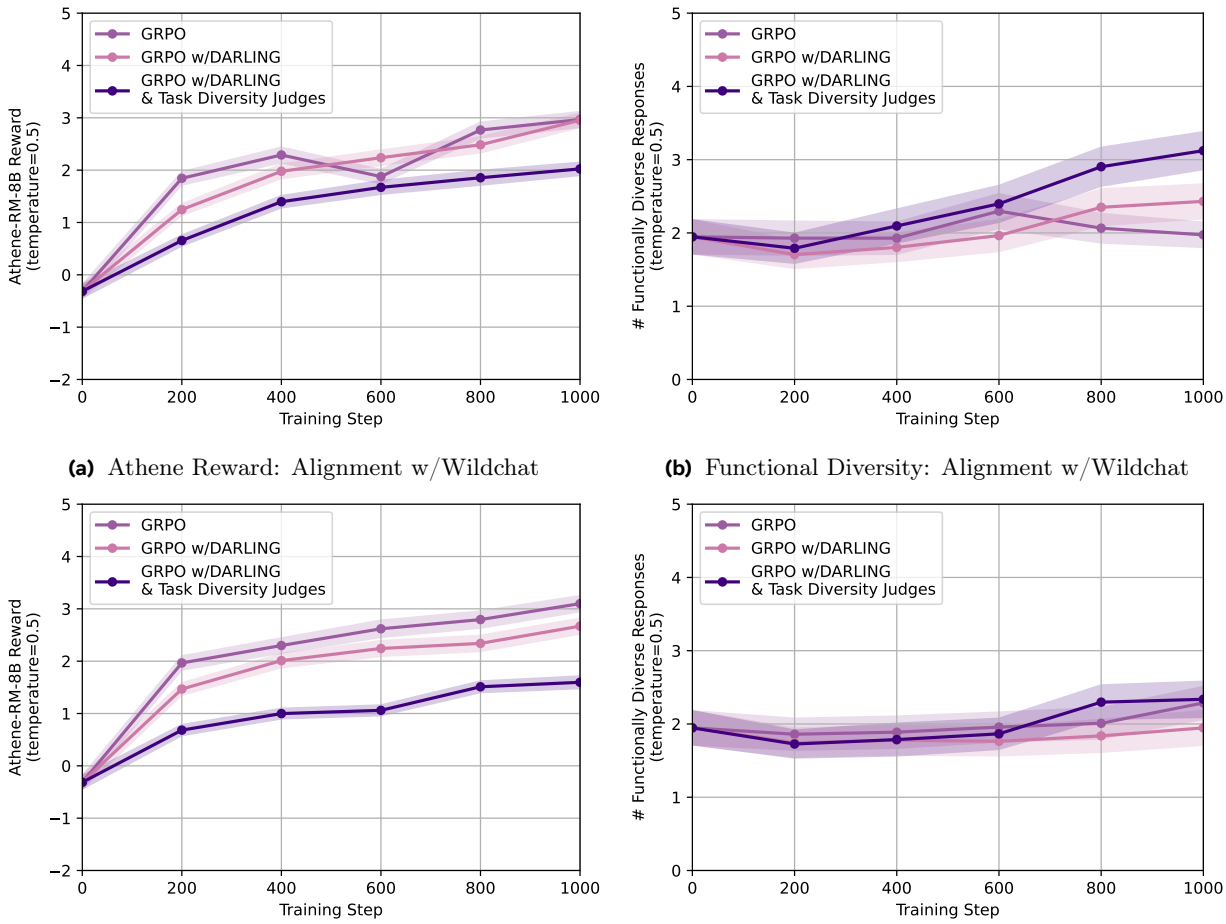


(a) Preference alignment using Wildchat prompts.



(b) Preference alignment using Ultrafeedback prompts.

Figure 18 Number of functionally diverse responses generated by Llama-3.1-8B-Instruct, after preference alignment with DARLING (Li et al., 2025b). GRPO and DARLING results with $\beta = 0.001$. DARLING with task diversity judges uses GPT-4o as the task-dependent functional diversity judge. DARLING generally maintains or improves functional diversity over GRPO, and task diversity judges generally provide further improvement. All alignment methods undesirably reduce homogenization for category A (Well-Specified Objective).



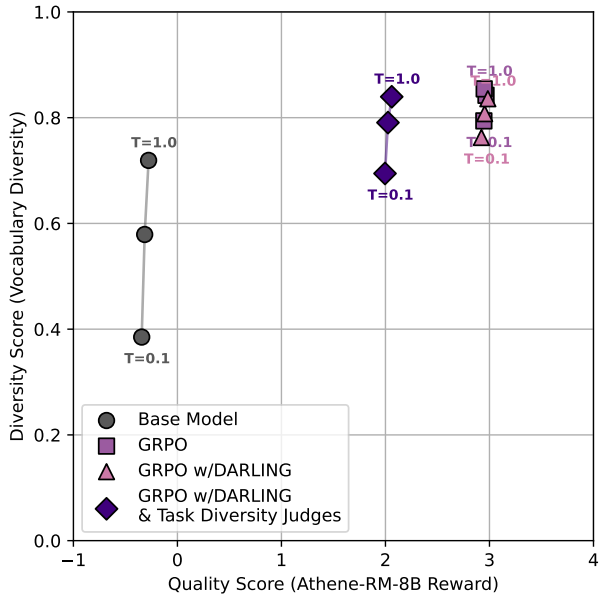
(a) Athene Reward: Alignment w/Wildchat

(b) Functional Diversity: Alignment w/Wildchat

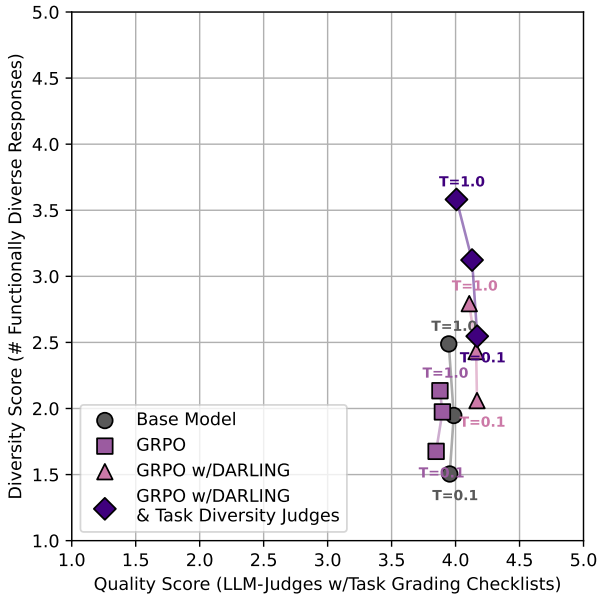
(c) Athene Reward: Alignment w/Ultrafeedback

(d) Functional Diversity: Alignment w/Ultrafeedback

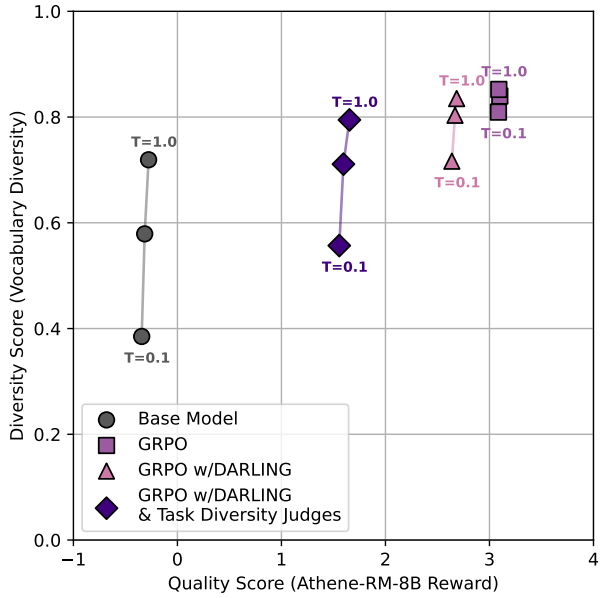
Figure 19 During alignment of Llama-3.1-8B-Instruct with DARLING using **Wildchat prompts**, both the reward and functional diversity generally increase. GRPO and DARLING use $\beta = 0.001$. Metrics avg. across all task categories except category A, where homogenization is desired.



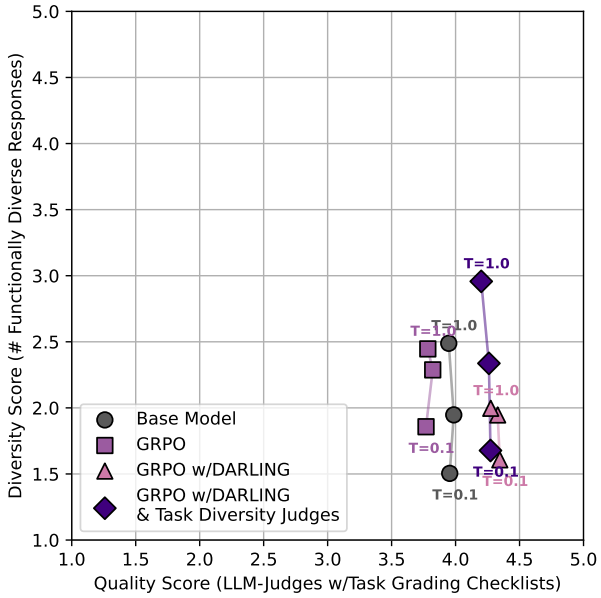
(a) General Metrics: Alignment w/Wildchat



(b) Task-Based Metrics: Alignment w/Wildchat



(c) General Metrics: Alignment w/Ultrafeedback



(d) Task-Based Metrics: Alignment w/Ultrafeedback

Figure 20 Diversity-quality tradeoff under general vs task-based metrics for Llama-3.1-8B-Instruct, after preference alignment with DARLING (Li et al., 2025b). GRPO and DARLING results based on 1000 training steps and $\beta = 0.001$. While general metrics do not show improvements, task-based metrics show that DARLING improves both diversity and quality compared to GRPO.