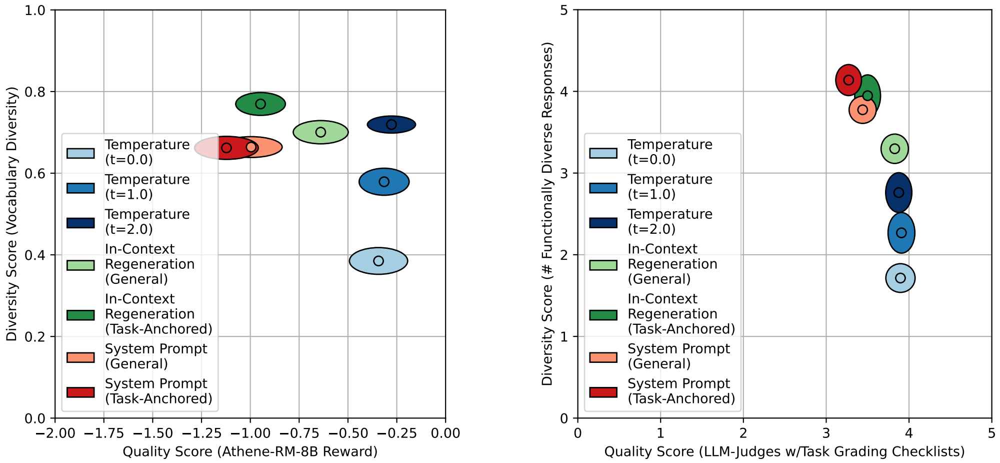
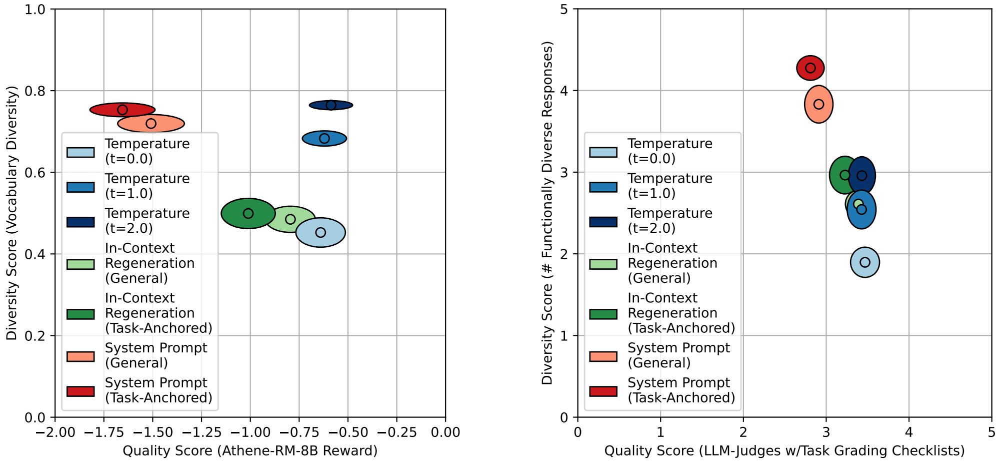


(a) General Metrics

(b) Task-Based Metrics

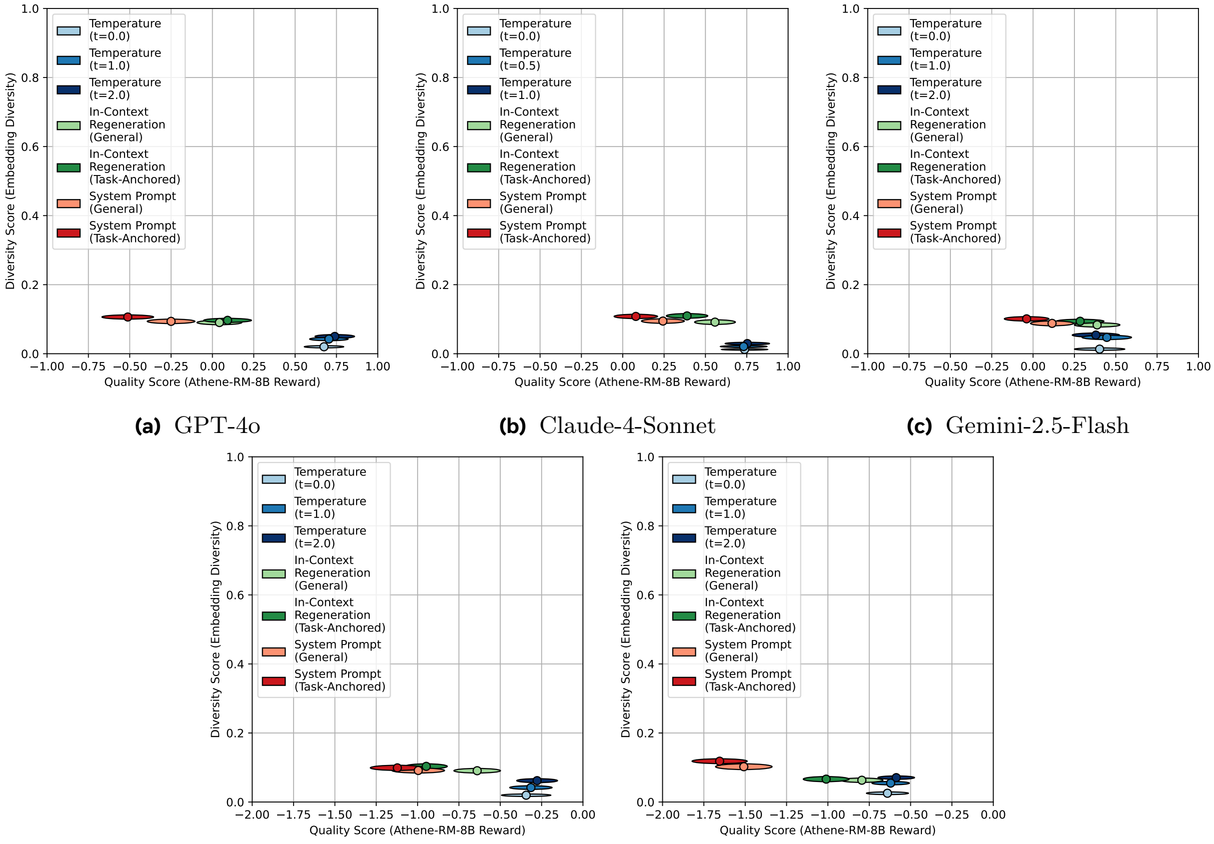
Figure 11 Diversity-quality tradeoff under general vs task-based metrics for **Llama-3.1-8B-Instruct**.



(a) General Metrics

(b) Task-Based Metrics

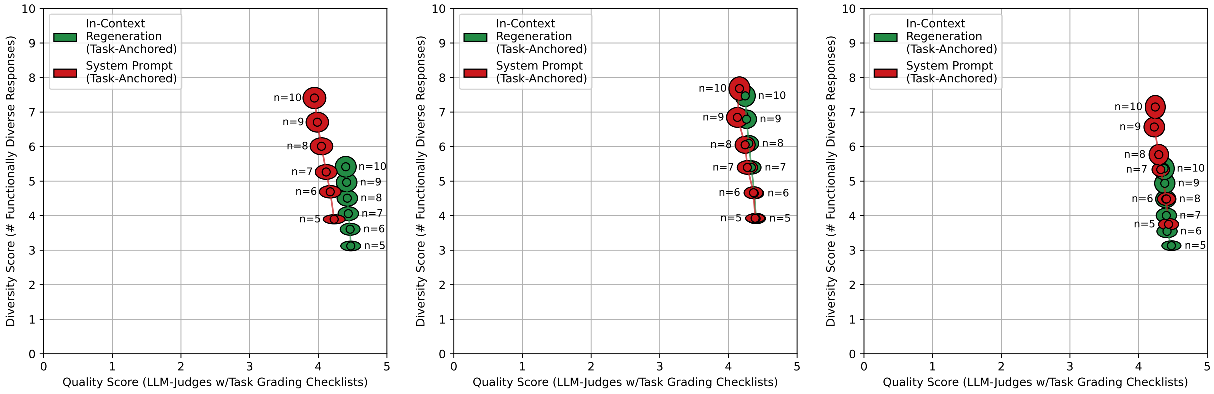
Figure 12 Diversity-quality tradeoff under general vs task-based metrics for **Mistral-7B-Instruct-v0.3**.



(a) GPT-4o **(b)** Claude-4-Sonnet **(c)** Gemini-2.5-Flash

(d) Llama-3.1-8B-Instruct **(e)** Mistral-7B-Instruct-v0.3

Figure 13 Diversity-quality tradeoff using embedding diversity.



(a) GPT-4o **(b)** Claude-4-Sonnet **(c)** Gemini-2.5-Flash

Figure 14 Diversity-quality tradeoff for varying number of generated responses ($n = 5$ to $n = 10$). Judge metrics based on GPT-4o only. The number of functionally diverse responses consistently increases with more generated responses. However, there appear to be small (statistically insignificant) decreases in checklist-based quality. The quality decrease is larger for system prompt sampling, possibly due to $n = 10$ approaching the max output length for a single generation.

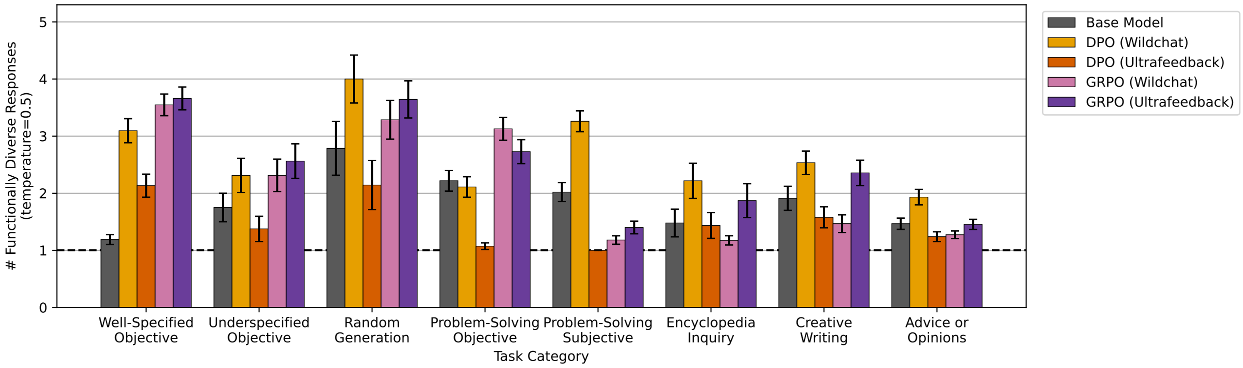
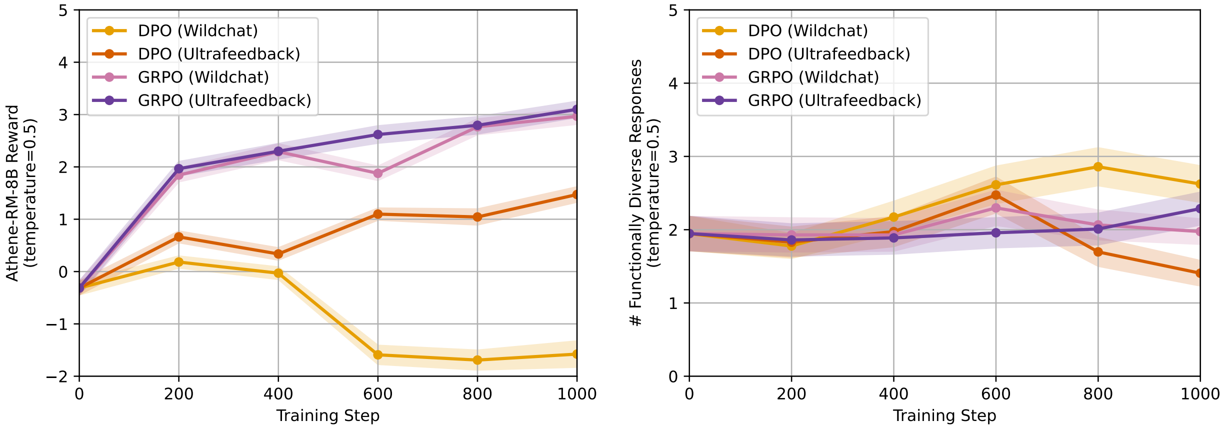


Figure 15 Number of functionally diverse responses generated by Llama-3.1-8B-Instruct, with and without preference alignment. DPO and GRPO results based on 1000 training steps and $\beta = 0.01$ and $\beta = 0.001$, respectively. Unlike prior results on token entropy (Lanchantin et al., 2025b), functional diversity does not collapse and sometimes increases post-alignment.



(a) Athene Reward by Training Step

(b) Functional Diversity by Training Step

Figure 16 During alignment of Llama-3.1-8B-Instruct, the reward generally increases without a collapse in functional diversity. DPO and GRPO results based on $\beta = 0.01$ and $\beta = 0.001$, respectively. Metrics avg. across all task categories except category A, where homogenization is desired.