LLM Output Homogenization is Task Dependent

Shomik Jain[1,2,†], **Jack Lanchantin**[1,*], **Maximilian Nickel**[1,*], **Karen Ullrich**[1,*], **Ashia Wilson**[2,*], **Jamelle Watson-Daniels**[1]

[1]FAIR at Meta, [2]Massachusetts Institute of Technology
[†]Work done during an internship at Meta, [*]Middle authors listed alphabetically

A large language model can be less helpful if it exhibits output response homogenization. But whether two responses are considered homogeneous, and whether such homogenization is problematic, both depend on the task category. For instance, in objective math tasks, we often expect no variation in the final answer but anticipate variation in the problem-solving strategy. Whereas, for creative writing tasks, we may expect variation in key narrative components (e.g. plot, genre, setting, etc), beyond the vocabulary or embedding diversity produced by temperature-sampling. Previous work addressing output homogenization often fails to conceptualize diversity in a task-dependent way. We address this gap in the literature directly by making the following contributions. (1) We present a task taxonomy comprised of eight task categories that each have distinct concepts of output homogenization. (2) We introduce task-anchored functional diversity to better evaluate output homogenization. (3) We propose a task-anchored sampling technique that increases functional diversity for task categories where homogenization is undesired, while preserving it where it is desired. (4) We challenge the perceived existence of a diversity-quality trade-off by increasing functional diversity while maintaining response quality. Overall, we demonstrate how task dependence improves the evaluation and mitigation of output homogenization.

Date: December 7, 2025
Correspondence: Jamelle Watson-Daniels at watsondaniels@meta.com

Meta

1 Introduction

Large language models (LLMs) often generate homogeneous outputs, but whether this is problematic depends on the specific task. Suppose a user asks for a joke and a model always responds with a "knock-knock" joke; such homogenization undermines the model's creative utility. By contrast, for tasks with verifiable solutions such as solving a math problem, consistency is not only acceptable but desirable, although variation in the explanation or problem-solving approach may still add value. Our central claim is that the implications of homogenization are task-dependent, and, therefore both the evaluation and mitigation of homogenization should also be task-dependent.

Existing approaches to reducing output homogenization rarely take task dependence into account. Several recent works propose methods that promote diversity in the alignment process or when sampling outputs at inference-time. However, these studies often fail to conceptualize diversity in a task-specific way. For example, some methods aim to increase token-level entropy or embedding-space variation in alignment (Chung et al., 2025; Lanchantin et al., 2025a; Slocum et al., 2025; Li et al., 2025b), while others promote diversity of viewpoints and perspectives when sampling multiple outputs (Wang et al., 2025b; Zhang et al., 2025a,b). Without a task-dependent approach, such methods may (1) fail to encourage diversity that is meaningful for a task, and/or (2) undesirably reduce homogenization in tasks where it is desired. We address this gap in the literature directly.

We introduce a task-anchored framework to evaluate and mitigate output homogenization. We build on the notion of *functional diversity* (Zhang et al., 2025b; Shypula et al., 2025), which asks whether a user would perceive two responses as meaningfully different for a given task. We argue that LLMs should be able to conceptualize functional diversity based on the task category. Consider the stakes: if a model wrongly conceptualizes functional diversity for a task that mimics an encyclopedia inquiry, the model could

misrepresent historical events in an attempt to naively reduce homogenization. Conversely, if a model wrongly conceptualizes functional diversity for a creative writing task, the model might repeat the same story arc no matter how many times a user asks the model to tell a story. We argue that task dependence should be incorporated into the way we address homogenization. Our contributions are as follows.

1. We present a *task taxonomy* of eight task categories each with distinct conceptualizations of output homogenization (§ 3.1). Our taxonomy extends the common distinction between verifiable and non-verifiable tasks. By introducing a more granular categorization, we aim to capture subtle nuances that may be overlooked if output homogenization is interpreted solely by the model. Although not exhaustive, our taxonomy effectively anchors task dependence. Note that if a prompt falls outside of our taxonomy, our approach can generalize to new task categories, or the model can resume its standard or default behavior.

2. We introduce *task-anchored functional diversity* to better evaluate output homogenization (§ 3.2). In our experiments, we compare to more general diversity metrics which are not task dependent (vocabulary and embedding differences). The results show that these general metrics fail to capture task-dependent diversity. Our task-anchored metric offers an alternative evaluation approach for future studies of output homogenization.

3. We propose a *task-anchored sampling technique* to increase functional diversity (§ 3.3), improving on previous sampling methods to promote diversity (Zhang et al., 2025a,b). Figure 1 offers a high-level illustration of our approach. We leverage our taxonomy to instruct models with task-dependent notions of diversity. Our approach increases functional diversity for task categories where homogenization is undesired, while preserving homogenization where it is desired (Figure 2).

4. We challenge the *perceived existence of a diversity-quality trade-off* (a common narrative in the literature) by adopting a quality measure (Lin et al., 2025; Wei et al., 2025) that also accounts for task-specific factors in quality. Figure 3 shows that the diversity-quality tradeoff prevalent in previous studies may simply be the result of mis-conceptualizing both diversity and quality. Our evaluation framework corrects both.

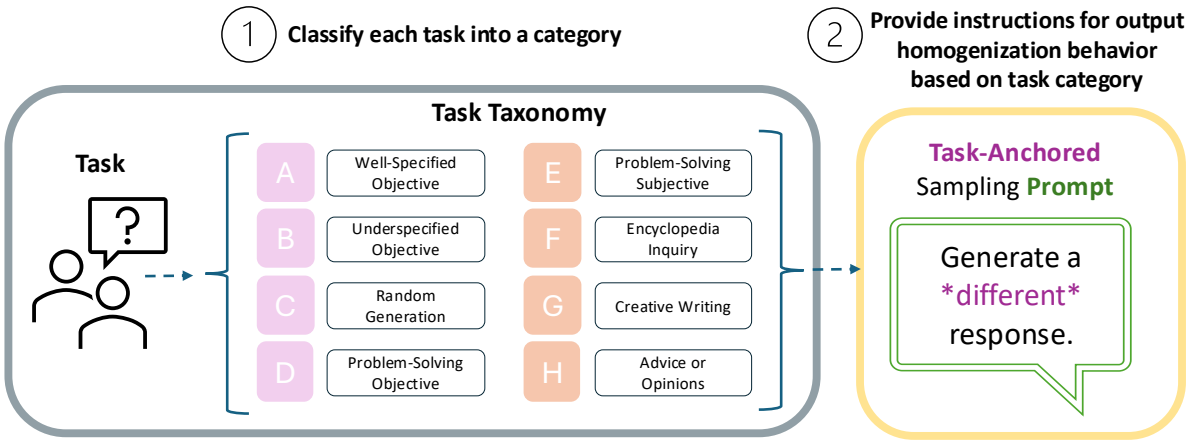

Figure 1 Our task-anchored sampling technique for improving output homogenization. The first step is to classify each input prompt into a task category. Note that if a prompt falls outside of the taxonomy, our approach can generalize to new task categories, or the model may resume its default behavior. The second step is task-anchored sampling where we clarify the concept of functional diversity in the instruction to generate "different" responses at inference-time. The taxonomy is outlined in § 3.1 and our task-anchored sampling technique is detailed in § 3.3.

2 Background

2.1 Homogenization in Aligned Models

Task Dependence Several works show that aligned LLMs exhibit output homogenization across a variety of tasks, such as creative writing (Moon, 2024; Wu et al., 2025), political discussions (Durmus et al., 2023; Santurkar et al., 2023), and math problem-solving (Slocum et al., 2025). Zhang et al. (2024, 2025b) further show that in question-answering, models often produce the same answer, even when the question is underspecified and multiple valid answers exist. These studies often evaluate homogenization in specific task domains, suggesting that problematic notions of homogenization are task-dependent. Our proposed taxonomy compares a variety of these task categories.

Representation Concerns In some tasks, homogeneous outputs may raise concerns about *representation*. The literature on pluralistic alignment (Sorensen et al., 2024; Chen et al., 2024; Zhang et al., 2025a) highlights representational harms, particularly when users seek advice or opinions from LLMs. However, pluralistic alignment discussions tend to operate in contexts where representation or diversity is presumed to be desirable. These discussions should recognize the task dependent nature of diversity, as we discuss in this work.

Causes of Homogenization There are many causes of output homogenization, such as limited diversity in training data and model design choices (Zhang et al., 2025a; Fazelpour and Fleisher, 2025). In particular, the alignment process is well-known to amplify homogenization in LLM outputs (Kirk et al., 2024; Lanchantin et al., 2025a). Models are typically aligned using methods such as Reinforcement Learning with Human Feedback (RLHF) (Ziegler et al., 2019) or Direct Preference Optimization (DPO) (Rafailov et al., 2023). These methods involve training on a dataset of pairwise preferences $\{(x, y^+, y^-)\}$, where x is a prompt, y^+ is a preferred response, and y^- is a dispreferred response. When there are conflicting preferences in the training data, such that both $y^+ \succ y^- \,|\, x$ and $y^- \succ y^+ \,|\, x$ coexist, the RLHF and DPO objectives implicitly reward putting all sequence-level probability on the majority preference (Slocum et al., 2025; Yao et al., 2025). Preference pairs with larger semantic differences also exert a stronger influence on the behavior of the aligned model (Chung et al., 2025; Shen et al., 2024). While this line of research is important, the present work focuses on how output homogenization should be conceptualized, not why it occurs.

Outcome Homogenization Another type of homogenization occurs when multiple models produce similar outputs (Kim et al., 2025; Wenger and Kenett, 2025). For example, outcome homogenization in decision-making refers to when individuals receive similar decisions from separate AI models (Bommasani et al., 2022; Jain et al., 2024b). In this work, we focus on the single-model case and do not deal directly with homogenization across different models. But reducing homogenization within a single model likely affects homogenization across models (Jain et al., 2024a, 2025).

2.2 Diversity-Promoting Methods

Alignment Methods A growing body of literature explores methods to reduce homogenization in aligned LLMs. Several studies propose modifying the alignment process, either by altering the construction of preference datasets or by adjusting alignment objectives (Lanchantin et al., 2025a; Slocum et al., 2025; Chung et al., 2025). All of these methods substantially increase diversity during alignment, as measured by token-level entropy or embedding-space variation. However, we highlight how these metrics may not capture meaningful, task-dependent notions of diversity.

Inference-Time Methods While most evaluations of homogenization examine temperature-sampled outputs, a few studies explore prompt-based strategies to explicitly sample diverse outputs at inference-time. For example, Zhang et al. (2025a) use a *system prompt* that explicitly tells LLMs to generate k responses in a single output that represent "diverse values." Zhang et al. (2025b) propose *in-context regeneration*, where models are prompted to produce a different response while retaining all previous responses in the conversation context. Other works also use implicit techniques, such as persona-based or multilingual prompting (Wang et al., 2025a,b). Our work improves inference-time methods by explicitly clarifying the notion of "diversity" in model instructions.