Task Category	Task Definition	Example Task	Functional Diversity	Reward Type	Example Dataset
A. Well-Specified Singular Objective	Tasks with a single verifiable correct answer	What is the largest Spanish-speaking country?	None	Verifiable	SimpleQA Wei et al. (2024)
B. Underspecified Singular Objective	Tasks with many verifiable correct answers	Name one Spanish-speaking country.	Different correct answers	Verifiable (Multiple)	NoveltyBench Zhang et al. (2025b)
C. Random Generation	Tasks that involve randomizing over a set of options	Roll a make-believe 6-sided dice.	Different pseudo-random options	Verifiable (Multiple)	NoveltyBench Zhang et al. (2025b)
D. Problem-Solving Objective	Tasks to solve a problem with a verifiable solution	How many positive whole-number divisors of 196?	Different solution strategies	Verifiable	MATH-500 Lightman et al. (2023)
E. Problem-Solving or Design Subjective	Tasks to solve a problem with many partially verifiable solutions	Design a room that minimizes energy consumption while maintaining comfort.	Different solutions or strategies	Partially Verifiable	MacGyver Tian et al. (2024)
F. Encyclopedia Inquiry	Tasks to provide information about real-world societies, traditions, events where there are credible references	Why is Isaac Newton famous?	Different factual perspectives	Partially Verifiable	Community Alignment Zhang et al. (2025a)
G. Creative Writing	Tasks that require creative expression	Tell me a riddle.	Different creative elements	Non-Verifiable	WildBench Lin et al. (2025)
H. Advice or Opinions	Tasks that solicit advice, opinions or feedback on specific topics/scenarios	What is a good Mother's day gift?	Different views or perspectives	Non-Verifiable	Community Alignment Zhang et al. (2025a)

Table 1 Taxonomy of Task Categories. Categories are distinguished by their concept of functional diversity. While non-exhaustive, task categories are a useful mechanism to clarify what elements of responses should be homogeneous and what meaningful elements of responses may vary.

3 Framework

3.1 Task Taxonomy

We begin by outlining the task categories used in our *task-anchored* framework (Table 1). Each of the 8 categories are distinguished by their conceptualization of output homogenization. The first four categories (A, B, C, D) capture prompts that elicit *verifiable* solution(s) that might be considered objective in nature, yet may still have more than one verifiable answer or explanation[1]. The second four categories (E, F, G, H) capture prompts that elicit more open-ended solution(s) that may be considered *non-verifiable* or only have partially verifiable components. Our taxonomy offers a more granular categorization of reward verifiability than the binary distinction (verifiable vs non-verifiable) in the literature (Lambert et al., 2024; Lanchantin et al., 2025b). Our task categories capture many real-world LLM usecases identified in recent work (Tamkin et al., 2024; Chatterji et al., 2025) (c.f. Appendix A.1). The categories are also motivated by recent studies that evaluate homogenization in specific task domains (§ 2.1). Though our taxonomy is non-exhaustive, we illustrate how task categories are a useful mechanism to appropriately conceptualize output homogenization.

Each task category corresponds to a degree of *reward verifiability.* Categories help clarify: what elements of responses are verifiable and should remain homogeneous for a given task? At one extreme, *Well-Specified Objective* (category A) captures prompts that have only one verifiable answer. Whereas, *Creative Writing* (category G) captures prompts that have an infinite number of non-verifiable answers. When we consider different types of verifiability, we realize that tasks may allow for multiple verifiable answers (categories B, C & E) as well as multiple explanations for those answers (categories D & E). For instance, *Problem Solving Objective* (category D) captures tasks that have a single verifiable answer, but multiple explanations available for arriving at that verifiable answer. Subjective problem-solving tasks (category E) may have multiple verifiable answers, as well as multiple explanations for those answers.

[1] We view objective/subjective and underspecified/well-specified not as discrete categories, but as spectra that task categories span. While the terms "objective" and "subjective" invite philosophical debate, such discussions are beyond this paper's scope. Likewise, we use "underspecified" and "well-specified" loosely, as clarifying terms, not precise definitions of the answer spaces.

Each task category further corresponds to a specific type of response variation or *functional diversity*. Previous works define two responses to be functionally diverse if a user would perceive them to be meaningfully different (Zhang et al., 2025b; Shypula et al., 2025). In this work, we define functional diversity based on our task categories, clarifying how responses could be meaningfully different for each task. For example, in *Problem Solving Objective* (category D), functional diversity is in solution strategies, not in the final answer. Whereas, in *Creative Writing* (category G), functional diversity is in the key creative elements (e.g. plot, genre, setting, etc.), not just in character names or vocabulary.

Functional diversity further depends on the level of *specification in the prompt*. For example, *Underspecified Singular Objective* (category B) may have multiple correct answers that could be generated, but the prompt does not specify a distribution over these answer options. An example task in this category is “Name one Spanish-speaking country.” In this prompt, it might be acceptable for the model to over-index on the most popular countries, but the prompt does not specify. Compare this to *Random Generation Objective* (category C) where an example task is “Roll a make-believe 6-sided dice.” The output distribution here is clearly specified by the prompt and not meant to be determined by the model. We aim to capture these nuances in how specified a task is. Ultimately, our taxonomy is simple yet powerful as a categorization that clarifies different types of reward verifiability and functional diversity that models may not inherently conceptualize on their own.

3.2 Evaluating Task-Anchored Diversity

Next, we formalize *task-anchored functional diversity*. We let $\mathcal{P}$ denote the set of possible input prompts or tasks and we let $\mathcal{Y}$ denote the set of possible outputs (e.g. sequences of tokens). We adopt the simple notation that a *language model* $\mathcal{M}$ is a stochastic function $\mathcal{M} : \mathcal{P} \to \mathcal{Y}$ that maps each prompt $p \in \mathcal{P}$ to an output $y \in \mathcal{Y}$. For a given prompt p, we assume $d(p, y_a, y_b) \in [0, 1]$ to be a pairwise diversity metric that indicates whether two responses y_a and y_b differ. To specify *functional diversity*, we anchor the definition of $d(p, y_a, y_b)$ on the task category $c(p) \in \mathcal{T}$, where each prompt p is is associated with a task category based on Table 1.

Definition 3.1 (Task-Anchored Functional Diversity)**.** Given a prompt $p \in \mathcal{P}$ with associated task category $c(p) \in \mathcal{T}$, and two responses $y_a, y_b \in \mathcal{Y}$, the *task-anchored functional diversity* is

$$d(p, y_a, y_b) := \mathbb{1}_{c(p)}[y_a \neq y_b],$$

where $\mathbb{1}_{c(p)}[y_a \neq y_b]$ is an indicator function that returns 1 if y_a and y_b are functionally different with respect to the task category $c(p)$, and 0 otherwise.

For example, consider whether two responses are functionally diverse in the *Problem-Solving Objective* task (category D). Here, $d(p, y_a, y_b)$ represents whether responses have different solution strategies, and assumes the single verifiable answer to be the same. In practice, evaluating functional diversity requires human annotation or LLM-judges. Based on the pairwise indicators of functional diversity, a set of responses can be partitioned into functionally distinct response groups as follows.

Definition 3.2 (Number of Functionally Unique Responses)**.** Let $\mathcal{Y}_p = \{y_1, \ldots, y_n\}$ be a set of responses for prompt p. Define an undirected graph $G = (\mathcal{Y}_p, E)$ where the edge set is $E := \{(y_a, y_b) \in \mathcal{Y}_p \times \mathcal{Y}_p : d(p, y_a, y_b) = 0\}$. Then, the *number of functionally unique responses* is $|\pi_0(G)|$ where $\pi_0(G)$ denotes the set of path-connected components of G.

Previous studies often evaluate response diversity with general diversity metrics. By general, we mean that the metric does not reference or depend on a predefined task category. For example, vocabulary diversity quantifies the extent to which two responses use different words where higher values mean more unique words or less words shared. Embedding diversity measures the difference in semantic content according to cosine distance in an embedding vector space where higher values mean more semantic difference. Appendix A.5 provides formal definitions for these metrics.

Method	**Previous Works** (No Task Dependence)	**Problem-Solving Objective** (Task-Anchored)	**Creative Writing** (Task-Anchored)
System Prompt	Generate `{num_responses}` responses that represent diverse values. (Zhang et al., 2025a)	The following problem has a single correct answer, but can be solved using different problem-solving strategies. Generate `{num_responses}` different solutions, each with a different problem-solving strategy.	The following prompt is asking for creative expression, so there are many possible subjective responses. Generate `{num_responses}` unique responses by varying the key creative elements such as tone, genre, point of view, theme, structure, etc. Each response should have different creative elements and reflect a distinct creative expression.
In-Context Regeneration	Can you generate a different answer? (Zhang et al., 2025b)	Can you solve the problem using a different strategy? The problem has a single correct answer, but can be solved using different problem-solving strategies.	Can you generate a new response with different creative elements? The prompt is asking for creative expression, so there are many possible subjective responses. Your new response should change the key creative elements such as tone, genre, point of view, theme, structure, etc.

Table 2 Prompt-Based Sampling Strategies. We modify prompt-based sampling methods in previous works (Zhang et al., 2025a,b) to promote task-anchored functional diversity.

3.3 Promoting Task-Anchored Diversity

We introduce a *task-anchored sampling technique* which modifies existing prompt-based methods for promoting diversity (c.f. Figure 1). Prompt-based sampling strategies are inference-time methods to generate multiple responses. We focus on two existing methods: *system prompt sampling*, which generates multiple responses in a single generation (Zhang et al., 2025a), and *in-context regeneration*, which iteratively generates multiple responses (Zhang et al., 2025b). Both these methods instruct the model to generate “different” or “diverse” responses. We modify these methods by clarifying in the instruction what is meant by “different” or “diverse”, using the functional diversity concepts in our taxonomy (Table 2, Appendix A.4). To reduce homogenization at inference-time, the model could sample over these responses, or choose a response based on other alignment criteria.

4 Experiments

In this section, we operationalize our framework for evaluating and mitigating task-anchored output homogenization. We use prompts from benchmark datasets that cover the task categories in our taxonomy (Table 1). In our task-anchored sampling technique (c.f. Figure 1), models first classify the prompt into a task category. Based on this classification, we explicitly instruct models to generate *different* responses where the instruction clarifies the task-specific concept of functional diversity. For comparison, we also sample *different* responses without task clarity (temperature sampling and general prompt-based sampling). With these experiments, we explore the following questions:

1. Compared to general sampling strategies, to what extent does our task-anchored sampling technique improve functional diversity across task categories?
2. How well do general diversity metrics capture task-anchored functional diversity?
3. With improved diversity, does our task-anchored sampling technique decrease the quality of responses?

4.1 Experiment Details

Models We evaluate responses from five models: *GPT-4o*, *Claude-4-Sonnet*, *Gemini-2.5-Flash*, *Llama-3.1-8B-Instruct*, and *Mistral-7B-Instruct-v0.3*. Separately, we use *GPT-4o*, *Claude-4-Sonnet*, and *Gemini-2.5-Flash* as LLM-judges (independent of response generation). When reporting LLM-judge metrics, we average the outputs across these three judge models.