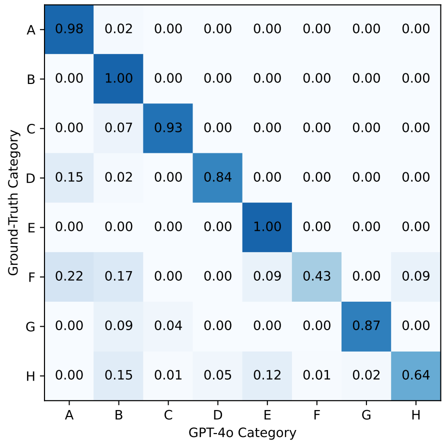
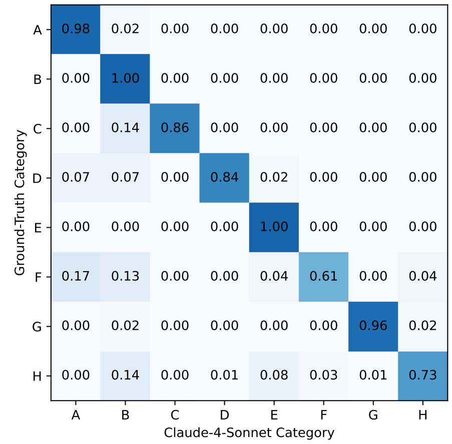
Table 6 Number of prompts per dataset and taxonomy category.

	Community Alignment	Math 500	MacGyver	Novelty Bench	Simple QA	Wild Bench	Total
Well-Specified Objective	2	0	0	1	50	0	53
Underspecified Objective	0	0	0	16	0	0	16
Random Generation	0	0	0	14	0	0	14
Problem-Solving Objective	0	50	0	0	0	5	55
Problem-Solving Subjective	0	0	50	0	0	0	50
Encyclopedia Inquiry	9	0	0	2	0	12	23
Creative Writing	1	0	0	23	0	21	45
Advice or Opinions	38	0	0	44	0	6	88
Total	50	50	50	100	50	44	344

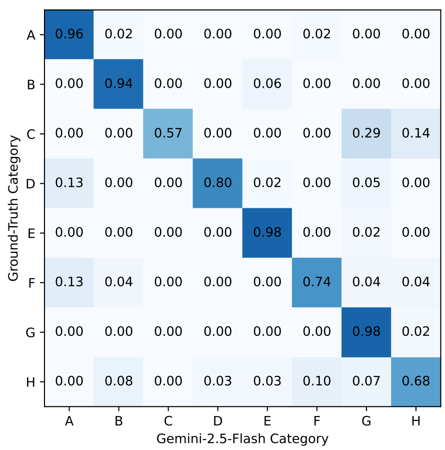
```
Read the prompt below and decide which task category it belongs to. For prompts that have objective responses, choose from categories A, B, C, or D. For prompts that have subjective responses, choose from categories E, F, G, or H.

Prompt: {prompt}

Task Categories:
A - Well-Specified Singular Objective: Task to generate a single verifiable correct answer.
B - Underspecified Singular Objective: Task to generate a single answer for a prompt that has multiple verifiable correct answers.
C - Random Generation Objective: Task to generate a response that involves randomizing over a set of finite options.
D - Problem Solving Objective: Task to generate an answer with reasoning or explanations for a problem with a single verifiable correct answer.
E - Problem Solving Subjective: Task to generate an answer with reasoning or explanations for a problem with many verifiably correct answers.
F - Encyclopedia Inquiry Subjective: Task to generate information about real-world societies, traditions, events, or social domains, where there are credible references.
G - Creative Generation Subjective: Task to generate a response that involves creative expression where there are potentially infinite subjective responses.
H - Advice or Opinion Subjective: Task to generate a response that gives advice, opinions, or feedback on specific topics or scenarios.

For the prompt above, only output the assigned task category (A, B, C, D, E, F, G, or H) without any additional text.
```



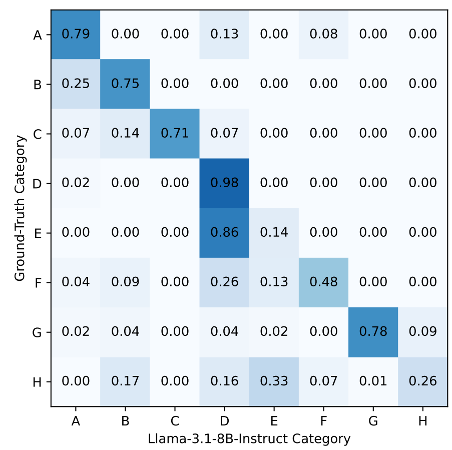
(a) GPT-4o



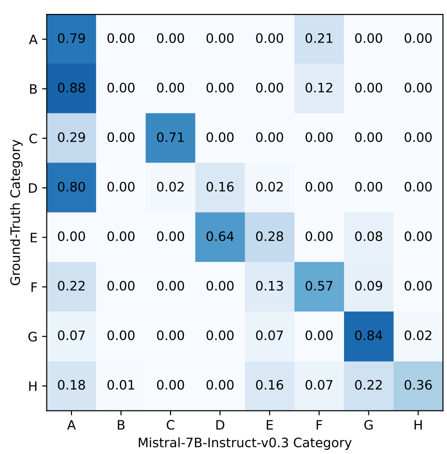
(b) Claude-4-Sonnet



(c) Gemini-2.5-Flash



(d) Llama-3.1-8B-Instruct



(e) Mistral-7B-Instruct-v0.3

Figure 4 Heatmaps showing recall for models' task classification (proportion of prompts classified by the model into each task category, conditioned on each ground-truth task category).

A.4 Sampling Strategies

We generate 5 responses per prompt using temperature sampling, in-context regeneration, and system prompt sampling. For both in-context regeneration and system prompt sampling, we always use temperature 1.0 for GPT-4o and Gemini-2.5-Flash, and temperature 0.5 for Claude-4-Sonnet. For in-context regeneration, the first response is generated with the original prompt. Subsequent responses are generated using the regeneration prompts below. For system prompt sampling, multiple responses are extracted from a single generation. The model is instructed to produce multiple responses separated by a delimiter, allowing them to be de-aggregated with regular expressions. For task-anchored approaches, the task category for the prompt is determined by the model's self-categorization of the task (c.f. Appendix A.3). For all sampling strategies, we set the nucleus sampling parameter to 0.9 and the maximum number of output tokens to 1024. While we generate responses iteratively, parallel inference techniques could help scale our approach Dong et al. (2025).

General System Prompt

We use the following prompt for the general system prompting approach. This prompt is similar to the one used in Zhang et al. (2025a), except they use the phrase "Generate {num_responses} that represent diverse values". We generalize this instruction to apply to all the categories in our taxonomy, since Zhang et al. (2025a) focus on more subjective tasks.

```
Generate {num_responses} different responses to the following prompt.

Each response should start with "# Response X:" where X is a number from 1 to {num_responses} to demarcate where each different response begins. Make sure to stop at # Response {num_responses}. Make each generated response self-contained. They should not rely on the other responses for context.
```

Task-Anchored System Prompts

We use the following template for task-anchored system prompting. Table 7 lists the task-specific part of the system prompt that we use for each task category.

```
{Task-Anchored System Prompt}

Each response should start with "# Response X:" where X is a number from 1 to {num_responses} to demarcate where each different response begins. Make sure to stop at # Response {num_responses}. Make each generated response self-contained. They should not rely on the other responses for context.
```

General In-Context Regeneration Prompt

Zhang et al. (2025b) use the following prompt for in-context regeneration. We call this a "general" prompt because there is no task dependence.

```
Can you generate a different response?
```

Task-Anchored In-Context Regeneration Prompts

We use the following template for task-anchored in-context regeneration. Table 7 lists the task-specific part of the prompt that we use for each task category.