Datasets We evaluate $n = 344$ prompts from a variety of datasets that achieve reasonable coverage across the task categories in our taxonomy (c.f. Appendix Table 6). The 6 datasets used in our evaluation are: *Community Alignment* (Zhang et al., 2025a), *MacGyver* Tian et al. (2024), *MATH-500* (Lightman et al., 2023), *NoveltyBench* (Zhang et al., 2025b), *SimpleQA* (Wei et al., 2024), and *WildBench* (Lin et al., 2025). These datasets represent a mix of user-generated and curated prompts. Appendix A.2 includes more details about each dataset and how we sampled prompts.

Sampling Strategies To evaluate homogenization over multiple responses, we compare three sampling strategies: temperature sampling, system prompt sampling, and in-context regeneration. For each sampling strategy, we sample 5 responses per prompt. For *temperature sampling*, we consider three temperature levels for each model based on its permitted range: low ($t = 0.0$ for GPT, Claude, and Gemini, $t = 0.1$ for Llama and Mistral), medium ($t = 1.0$ for GPT and Gemini, $t = 0.5$ for Claude, Llama, and Mistral), and high ($t = 2.0$ for GPT and Gemini, $t = 1.0$ for Claude, Llama and Mistral). We further evaluate both general and task-anchored approaches to system prompt sampling and in-context regeneration (§ 3.3). For the general approach, we use a variation of the prompts used previous works (Zhang et al., 2025a,b). Our task-anchored approach modifies these prompts to specify the functional difference relevant to each task category in our taxonomy, based on the model's self-categorization of the task. Appendix A.4 provides all our task-anchored prompts for system prompt sampling and in-context regeneration. For both system prompt sampling and in-context regeneration, we use the medium temperature values.

Diversity Metrics We compute four diversity metrics: task-anchored functional diversity (Def. 3.1), vocabulary diversity (Def. A.1), embedding diversity (Def. A.2), and compression diversity (Def. A.3). To calculate *functional diversity*, we use LLM-judges[2], where the judge prompt includes the functional diversity concept for the ground-truth task category (see Appendix A.5 for judge prompts). We determine the ground-truth task category for each prompt based on the source dataset and human annotation (Appendix A.3). We then determine the *number of functionally diverse responses* (Def 3.2) out of the 5 responses[3] generated per prompt and sampling strategy. To compute embedding diversity, we generate response embeddings using the *gemini-embedding-001* model. To compute compression diversity, we use Gzip following the method in Shaib et al. (2024).

Quality Metrics We evaluate the quality of responses in two ways. (1) *Reward Model Quality:* Following many recent work that evaluates the diversity-quality trade-off (Lanchantin et al., 2025a; Slocum et al., 2025), we measure quality in terms of reward scores assigned by a reward model. We use *Athene-RM-8B*, which is to date empirically validated as one of the best reward models for human preferences (Frick et al., 2025). (2) *Checklist-Based Quality:* We also measure quality following prior work that uses *LLM-judges with grading checklists* (Lin et al., 2025; Wei et al., 2025). In this approach, the LLM-judge first generates a checklist of 3-5 key factors for response quality in a given prompt. This prompt-specific checklist is then used by the LLM-judge to score a particular response on a Likert scale from 1 to 5, where 1 indicates that none of the checklist criteria are met and 5 indicates that all criteria are satisfied. We manually review all generated checklists, and include the judge prompts and examples of generated checklists in Appendix A.6.

4.2 Functional Diversity

We first report our evaluation results on functional diversity. Our main finding is that our task-anchored sampling technique outperforms the more general sampling techniques in previous work (Zhang et al., 2025a,b). Figure 2 shows how we significantly increase functional diversity for task categories where homogenization is undesired, while preserving homogenization where it is desired (for GPT-4o, all results in Appendix B). Below, we explore results across task categories.

[2] We validate the LLM-judges on a stratified random sample of 225 response pairs across models, task categories, and sampling strategies. Two authors independently labeled these responses for functional diversity, and agreed 79% with the LLM-judges (80% agreement between annotators). See Appendix Table 8 for details.

[3] Pairwise comparisons grow quadratically, which is why we only generate 5 responses per prompt.

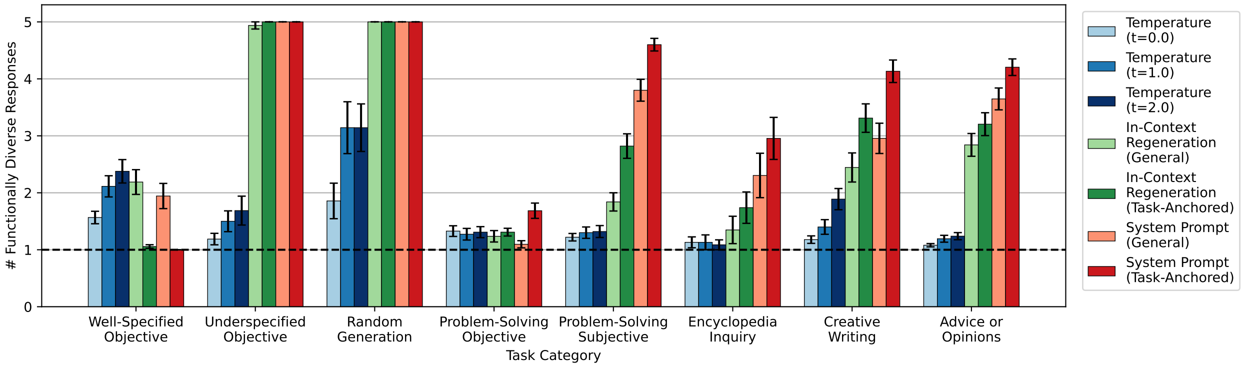


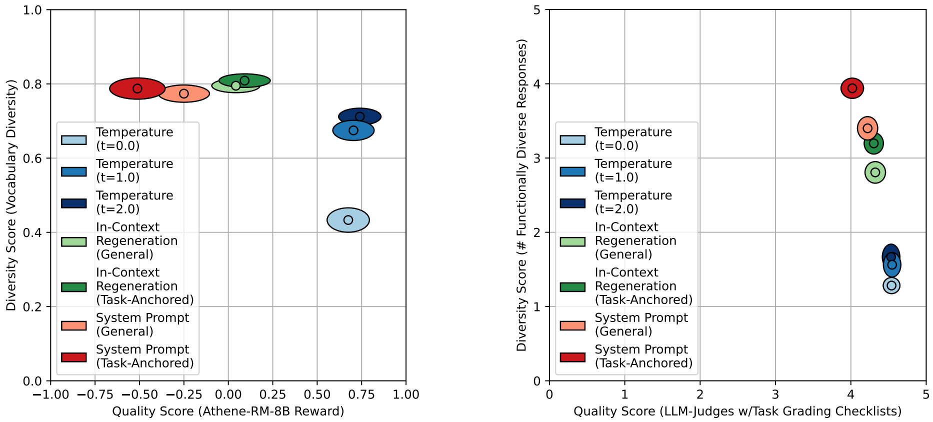
Figure 2 Our task-anchored sampling increases functional diversity for task categories where homogenization is undesired, while preserving homogenization where it is desired. We plot the average number of functionally diverse responses generated by GPT-4o for each sampling strategy and task category (with standard error). For the first category (Well-Specified Objective), bars closer to 1 reflect the preservation of output homogenization that is expected. For all other categories, bars closer to 5 reflect maximum functional diversity.

Well-Specified Objective Tasks (Category A) Tasks in this category have a single verifiably correct answer; thus, no functional diversity is expected. However, when employing general diversity-promoting sampling methods, homogenization is undesirably reduced, as evidenced by the generation of multiple unique answers (2 on average). Increasing temperature also undesirably reduces homogenization. In contrast, our task-anchored sampling method maintains homogenization, consistently producing one unique answer per task for GPT-4o, Gemini-2.5-Flash, and Claude-4-Sonnet.

Underspecified Objective and Random Generation Tasks (Categories B & C) Tasks in these categories are characterized by the existence of multiple verifiably correct answers, which suggests that models may easily conceptualize what difference means here. Consequently, we observe no significant differences between task-anchored and general sampling approaches, as the concept of diversity—defined as producing distinct correct answers—is inherently straightforward in this context. Both methods yield nearly maximal functional diversity, with approximately 5 unique responses out of 5 generations. In contrast, higher temperature settings result in suboptimal functional diversity, producing only 2 to 3 unique responses on average.

Problem-Solving Objective Tasks (Category D) Tasks in this category are defined by the presence of a single correct answer, but allow for multiple valid explanations or solution strategies. In this setting, we find that general prompt-based strategies are not effective in eliciting responses with diverse solution strategies. In contrast, both task-anchored system prompts and in-context regeneration sampling are able to generate approximately 2–3 distinct solution strategies. This relatively low number may be attributable to the inherent difficulty of the MATH-500 benchmark, which poses significant challenges for LLMs in producing even a single correct solution (Hendrycks et al., 2021).

Partial and Non-Verifiable Tasks (Categories E, F, G, H) Tasks in these categories cover subjective problem-solving, encyclopedia inquiries, creative writing, and requests for advice or opinions. Across all five models, our task-anchored sampling methods – both system prompt and in-context regeneration – reduce homogenization compared to their respective general approaches). For GPT-4o, Gemini-2.5-Flash, and Mistral-7B-Instruct-v0.3, task-anchored system prompting yields the highest number of functionally diverse responses. In contrast, for Claude-4-Sonnet and Llama-3.1-8B-Instruct, both task-anchored methods demonstrate comparable performance in promoting response diversity. Among temperature-sampled outputs, smaller open-weight models tend to have less homogenization than larger commercial models, possibly due to their less extensive alignment.



(a) General Metrics (GPT-4o Responses) **(b)** Task-Based Metrics (GPT-4o Responses)

Figure 3 With task-based metrics, diversity is improved with no significant drop in quality. We plot quality on the x-axis and diversity on the y-axis and compare the tradeoff under general metrics vs task-based metrics. In (a), there is a large tradeoff between vocabulary diversity (Def. A.1) and quality scores determined by a reward model. In (b), there is a negligible tradeoff between task-anchored functional diversity (Def. 3.1) and LLM-judges with task-based grading checklists. Note that the checklist-based quality difference between score 4 and 5 is "good" vs "very good". Plots show the mean and standard error of all metrics averaged across all task categories except category A, which we exclude because it is the only category where output homogenization is desired.

4.3 Diversity-Quality Tradeoff: Comparing General & Task-Based Metrics

We find that improved functional diversity from our task-anchored sampling often maintains the quality of responses, when the task-dependent nature of quality is captured in the quality metric.[4] Recent proposals for measuring quality using task-specific checklists align with our discussions around task-based metrics for diversity (Lin et al., 2025; Wei et al., 2025). Whereas, when quality is determined by a reward model, the scores do not inherently reflect task differences (e.g. the quality of a creative writing response is measured in the same way as the quality of a math problem-solving response). For GPT-4o responses, Figure 3 shows that the diversity-quality tradeoff prevalent in previous studies may simply be the result of mis-conceptualizing both diversity and quality. When evaluating general metrics, there appears to be a large diversity-quality tradeoff between vocabulary diversity and reward quality (Figure 3a), and the tradeoff is similarly large with embedding diversity (Appendix Figure 13). When we compare task-anchored functional diversity with checklist-based quality, there is a negligible diversity-quality tradeoff (Figure 3b). These results are similar for Claude-4-Sonnet and Gemini-2.5-Flash (Appendix Figures 9-10).

With task-based metrics, the diversity-quality tradeoff is more noticeable for smaller open-weight models (Appendix Figures 11-12), but still small (∼0.5 on our 5-point scale). One reason may be that smaller open-weight models have much lower task classification accuracy (only about 50% compared to 85% for commercial models). For commercial models, the quality slightly decreases when generating more than 5 responses using prompt-based methods (Appendix Figure 14). In particular, system prompt sampling may have lower response quality as the number of generated responses approaches the maximum output length for a single generation. Overall, our task-anchored sampling maintains the same level of quality as the general prompt-based strategies in previous work (Zhang et al., 2025a,b), while improving functional diversity.

[4] For tasks with singular verifiable rewards (Simple-QA & MATH-500), we separately validate the accuracy of responses (Appendix Table 26-27). Overall, our task-anchored sampling approaches often maintain and sometimes improve accuracy compared to temperature sampling. For MATH-500, system prompt sampling has the best accuracy for all models except Mistral. For Simple-QA, in-context regeneration performs the best for Gemini and Mistral, while system prompt sampling performs the best for Claude and Llama.