A Additional Experiment Details

A.1 Taxonomy Crosswalks

Our taxonomy is grounded in existing literature on LLM output homogenization and diversity. Specifically, we observe that many studies evaluate homogenization in specific task domains, suggesting that problematic notions of homogenization are task dependent. We designed our taxonomy to cover a variety of these task categories. Table 3 maps our taxonomy to related works that have evaluated output homogenization for each task category. To our knowledge, discussions of response variance in well-specified objective prompts (category A) are often found in studies of confabulation, not in the homogenization literature.

We further show that our task taxonomy captures many real-world LLM use cases identified in recent work. Chatterji et al. (2025) create a task taxonomy based on ChatGPT usage, which we map to our taxonomy in Table 4. Similarly, Tamkin et al. (2024) provide a list of common tasks based on Claude usage, which we map to our taxonomy in Table 5. We show that all real-world task categories in these previous works map to at least one category in our taxonomy (for text-based tasks). Many real-world task categories appear to correspond with multiple task categories in our taxonomy. For example, the "Practical Guidance" category in Chatterji et al. (2025) may correspond to Problem-Solving Subjective (Category E) or Advice/Opinions (Category H). This illustrates how our task categories are meant to capture different functional diversity concepts, while the categories in these other works are meant to summarize usage trends. For "Practical Guidance" tasks, we would consider them as Problem-Solving Subjective if responses span different partially verifiable solutions, and as Advice/Opinions if responses span different non-verifiable perspectives or views.

Ultimately, our task categories represent one categorization of functional diversity concepts, and a task may fall outside our taxonomy or correspond to multiple categories. In these cases, the model may resume its default behavior instead of using our task-anchored sampling technique, or promote diversity based on one of the applicable categories. Our approach is further generalizable to alternative taxonomies or task categories.

Table 3 Crosswalk of Our Taxonomy and Previous Output Homogenization Studies

Task Category	**Previous Works** (Non-Exhaustive)
A. Well-Specified Singular Objective	Wei et al. (2024)
B. Underspecified Singular Objective	Zhang et al. (2025b)
C. Random Generation	Hopkins et al. (2023); Zhang et al. (2025b)
D. Problem-Solving Objective	Lee and Lai (2024); Slocum et al. (2025); Wu et al. (2025)
E. Problem-Solving or Design Subjective	Ma et al. (2024); Yang et al. (2025)
F. Encyclopedia Inquiry	Sharma et al. (2024); Sui et al. (2025); Wright et al. (2025)
G. Creative Writing	Anderson et al. (2024); Doshi and Hauser (2024); Lanchantin et al. (2025a); Moon (2024); Moon et al. (2025); Padmakumar and He (2024); Wu et al. (2025); Zhang et al. (2025b)
H. Advice or Opinions	Agarwal et al. (2025); Durmus et al. (2023); Santurkar et al. (2023); Shahid et al. (2025); Zhang et al. (2025a)

Table 4 Crosswalk of Task Categories for ChatGPT Usage with Our Taxonomy

ChatGPT Task Category c.f. Table 3 in Chatterji et al. (2025)	**Categories in Our Taxonomy**
Writing (Edit or Critique Provided Text, Personal Writing or Communication, Translation, Argument or Summary Generation, Write Fiction)	Underspecified Objective (B), Creative Writing (G), Advice or Opinions (H)
Practical Guidance (How-To Advice, Tutoring or Teaching, Creative Ideation, Health, Fitness, Beauty, or Self-Care)	Problem-Solving Subjective (E), Advice or Opinions (H)
Technical Help (Mathematical Calculation, Data Analysis, Computer Programming)	Problem-Solving Objective (D), Problem-Solving Subjective (E)
Multimedia (Create an Image, Analyze an Image, Generate or Retrieve Other Media)	N/A (Our taxonomy is limited to text-based tasks)
Seeking Information (Specific Info, Purchasable Products, Cooking & Recipes)	Well-Specified Objective (A), Underspecified Objective (B), Encyclopedia Inquiry (F), Advice or Opinions (H)
Self-Expression (Greetings & Chitchat, Relationships & Personal Reflection, Games & Role Play)	Creative Writing (G), Advice or Opinions (H)

Table 5 Crosswalk of Top 10 Task Categories in Claude Usage with Our Taxonomy

Claude Task Category c.f. Figure 6 in Tamkin et al. (2024)	**Categories in Our Taxonomy**
Web and mobile application development assistance	Problem-Solving Objective (D), Problem-Solving Subjective (E)
Content creation and communication assistance across disciplines	Creative Writing (G)
Multidisciplinary academic research and writing assistance	Well-Specified Objective (A), Underspecified Objective (B), Encyclopedia Inquiry (E)
Education and career development assistance	Advice or Opinions (H)
Implement and optimize diverse AI/ML technologies and applications	Problem-Solving Objective (D), Problem-Solving Subjective (E)
Business strategy and operations assistance across industries	Problem-Solving Subjective (E), Advice or Opinions (H)
Multilingual NLP, translation, and linguistic analysis services	Underspecified Objective (B), Creative Writing (G)
DevOps and cloud infrastructure implementation and troubleshooting	Problem-Solving Objective (D), Problem-Solving Subjective (E)
Digital marketing and SEO optimization assistance	Problem-Solving Subjective (E), Advice or Opinions (H)
Data analysis, visualization, and management assistance	Problem-Solving Subjective (E), Advice or Opinions (H)

A.2 Evaluation Datasets

We sample 350 total prompts from the following datasets to use in evaluation of output homogenization. These datasets were chosen to achieve coverage across our task taxonomy (c.f. Table 6). For random sampling, we first shuffle the dataset using a random seed of 38, then select the required number of prompts in order from the shuffled dataset.

- **Community Alignment** (Zhang et al. (2025a)): A diverse human preference dataset containing user-generated prompts. We use 50 randomly-sampled prompts from the subset of user-generated first-turn prompts in English. Users were instructed to "ask, request, or talk to the model about something important to you or that represents your values. This could be related to work, religion, family, relationships, politics, or culture."
- **MacGyver** (Tian et al. (2024)): A dataset of creative problem-solving tasks. We use 50 randomly-sampled prompts from the subset of "solvable" problems that require "unconventional" solutions.
- **MATH-500** (Lightman et al. (2023)): A subset of the MATH dataset Hendrycks et al. (2021). We use 10 randomly-sampled prompts from each of the 5 difficulty levels.
- **NoveltyBench** (Zhang et al. (2025b)): A dataset of creative tasks where multiple distinct and high-quality outputs are expected. We use their entire curated dataset of 100 prompts.
- **SimpleQA** (Wei et al. (2024)): A dataset of short, fact-seeking queries across diverse topics. The prompts were created to be challenging for frontier models (e.g. GPT-4o accuracy $< 40\%$). We use 50 randomly-sampled prompts.
- **WildBench** (Lin et al. (2025)): A subset of the WildChat dataset (Zhao et al., 2024). WildChat is a corpus of 1 million user-ChatGPT conversations. WildBench is a filtered subset of WildChat such that tasks are diverse and challenging for models. We use 50 randomly-sampled prompts from the WildBench-V2.

A.3 Task Classification Into Our Taxonomy

When calculating functional diversity, we use ground-truth task categories for each prompt based on the source dataset and human-annotation. When generating responses in our task-anchored sampling technique, we use the model's task categorization of the prompt.

Table 6 shows the number of prompts by ground-truth category and dataset. For the SimpleQA and MATH-500 datasets, we classify prompts as category A (well-specified objective) and category D (problem-solving objective), respectively. For Community Alignment, NoveltyBench, and WildBench, two authors independently classified these prompts into categories. 11 prompts received disagreeing labels, which the annotators resolved after discussion. 6 prompts did not fit into our taxonomy (all from WildBench) due to missing information (e.g. prompts that referenced an unattached document) or language (we only evaluate English prompts).

We determine models' task categorization of each prompt using the judge prompt below. The agreement rate with ground-truth categories is 82% for GPT-4o, 86% for Claude-4-Sonnet, 84% for Gemini-2.5-Flash, 56% for Llama-3.1-8B-Instruct, and 46% for Mistral-7B-Instruct-v0.3.