{Task-Anchored In-Context Regeneration Prompt}

Do not include any starting phrases or reasons for why your new response is different. Your response should be self-contained, as if the prompt was the first thing that I asked.

Remember, the prompt is: {prompt}

Table 7 Prompts for Task-Anchored Sampling Strategies

Category	Task-Anchored System Prompt	Task-Anchored In-Context Regeneration Prompt
Well-Specified Objective	The following prompt has a single correct answer. Generate {num_responses} responses. If relevant, slight variation in wording is allowed but the answer should remain the same.	Can you generate a different response? The prompt has a single correct answer, so your answer should remain the same. If relevant, slight variation in wording is allowed.
Underspecified Objective	The following prompt is underspecified and has many correct answers. Generate {num_responses} responses, each with a different correct answer.	Can you generate a different correct answer? The prompt is underspecified and has many correct answers.
Random Generation	The following prompt is asking you to randomize over a set of finite options. Generate {num_responses} responses, each with a different pseudo-random option.	Can you generate a different pseudo-random response? The prompt is asking you to randomize over a set of finite options.
Problem-Solving Objective	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.	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.
Problem-Solving Subjective	The following problem has multiple correct answers, and may be solved using different problem-solving strategies. Generate {num_responses} different solutions, each with a different answer or problem-solving strategy.	Can you solve the problem using a different strategy? The problem has multiple correct answers, and may be solved using different problem-solving strategies.
Encyclopedia Inquiry	The following prompt is asking for information about the real-world, where there may be different factual perspectives. Your response must be grounded in credible references though references do not need to be mentioned. Generate {num_responses} responses that reflect different perspectives.	Can you generate a new response that reflects a different factual perspective? The prompt is asking for information about the real-world, where there may be different factual perspectives. Your response must be grounded in credible references though references do not need to be mentioned.
Creative Writing	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.	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.
Advice or Opinions	The following prompt is asking for advice or opinions, so there are many possible subjective responses. Generate {num_responses} unique responses where each response expresses a different viewpoint or perspective.	Can you generate a new response with a different viewpoint or perspective? The prompt is asking for advice or opinions, so there are many possible subjective responses.

A.5 Measuring Functional Diversity

We use the following prompt template to evaluate functional diversity using LLM-judges. Table 9 lists the task-specific parts of the prompt that we use for each task category.

We validate the LLM-judges on a stratified random sample of 225 response pairs across models, tasks, and sampling strategies. Two authors independently labeled these prompts for functional diversity. Table 8 reports the inter-rater agreement between annotators and each LLM-judge. When compared to the ensemble of LLM-judges, annotator 1 achieved 77.7% agreement and annotator 2 achieved 79.0% agreement. This is comparable to Zhang et al. (2025b)'s evaluation of functional diversity in creative tasks, where a fine-tuned judge achieved 79% agreement with human annotators.

```
For the given prompt and two responses, determine if the responses are functionally equivalent.
Functional equivalence means a user who has seen one response would find the other response to be
redundant.

{Task-Anchored Functional Diversity Definition}

###
Prompt: {prompt}
Response 1: {response 1}
Response 2: {response 2}
###

Are the responses functionally equivalent?
{Task-Anchored Diversity Judge Options}
Only output YES or NO.
```

To compare with functional diversity, we also measure diversity using three general diversity metrics that do not have task-dependence. For embedding diversity, we generate response embeddings using the gemini-embedding-001 model (with 3072-dimensional embeddings). For compression diversity, we use gZip to compress the concatenated text of response pairs Shaib et al. (2024).

Definition A.1 (Vocabulary Diversity)**.** Given two responses $y_a, y_b \in \mathcal{Y}$, let $\mathcal{V}_a$ and $\mathcal{V}_b$ denote the sets of unique words in y_a and y_b, respectively. The *vocabulary diversity* between y_a and y_b is

$$d_{\text{vocab}}(y_a, y_b) := 1 - \frac{|\mathcal{V}_a \cap \mathcal{V}_b|}{|\mathcal{V}_a \cup \mathcal{V}_b|},$$

where $|\mathcal{V}_a \cap \mathcal{V}_b|$ is the number of shared words and $|\mathcal{V}_a \cup \mathcal{V}_b|$ is the total number of unique words in both responses.

Definition A.2 (Embedding Diversity)**.** Given two responses $y_a, y_b \in \mathcal{Y}$, let $e(y)$ denote the embedding vector for response y. The *embedding diversity* between y_a and y_b is

$$d_{\text{embed}}(y_a, y_b) := 1 - \cos(e(y_a), e(y_b)),$$

where $\cos(e(y_a), e(y_b))$ is the cosine similarity between the embedding vectors of y_a and y_b.

Definition A.3 (Compression Diversity)**.** Given two responses $y_a, y_b \in \mathcal{Y}$, the *compression diversity* between y_a and y_b is

$$d_{\text{compression}}(y_a, y_b) := \frac{\text{compressed size of } y_a||y_b}{\text{size of } y_a||y_b},$$

where $y_a||y_b$ is the concatenated text of responses y_a and y_b.

Table 8 Annotator Agreement

	Annotator 1	Annotator 2	GPT-4o	Claude-4-Sonnet	Gemini-2.5-Flash
Annotator 1	-	79.9%	75.0%	77.2%	77.2%
Annotator 2	79.9%	-	79.0%	77.7%	80.4%
GPT-4o	75.0%	79.0%	-	90.6%	88.8%
Claude-4-Sonnet	77.2%	77.7%	90.6%	-	93.8%
Gemini-2.5-Flash	77.2%	80.4%	88.8%	93.8%	-

Table 9 Prompts for Functional Diversity LLM-Judge

Category	Task-Anchored Functional Diversity Def.	Diversity Judge Options
Well-Specified Objective	The prompt has a single correct answer. Responses are functionally equivalent if they represent the same answer.	Output YES if the responses represent the same answer. Output NO if the responses represent different answers.
Underspecified Objective	The prompt is underspecified and has many correct answers. Responses are functionally equivalent if they represent the same answer.	Output YES if the responses represent the same answer. Output NO if the responses represent different answers.
Random Generation	The prompt is asking for a random response over a set of finite options. Responses are functionally equivalent if they represent the same pseudo-random option.	Output YES if the responses represent the same pseudo-random option. Output NO if the responses represent different pseudo-random options.
Problem-Solving Objective	The prompt involves solving a problem with a single correct answer, but it can be solved using different problem-solving strategies. Responses are functionally equivalent if they represent the same problem-solving strategy.	Output YES if the responses represent the same problem-solving strategy. Output NO if the responses represent different problem-solving strategies.
Problem-Solving Subjective	The prompt involves solving a problem with multiple correct answers, and may be solved using different problem-solving strategies. Responses are functionally equivalent if they represent the same answer and problem-solving strategy.	Output YES if the responses represent the same answer and problem-solving strategy. Output NO if the responses represent different answers or problem-solving strategies.
Encyclopedia Inquiry	The prompt is asking for information about the real-world, where there may be different factual perspectives. Responses are functionally equivalent if they represent similar factual perspectives.	Output YES if the responses represent similar perspectives. Output NO if the responses represent different perspectives.
Creative Writing	The prompt is asking for creative expression where there are many possible subjective responses. Responses are functionally equivalent if the key creative elements (such as tone, genre, point of view, theme, structure, etc.) are the same.	Output YES if the responses have similar key creative elements. Output NO if the responses have different key creative elements.
Advice or Opinions	The prompt is asking for advice or opinions. Responses are functionally equivalent if they express the same viewpoint or perspective, even if they are worded differently.	Output YES if the responses have similar perspectives. Output NO if the responses have different perspectives.