A.7 Alignment Experiments

We run exploratory alignment experiments to further show the usefulness of our task-dependent framework for evaluating and reducing homogenization. All alignment experiments use *Llama-3.1-8B-Instruct* and the *Athene-RM-8B* reward model, and LLM-judge metrics are only calculated with GPT-4o. First, we demonstrate that functional diversity does not collapse and sometimes increases after preference tuning with Direct Preference Optimization (DPO) (Rafailov et al., 2023) and Group Relative Policy Optimization (GRPO) (Shao et al., 2024) (Figures 15-17). Second, we show how task-dependent evaluation clarifies the impact of diversity-promoting methods in alignment by evaluating DARLING (Diversity-Aware Reinforcement Learning) (Li et al., 2025b) (Figures 18-20). We further explore how DARLING could be modified to account for task-dependence.

DPO and GRPO We run online DPO and GRPO following the training recipe for non-verifiable rewards in Lanchantin et al. (2025b) and Li et al. (2025b), respectively. In particular, we use a learning rate of $1e{-}6$, batch size of 32, and train for 1000 steps. At each step, we generate 8 responses per prompt with temperature 1.0 and 1024 max tokens. For DPO, preference pairs are constructed based on the responses with the maximum and minimum reward. For GRPO, all 8 responses are used to calculate advantage. We explore two values of β: 0.01 and 0.1 for DPO, and 0.001 and 0.01 for GRPO.

We also explore using two datasets for preference tuning: Wildchat (Zhao et al., 2024) and Ultrafeedback (Cui et al., 2023). For Wildchat, we use 10,000 randomly sampled prompts. The majority of these prompts (5,535) corresponded to category E (Creative Writing), based on task classification by GPT-4o. Thus, for Ultrafeedback, we try a stratified random sample of 10,000 prompts based on each task category in our taxonomy and GPT-4o as the task classification judge. Specifically, we sample 2,500 prompts for each task category, excluding category E (Problem-Solving Subjective) and combining categories B and C (Underspecified Objective and Random Generation) due to prompt availability. In both datasets, we exclude prompts with more than 512 tokens.

DARLING Li et al. (2025b) propose DARLING (Diversity-Aware Reinforcement Learning), which modifies the GRPO reward to jointly reinforce diversity and quality. Specifically, they scale reward by the diversity $d(y_i|y_1, \ldots, y_n)$ of a generation y_i, which they define as the average pairwise distance between y_i and all other generations y_j $(j \neq i)$, normalized to be between 0 and 1. They implement their method using "semantic uniqueness" as their distance metric, which represents a general notion of functional diversity (not task-dependent). They fine-tune a *ModernBERT-base* model to predict semantic uniqueness based on 1000 human annotations from NoveltyBench (Zhang et al., 2025b).

We explore modifying DARLING to account for task-dependence by using GPT-4o as a task-dependent functional diversity judge, in place of the fine-tuned ModernBert classifier. For prompts in category A (Well-Specified Objective), we also modify DARLING to scale the reward by $1 - d(y_i|y_1, \ldots, y_n)$, which promotes homogenization instead of diversity for category A.

B Additional Experiment Results

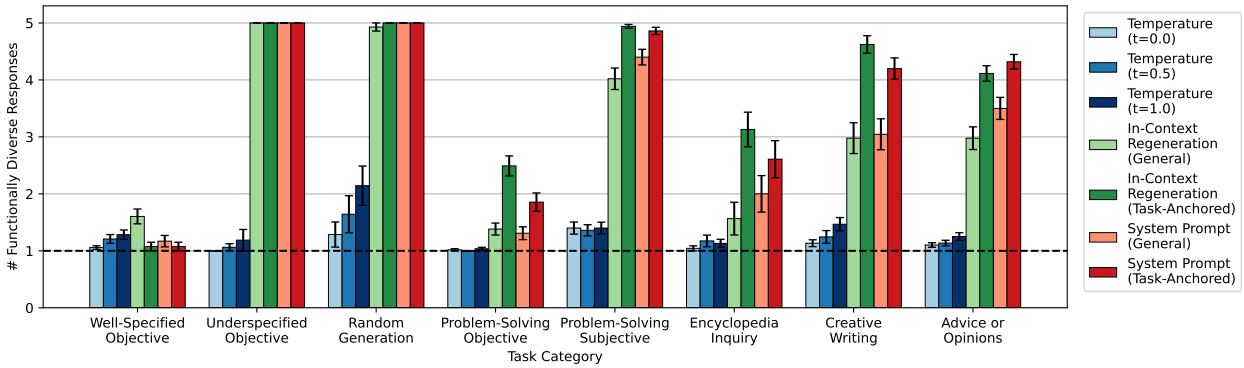


Figure 5 Number of functionally diverse responses generated by **Claude-4-Sonnet** (c.f. Figure 2).

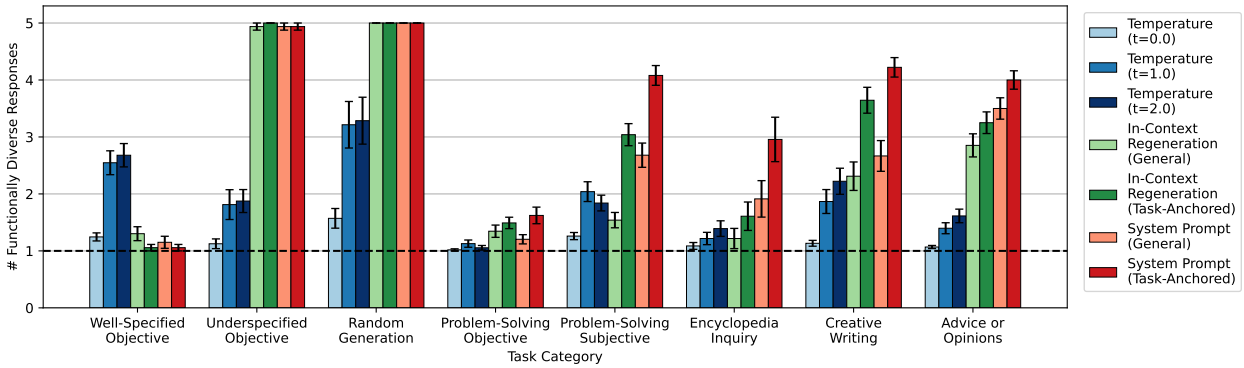


Figure 6 Number of functionally diverse responses generated by **Gemini-2.5-Flash** (c.f. Figure 2).

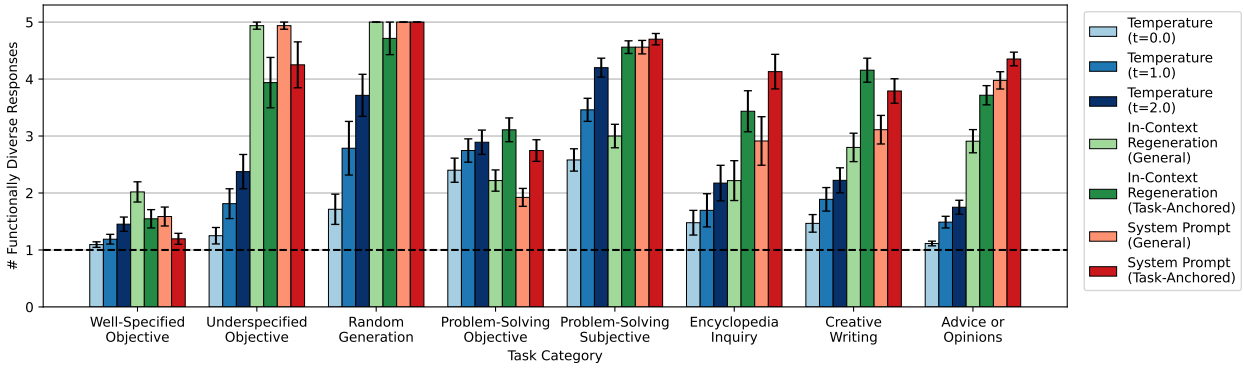


Figure 7 Number of functionally diverse responses generated by **Llama-3.1-8B-Instruct** (c.f. Figure 2).

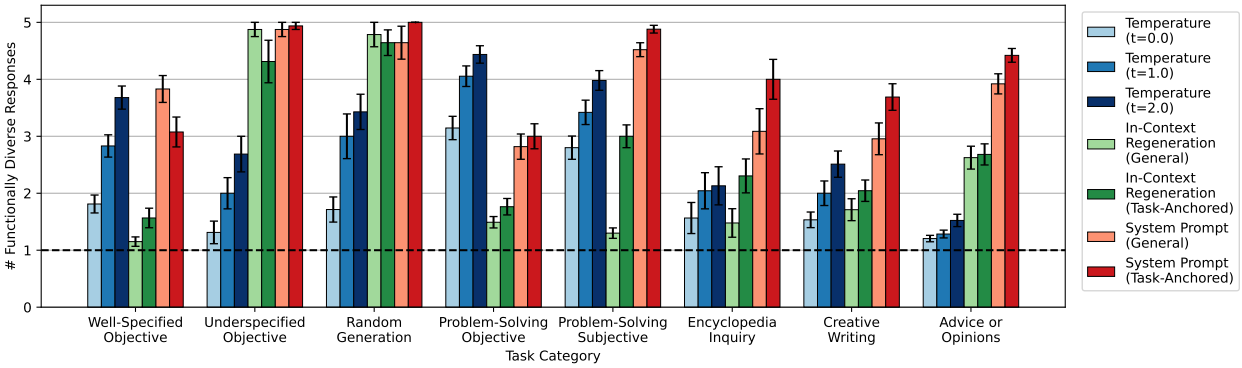
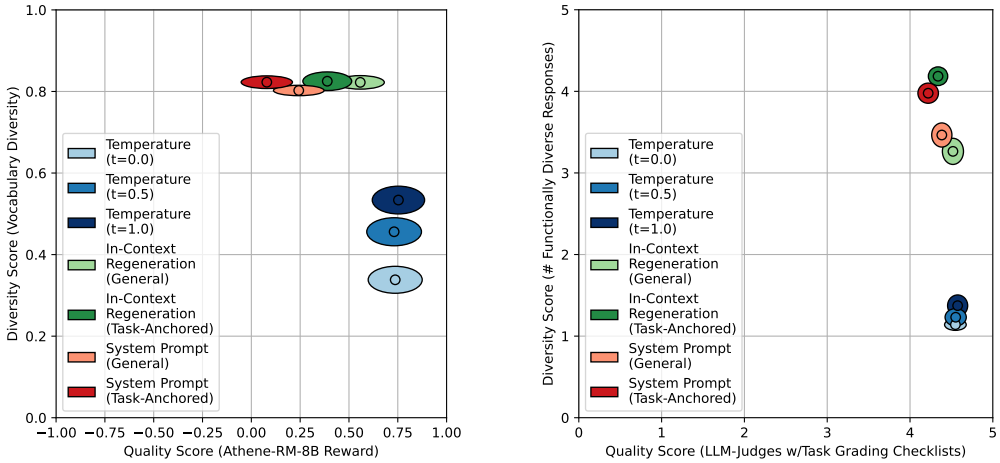
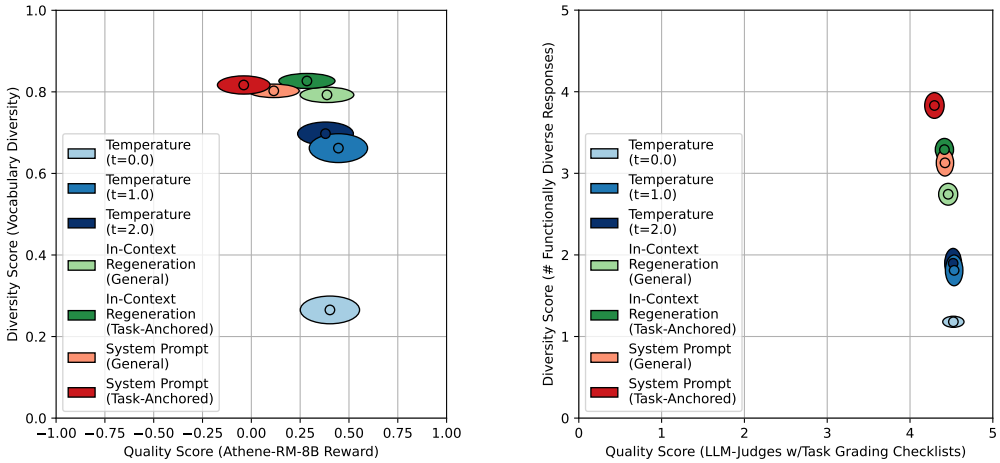


Figure 8 Number of functionally diverse responses generated by **Mistral-7B-Instruct-v0.3** (c.f. Figure 2).



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

Figure 9 Diversity-quality tradeoff under general vs task-based metrics for **Claude-4-Sonnet**.



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

Figure 10 Diversity-quality tradeoff under general vs task-based metrics for **Gemini-2.5-Flash**.